CHAPTER 13

Treatment Research

Mark D. Rapport, Michael J. Kofler, Jennifer Bolden, and Dustin E. Sarver

This chapter provides a broad overview of principles, practices, and issues related to planning, conducting, and evaluating treatment research with children. We begin by discussing relevant issues, considerations, and obstacles in conducting child research. These include developmental considerations, measurement issues, informed consent, and confidentiality. The ensuing sections recapitulate the evolving debate concerning clinical efficacy and clinical effectiveness, and the evolution of empirically supported treatments (ESTs) for children. Methodological approaches that address the selection and evaluation of outcome measures appropriate for children—including the clinical significance of findings—are highlighted afterward. The last two sections provide broad coverage of research designs frequently used in child research and discuss factors relevant to identifying a research sample, including sample size and power.

MEASUREMENT ISSUES AND DEVELOPMENTAL CONSIDERATIONS

Measurement Issues

Conducting research with children requires broad knowledge of measurement issues related to maturation, selection of dependent measures, and the degree to which informants (parents, teachers, observers) agree with one another when providing information about children's behavior. This is due to several factors. Children's behavior is qualitatively different from that of adults and thus requires special consideration with respect to measurement and observation. Extrapolating adult measures for use in child research—by changing wording, modifying administration procedures, and making other adjustments—frequently fails to capture the intended underlying construct. Childhood behavior problems typically entail a broad range of symptoms, and these difficulties can affect functioning in multiple situations and settings. Observations of core and secondary symptoms may consequently vary significantly depending on setting and task demands. For example, children with Attention-Deficit/Hyperactivity Disorder (ADHD) typically experience chronic behavioral and academic difficulties at home and at school, and there is generally good agreement among observers. In contrast, children with internalizing disorders are frequently underrecognized until daily functioning is sufficiently

impaired to render their suffering more transparent. Multiproblem and multisituation assessment and analysis are thus the rule rather than the exception in child research.

Finally, children's rapidly emerging and changing verbal and cognitive abilities require careful consideration for research studies in which children are asked questions about themselves, others, and internal (e.g., mood/affect, anxiety, concentration) or external (e.g., activity level, impulsiveness) states. This consideration is complicated by extant research demonstrating differences among raters (e.g., parents, teachers, and trained observers), across settings (e.g., home and school), and for different types of behavior problems, as discussed later in the chapter.

Developmental Considerations

Treatment research with children is an intriguing but complicated domain of clinical psychology due to developmental factors that play a key role in our understanding of and assessing maladaptive behavior. For example, many behaviors associated with or mistaken as maladjustment and emotional disturbance are relatively common during childhood. Estimated prevalence rates based on parent reports indicate that fears and worries (43%), temper tantrums (80%), bedwetting (17%), and restlessness (30%) occur frequently in 6- to 12-year-old children (Lapouse & Monk, 1959) but do not necessarily portend later psychological dysfunction. In a similar vein, most children exhibit stranger anxiety around 8 months of age that mirrors their ability to discriminate a familiar face from an unfamiliar face. Separation anxiety becomes evident shortly after stranger anxiety begins, characterized by distress and an inability to be readily comforted by others in the absence of the child's parents or primary care providers. This is a normal response pattern in very young children, whereas excessive anxiety concerning separation from major attachment figures in later years accompanied by other behaviors—such as reluctance to attend school, unrealistic worry that an untoward calamitous event will separate the child from his or her parents, and somatic complaints—may signal the onset of a clinical disorder. Oppositional behavior is a normal developmental phenomenon in 18- to 36-month-old children; however, its persistence and accompaniment by other behavior problems is considered maladaptive in later years.

Behavioral and emotional problems evident in less severe forms during different developmental stages frequently represent normal developmental progressions in children, whereas their persistence may predict developing psychopathology. Consider the heterotypic continuity of ADHD symptoms as an example. Symptoms of ADHD arise early in childhood; however, their presence does not necessarily portend a persistent pattern of ADHD beyond 3 years of age in an estimated 50% to 90% of children so characterized (Palfrey, Levine, Walker, & Sullivan, 1985). Continuation of early ADHD-like symptoms to 4 years of age, however, is highly predictive of clinical hyperactivity at 9 years of age (Campbell, 1990). Thus, the early onset, degree, and persistence of symptoms past 4 years of age is highly predictive of a clinical diagnosis (indicating continuing and worsening impairment) and continuing difficulties throughout adolescence and early adulthood.

Other problem behaviors show clear developmental trends and are not considered maladaptive until late childhood. Lying and destructiveness are common exemplars. An estimated 50% of boys and girls engage in lying by age 6, based on parent report. Frequency of lying decreases to approximately 25% and 13% by age 7 in boys and girls, respectively, and continues to follow a downward trend as a function of increasing age in most children. Destructive behavior shows a similar developmental trend, peaking at 3 and 5 years of age in girls and boys, respectively, with a clear downward trend for both sexes through age 13 (MacFarlane, Allen, & Honzik, 1954).

Some maladaptive behaviors remain stable over time in terms of their frequency, but change with respect to topography or form. The manifestation of childhood aggression is a good example. Threatening, pushing, and shoving in young children frequently evolve into verbal and physical assault in older children, who continue to manifest aggressive behavior (Dishion & Patterson, 2006).

The foregoing examples inform us that many problematic behaviors in children will diminish or remit over the course of normal development and may not be appropriate or high priorities for intervention unless the behavior is sufficiently impairing and more rapidly alleviated by clinical treatment. They also highlight the importance of understanding the typical frequency and topography with which specific behavioral and emotional problems occur in children, and how base rates change over the course of development. Failure to appreciate and control for these variables may inadvertently result in attributing change to treatment rather than maturation or other historical events. Study design must also plan for possible topographical changes in maladaptive behavior if outcome assessment spans several years (e.g., in some cross-sectional and longitudinal studies). This may necessitate using different age norms or even different measurement instruments, which introduces unwanted error into the protocol.

A different set of developmental considerations comes into play when research studies require investigators to assign children to groups based on the presence or absence of a clinical disorder. Structured or semi-structured clinical interviews, coupled with other instruments and measures, are characteristically used to render a clinical diagnosis for purposes of group assignment (for reviews, see Orvaschel, 2006; Rapport, Timko, & Wolfe, 2006). Skilled interviewers have little difficulty interviewing adult informants, whereas interviewing children about their emotional state can be extremely difficult, and at times impractical, depending on developmental factors. Children's emotional development significantly affects their ability to respond to clinical interview questions about themselves and how they feel. The ability to label emotions and talk to others about feelings develops by about 2 years of age, and children have a reasonably well-developed range of emotional displays and labels by school age. Children can report changes in positive emotions before negative ones, but most experience difficulty describing variability in negative emotions until late childhood or early adolescence. Conscious awareness of and the ability to report on emotions as a state or system is not fully developed until middle or late adolescence (Haviland-Jones, Gebelt, & Stapley, 1997).

MEASURING AND ASSESSING CHILDREN'S BEHAVIOR

Informant Ratings of Children's Behavior

Behavior checklists and rating scales play a prominent role in child research. They serve as an important source of information concerning a child's behavior in different settings, how others judge behavior, and the extent to which behavior deviates from age- and gender-related norms. Information gleaned from rating scales contributes to the diagnostic process, and several scales serve as treatment efficacy measures.

Standardized scales are available that provide broad (e.g., internalizing, externalizing behavior problems) and narrow (e.g., measures of particular clinical disorders or states such as depression) indices of behavior, as well as for particular constructs (e.g., self-esteem, perceived competence) and other types of functioning (e.g., classroom performance, adaptive behavior, peer perceptions). Incorporating rating scales in research necessitates examination of the scale's psychometric properties (i.e., whether the instrument provides valid and reliable information with respect to what it purports to measure), general knowledge concerning the degree to which different informants can be counted on to provide valid information about a child's behavior, and factors that influence informant ratings. A judicious review of an instrument's psychometric properties is a necessary first step in constructing a research protocol. This is accomplished by checking with the publisher or author of the instrument, conducting a literature search of the instrument's psychometric properties, and researching specialty texts that cover tests and rating scales (e.g., *The Fifteenth Mental Measurements Yearbook*, edited by Plake, Impara, & Spies, 2003).

Extant research examining the degree to which raters agree concerning the presence of behavior problems reveals several, relatively consistent trends. In general, informants who interact with the child in the same environment (e.g., parents) tend to show better agreement in their reports of behavior than those who interact with the child in different environments (e.g., parents versus teachers versus mental health workers; Achenbach, McConaughy, & Howell, 1987). Agreement between parents, however, is less than ideal, as highlighted in a recent meta-analytic review (Duhig, Renk, Epstein, & Phares, 2000). Correspondence between mother and father ratings of children's behavior problems varied depending on the category of problem behaviors examined. For example, mean correspondence between parents for internalizing behavior problems was .45 as opposed to .63 and .70 for externalizing and total behavior problem scores, respectively—perhaps because externalizing behavior problems are easier to judge and/or more consistent across situations or settings (Achenbach et al., 1987; Walker & Bracken, 1996). Although a correlation of .45 is considered a moderate level of agreement between raters, it nevertheless indicates that only 20% ($.45^2$) of the variability in one parent's ratings can be explained (predicted) by the variability in the other parent's ratings. Higher correspondence between parents was reported for adolescents than for younger children when examining both internalizing and externalizing behavior problems, and the family's socioeconomic status appears to exert small but significant effects

on parent ratings. Agreement among raters also varies for clinical and nonclinical samples, is more variable for particular types of behavior and emotional problems, and is unacceptably low when comparing children's self-ratings to adult ratings (Achenbach et al., 1987).

The foregoing summary indicates that parents and teachers can be relied on to provide reasonably reliable ratings of children's behavior in the context in which they observe children for broad indices of behavior problems such as internalizing and externalizing behavior problems. Diminished correspondence between informants is evident, however, for ratings of more discrete types of behavior problems, when informant ratings are based on different settings, and when comparing children's own ratings to those of adults. Investigators need to consider these issues during the research design stage and minimize error variance by utilizing a multitrait, multimethod approach. This approach recognizes the undesirable shared overlap (variance) among similar measures and raters of specific constructs and behaviors and informs the investigator how to maximize the unique contribution of measures relevant to the variables studied (cf. Burns & Haynes, 2006, for a review).

Measurement Limitations

Limitations common to most rating scales include their reliance on subjective judgments and multiple threats to internal validity. These include halo effects, response bias, intensity and immediacy effects, and rater expectation bias (Harris & Lahey, 1982; McClellen & Werry, 2000). The underlying assumptions that Likert rating formats reflect interval-level measurement (i.e., that the unit of measure between 2 and 4 is identical to the difference between 1 and 3, and that these behavioral units are consistent across scales) and that all behavioral and emotional problems should be equally weighted (count the same when endorsed at the same level) represent additional psychometric challenges to clinical rating scales. For example, two children could receive identical Conduct Disorder rating scale scores, yet be dramatically different with respect to behavioral disturbance. One child might receive four 2-point ratings (indicating the highest level of severity or frequency) for items such as lying, stealing, cursing, and truancy, whereas the other child receives four 2-point ratings for breaking and entering, rape, firearm violation, and physical aggressiveness. Their total scores would be identical, yet the second child's symptoms are clearly more serious than the former's. Other factors, such as the presence of psychopathology among informants, the informant's experience with children, and other demand characteristics, represent additional challenges to clinical child rating scales.

Several desirable psychometric properties have not been thoroughly established for many child rating scales. For example, scale validity is usually accomplished by demonstrating that scores derived from one instrument correlate with those derived from an already established scale of the same latent variable, or by demonstrating that children known to be highly active (e.g., ADHD) score significantly higher than normal peers on the scale. Most scales meet at least one of these two criteria. Extant research, however, reveals that even when scale scores are correlated, they may be unrelated to objective measures of the same trait (e.g., Rapoport, Abramson,

Alexander, & Lott, 1971; Stevens, Kupst, Suran, & Schulman, 1978). For example, when measured by step counter, nearly 64% of children rated as clinically hyperactive were less active than the most active child rated as being normal by the teacher (Tryon & Pinto, 1994). Collectively, these shortcomings limit the interpretability and usefulness of activity rating scales.

The diagnostic utility of most rating scales is unknown; however, this situation has improved considerably over the past several years. Four metrics address this concern: sensitivity, specificity, positive predictive power (PPP), and negative predictive power (NPP). Sensitivity and specificity indicate the proportion of the group with a target diagnosis who test positive and negative on a measure, respectively. These two indices are useful for examining the overall classification accuracy of rating scales and other instruments, but are not particularly valuable to clinicians who are unaware of a child's diagnostic standing prior to referral. The statistics most relevant for this purpose are PPP and NPP. As it applies to rating scale utility, PPP indicates the conditional probability that a child exceeding a rating scale cutoff score meets criteria for a particular diagnosis such as ADHD (i.e., the ratio of true positive cases to all test positives). In contrast, NPP indicates the conditional probability that a child who doesn't exceed an established cutoff score will not meet criteria for a particular clinical diagnosis (i.e., the ratio of true negative cases to all test negatives). High values (e.g., >.80) for all four indices are desirable.

Finally, it is becoming increasingly common for investigators to create specific measures for assessing outcome by borrowing from measures with established psychometric properties and reconfiguring them for the study. For example, items contained in adult rating scales may be reworded and rekeyed (i.e., changed from a 1–5 to a 1–3 metric scale) before administering them to children. Others borrow from different scales that reflect a similar latent construct (e.g., anxiety) and combine or select specific items thought to better reflect the construct. These well-intentioned efforts ignore the backbone of psychological measurement: psychometric theory. Altering a measure, no matter how strong its existing psychometric properties, requires the investigator to reestablish its psychometric properties with the intended population (e.g., children) before using it in research. The researcher must also consider relevant developmental phenomena when reconfiguring rating scales and other questionnaire instruments. For instance, young children are unable to describe variability in negative emotions until late childhood or early adolescence, and usually cannot reliably discriminate between behavioral descriptions on a 5-point frequency or severity scale.

Interview Measures of Children's Behavior

Conducting detailed clinical interviews with children and their caregivers is a common method for screening participants, determining group assignment, and obtaining study data for child research investigations. Structured and semi-structured clinical interviews provide unique information beyond the data gained from rating scales and currently represent the only assessment option that allows clinicians to probe for the onset, course, and duration of endorsed symptoms—a necessity for differential diagnosis. As a result, they are considered the gold standard for

assessment and diagnosis. The distinction between structured and semi-structured interviews lies in the degree of freedom granted to the clinician to stray from a given script and ask open-ended, probing questions in response to symptom endorsements made by the interviewee. As a general rule of thumb, semi-structured interviews evoke more detailed information concerning presenting symptomatology owing to the extensive probing and clarification permitted (e.g., by asking for examples) but require greater clinical acumen and training.

The development of structured and semi-structured interviews represents the recognition of both the questionable reliability and validity of unaided clinical interviews and the frequent disagreement between parent and child reports of symptom endorsement and severity. Structured and semi-structured clinical interviews result in decreased error compared to unstructured clinical interviews arising from both internal (e.g., differential training level of clinicians, clinician biases) and external (e.g., discrepancies between informants) sources. The interviews typically require between 1 and 2 hours to complete, depending on a range of factors. These include the clinician's experience, the informant's ability to remain focused and recall historical information, and the severity, range, and duration of presenting problems (Table 13.1). The time investment limits their practicality for repeated use (e.g., for assessing treatment effects). Financial investment varies significantly across interview schedules. Some are free and available online (e.g., Schedule for Affective Disorders and Schizophrenia for School-Age Children [K-SADS]), whereas others have an initial cost of $600 in addition to $2,000 fixed training costs (e.g., Child and Adolescent Psychiatric Assessment [CAPA]). All but one of the semi-structured interviews covers all major *DSM-IV* diagnoses for school-age children through age 18. Some of the semi-structured interviews provide separate versions for parents, children, and adolescents. None of the available semi-structured interviews includes teacher versions. Clinicians must use other instruments to obtain school-related information for purposes of establishing impairment across multiple settings.

Strong convergent and discriminant validity is reported for most of the available semi-structured interviews, and test-retest reliability suggests stability of diagnosis over 1 to 3 years in clinical samples (Pelham, Fabiano, & Massetti, 2005). High sensitivity and specificity are typically reported for the semi-structured interviews, with only limited information available pertaining to PPP and NPP. The lack of PPP and NPP metrics is due in part to the use of semi-structured clinical interviews as the gold standard from which the predictive power of other measures (e.g., rating scales) is established.

The CAPA appears to be the most extensively developed of the clinical interviews, but it requires up to 2 weeks of classroom instruction and an additional 1 to 2 weeks of practice to acquire the necessary certification (Angold & Costello, 2000). This training is estimated to cost $600, and there is an additional fixed cost of $2,000, which may limit widespread usage among clinicians, especially when schedules such as the K-SADS demonstrate adequate reliability and validity and are available online at no cost. The CAPA, however, may be superior to other available schedules due to several excellent features. The instrument provides for an intensity rating that varies by three symptom groupings: intrapsychic phenomena

Table 13.1 Clinical Child Diagnostic Interviews

Measures	Age Range	Time (min)	Test-Retest (Kappa)	Symptom History	Disorders Considered	Scoring Format	Instrument Cost ($)	Training Required
Structured Interviews								
Diagnostic Interview Schedule for Children IV (DISC-IV)	6–17 (P) 9–17 (C)	90–120 (P) 45–90 (C)	0.79 (P) 0.42 (C)	4 weeks/ 12 months	All major Dx	Y/N	150– 2,000	2–3 day training module
Diagnostic Interview for Children and Adolescents IV (DICA-IV)	6–17	60–120	NR (P) 0.32 (C) 0.59 (A)	4 weeks/ 12 months	All major Dx	Y/N	1,000	2–4 weeks
Children's Interview for Psychiatric Syndromes (ChIPS)	6–18	40	0.4	NR	All major Dx	Y/N	115	NDR
Semi-structured Interviews								
Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS)	6–17	30–90	0.63	6 months/ lifetime	All major Dx	0–3	Free online	CTR
Semi-structured Clinical Interview for Children and Adolescents (SCICA)	13–18	60–90	0.57 (AP Scale)	NR	Does not correspond with *DSM-IV*		110–295 25 for 50	NDR
Child and Adolescent Psychiatric Assessment (CAPA)	9–17	20–210 $M = 66$ (P) 22–150 $M = 59$ (C)	NR for ADHD 0.55 for CD	3 months	All major Dx	0–3	600 + 2,000 Fixed Costs	BA
Interview Schedule for Children and Adolescents (ISCA)	8–17	120–150 (P) 45–90 (C)	Between 0.64–1.0	NR	All major Dx	0–3		CTR

Notes: Properties are equivalent for parent and child/adolescent version unless otherwise indicated.
A = Adolescent; AP = Attention Problems; BA = Bachelor's level training; C = Child; CTR = Clinical training required; Dx = Diagnosis; NDR = No degree requirements; NR = Not reported; P = Parent.

such as worrying, qualitatively different symptoms such as psychosis, and conduct disturbances. Training and coding are based on a detailed glossary, and thoroughly investigated symptoms are matched to appropriate glossary definitions and levels of severity. Formal rules are provided for the use of screening, mandatory, and discretionary questions.

The Diagnostic Interview for Children and Adolescents, *DSM-IV* edition (DISC-IV), provides separate versions for children (ages 6 to 12) and adolescents (13 to 18) based on field testing of interview questions with different age, gender, and racial groups. Two to 4 weeks of training at a cost of approximately $1,000 are required to reach the desired level of competence. The duration of training is based on clinician experience and includes topics such as age-appropriate probe questions and maintaining a child's interest through techniques such as tone of voice and appropriate nonverbal gestures. Reliability estimates corroborate research findings indicating that children are less reliable reporters of externalizing symptoms but more reliable reporters on internalizing symptoms relative to their parents. Computer versions are available, but initial research suggests poor reliability compared to the standard interview format.

The K-SADS is currently the most widely used semi-structured clinical interview. The Present-Lifetime (PL) version collects information from the parent regarding both current symptomatology as well as symptomatology at its most frequent and severe levels in the past. A separate interview is conducted with the child, and a third pair of ratings is generated based on integration of the parent and child reports with historical information and other data (e.g., rating scales). The K-SADS-PL focuses on chronology, treatment, impairment, and severity of symptoms. The initial screening interview consists of 82 items covering all major *DSM-IV* diagnostic categories. Cutoff scores are used to determine the need to administer the in-depth supplementary sections available for each diagnostic category, thus shortening administration time by allowing the clinician to skip supplementary sections based on negative endorsement of key screening questions. Interrater reliability estimates for the K-SADS are among the highest of any of the semi-structured clinical interviews. Extensive knowledge of diagnostic and symptom subtleties is required owing to lack of formal training requirements, and clinicians must use their judgment to interject noncued verbal probes to clarify informant responses and elicit examples of problematic behavioral and emotional symptoms.

Observational Measures of Children's Behavior

Behavioral observation is a time-consuming but valuable approach for obtaining detailed information about children and the environments in which they learn, play, and interact with others. It refers to a process in which observers record motor, verbal, and interactive behaviors of children using carefully defined operational criteria as opposed to evaluative judgments (e.g., rating scales) or mechanically activated devices (e.g., movement monitors, physiological monitoring). As such, many consider it the sine qua non of child assessment. An advantage of using behavioral observations is that the technique is virtually unrestricted with respect

to context. Observations can be conducted in naturalistic settings (e.g., to record classroom, playground, or bus riding behavior), analogue settings (e.g., a clinic room furnished to resemble a classroom), or specialty clinics (e.g., to assess parent-child interactions as part of a treatment program for children with Conduct Disorder, to assist a diagnostic evaluation for autism).

Numerous observational coding schemas are available for research purposes. Some permit coding of a wide range of behaviors (e.g., parent-child interactions, general classroom deportment) and are best suited for particular settings (e.g., home or classroom observations). Others permit highly refined and detailed observation, recording, and measurement of more discrete types of behavior (e.g., gross motor activity). Assessing interobserver agreement is nearly always required for research studies and involves pretraining on the selected coding schema and arranging for multiple observers to code the targeted behaviors simultaneously while in the setting or based on taped recordings. Detailed information concerning how to select particular stimuli, responses, and specific recording techniques; how to operationalize particular types, forms, and classes of behavior; and how to calculate interobserver reliability is readily available in classic texts on child behavior assessment (Kazdin, 1982; Ollendick & Hersen, 1998).

Direct observations represent the gold standard for experimental and outcome research for many investigators, but their use in clinical assessment is limited by several factors. No two research teams or commercially available observation systems define behaviors in exactly the same way, and research indicates that differences in observational schema can produce widely discrepant results in collected data (see Kofler, Rapport, & Alderson, in press, for a meta-analytic review).

Commercially available observation systems shown in Table 13.2 are available for school-age children and typically require between 10 and 30 minutes of observation. Multiple observation days are required to produce representative and reliable data; however, some systems offer software versions of their product, allowing researchers to record behavior directly onto a personal digital assistant (PDA) or laptop computer.

Direct observations are commonly used in school and institutional settings that require a formal functional analysis to determine whether antecedent and consequential stimuli or events contribute to a child's maladaptive behavior. These specialized assessments are typically conducted by highly trained school psychologists and individuals specializing in applied behavior analysis.

Although direct observations by independent observers can provide more objective and valid data than any of the other assessment methods discussed here, their relatively high temporal cost, coupled with the general lack of norms, suggests that their usefulness for assessment and group assignment may be limited to situations where large discrepancies exist among informant reports.

An alternative and less costly procedure for obtaining information relevant to children's classroom functioning is to ask classroom teachers to answer specific questions about a child's day or to save work samples. For example, desk checks ("Is the child prepared for class?"), teacher records of verbally intrusive behavior, and the percentage of daily academic assignments completed correctly appear to be useful for discriminating between children with and without ADHD (Pelham

Table 13.2 Mechanical and Observational Assessment Tools

Instrument/ Distributor	Age Range	Recording Length	Norms	Software Available	Cost
Mechanical					
Actigraphs Ambulatory Monitoring MiniMitter MTI, Inc.	Any	22 days per 32 kilobytes of memory	N	Y	Starter: $1,000+ (with necessary software and reader interface); $500–$2,000 for each additional actigraph
Actometers[a] Model 108 Engineering Department Times Industries Waterbury, CT 06720	Any	Variable	N	Y[b]	*NR*
Pedometers (available at sporting goods stores) Stand-alone with data downloadable to PC	Any	Range: 99,999 steps (~5.25 miles) to 1,000 miles	N N	N Y	$10–$40 $125–$400+
Direct Observations					
ADHD BCS Barkley, 1990	*NR*	15 min.	N	N	*NR*
AET-SSBD Sopris West	School age	15 min.	Y	N	Kit: $108 (includes all three parts of SSBD)
ADHD-SOC Checkmate Plus, Ltd.	School age	16 min.	N	N	Kit: $25
BASC-2 SOS AGS, Inc.	School age	15 min.	N	Y	25 forms @ $33
BOSS Harcourt Assessment	School age	15 min.	N	Y	Kit: $120
COC Abikoff, 1977/1980	School age	32 min.	N	N	NR
DOF ASEBA	5–14	10 min.	Y	Y	50 forms @ $25
SECOS Saudargas, 1997	Grades 1–5	20 min.	Y	N	
Noldus Observer Noldus Information Technology	Any	Variable	N	Y	Observer Basic 5.0 $1,795 Observer Video Pro 5.0 $5,850

Notes: AET-SSBD = Academic Engaged Time Code of the SSBD; BCS = Behavior Coding System; BOSS = Behavioral Observation of Students in Schools; COC = Classroom Observation Code; DOF = Direct Observation Form; NR = not reported; SECOS = State-Event Classroom Observation System; SOC = School Observation Code; SOS = Student Observation System.

[a]Many studies report either using the Kaulins & Willis actometers (no longer manufactured) or enlisting a jeweler to modify a self-winding wristwatch as described by Schulman & Reisman (1959).

[b]Eaton, McKeen, & Saudino (1996) provide SAS syntax for performing group-level data analysis based on actometer readings.

et al., 2005). These methods have the potential for objectivity that characterizes the independent observation methodologies but await critical evaluation to determine their utility for rendering diagnostic judgments at the individual level.

Potential Obstacles to Conducting Research with Children

Knowledge of regulatory statutes, the role and function of institutional review boards (IRBs), and general procedures for ensuring and obtaining informed consent is essential for conducting research with children. Detailed information concerning procedures and regulatory statutes related to studies involving children can be obtained from the Department of Health and Human Services website (www.nihtraining.com/ohsrsite/guidelines/guidelines.html). These regulations provide the minimum standards for protecting human subjects, particularly under the subpart D section, which includes detailed information concerning the involvement of children in research.

Regulations require the *assent* of the child or minor and the permission or *consent* of the parents or legally authorized guardian when children are involved in research. Assent is a child's affirmative agreement to participate in research. The standard for assent is the ability to understand, to some degree, the purpose of the research and what will happen if one participates.

Consent refers to the permission or agreement of parents or guardian to the participation of their child or ward in research. There are two forms of parental consent. *Passive* parental consent is a procedure that requires parents to respond only if they *do not* want their child to participate in a research project. *Active* parental consent requires the parent to return a signed consent form indicating whether or not they are willing to allow their child to participate. The latter method is recommended, although many researchers prefer the passive consent procedure because of the larger participant pool typically achieved. As a general rule for research involving minimal risk, it is sufficient to obtain informed consent from one parent.

Informed consent consists of three primary elements: knowledge, volition, and competency. Parents or guardians of children must be knowledgeable about all aspects of a study, which includes a thorough description of the facts, plausible risks, and potential sources of discomfort associated with an experimental study that may affect their decision to permit their child to participate. This information must be presented to the child's parents or legal guardian in an understandable fashion, followed by an opportunity for them to ask questions and clarify any issues that might not be thoroughly understood.

Volition refers to the process in which participants (or, in the case of minor children, their parents or guardian) agree to participate in a study free of coercion or threat. Subjects (or, in the case of minor children, their parents or guardian) must be able to decline or withdraw from participating at any time preceding or during an experiment without penalty, and should be informed of this provision (i.e., participation must be entirely voluntary).

Ensuring competency in child research typically entails the parents' or guardian's ability to render an educated decision and give consent for their child to participate based on thorough knowledge and understanding of the study.

A consent form signed by the parents or guardian of children indicates their willingness to have their child participate in the study based on an informed decision. These forms traditionally undergo formal review by an IRB or agency committee that evaluates the research proposal, consent procedures, and the form itself. Consent forms vary by design, but must contain a thorough description of the study, inherent and potential risks and benefits, procedures, and issues concerning confidentiality and how the study data may be used. Components of an informed consent form are described in Table 13.3.

Institutional review boards serve to assess the risks, possible benefits, and discomforts associated with a research project. Before beginning a project, an appropriate IRB committee must consider and formally approve all research proposals. University-based projects will typically be reviewed by the institution's IRB panel, and additional approvals may need to be obtained from other agency boards (e.g., hospitals, school systems, National Institute of Mental Health), depending on the selection of participants and whether the project is funded or sponsored by private or federal agencies. Many universities and research centers currently mandate

Table 13.3 Components of Informed Consent Forms

Section of the Form	Purpose and Contents
Overview	Presentation of the goals of the study, why this is conducted, who is responsible for the study and its execution.
Description of procedures	Clarification of the experimental conditions, assessment procedures, requirements of the subjects.
Risks and inconveniences	Statement of any physical and psychological risks and an estimate of their likelihood. Inconveniences and demands to be placed on the subjects (e.g., how many sessions, requests to do anything, contact at home).
Benefits	A statement of what the subjects can reasonably hope to gain from participation, including psychological, physical, and monetary benefits.
Costs and economic considerations	Charges to the subjects (e.g., in treatment) and payment (e.g., for participation or completing various forms).
Confidentiality	Assurances that the information is confidential and will only be seen by people who need to do so for the purposes of research (e.g., scoring and data analysis), procedures to assure confidentiality (e.g., removal of names from forms, storage of data). Also, caveats are included here if it is possible that sensitive information (e.g., psychiatric information, criminal activity) can be subpoenaed.
Alternative treatments	In an intervention study, alternatives available to the client before or during participation are outlined.
Voluntary participation	A statement that the subject is willing to participate and can say no now or later without penalty of any kind.
Questions and further information	A statement that the subject is encouraged to ask questions at any time and can contact an individual (or individuals) (listed by name and phone number) who is available for such contacts.
Signature lines	A place for the subject as well as the experimenter to sign.

Source: Clinical Psychotherapy: Developing and Identifying Effective Treatments, by A. E. Kazdin, 1988, New York: Pergamon Press. Reprinted with permission from Allyn & Bacon and the author.

that all persons involved in the conduct of research—from the investigators to the research assistants—take a specific training course and pass an exam concerning the ethical and appropriate conduct of research. Several of these are offered online (e.g., www6.miami.edu/citireg/), and results are forwarded to the host university's IRB as a precondition for conducting research.

CHILD RESEARCH SETTINGS

General Dimensions

A research study may be conducted in a laboratory (e.g., clinic, university laboratory) or an applied setting (e.g., school, home, hospital, community mental health center). The advantage of most laboratory settings is that they typically allow for maximal control over the experiment. Their most apparent disadvantage is that the nature of the setting is usually quite different from most real-life situations, which calls into question the generality of results. The advantage of applied settings is that they frequently represent real-life situations or environments, which allows for greater generalization of results. Their most obvious disadvantage is that they usually afford less control than a laboratory setting. Typically, one thinks of true experimental and quasi-experimental designs as taking place in a laboratory setting and observational designs as taking place in an applied setting. This is because the researcher attempts to exert a high degree of control in true and quasi-experimental designs, whereas few or no controls over extraneous variables are implemented in observational designs. True or quasi-experiments, however, can be conducted in applied settings, and observational studies can occur in a laboratory setting. Single-case research designs are used in both settings.

Gaining Access to Educational Settings

Conducting research in applied settings such as a school involves striking a balance between maintaining the requisite degree of scientific rigor and respecting the mission and structure of the educational setting. Research protocols must be planned such that disruption of the school or classroom routine is minimized. The investigator must be mindful that in the educational setting, classes (as opposed to individuals) are frequently the units of analysis. For protocols involving individual administration, pulling a child from some class periods (e.g., recreational reading) may be less disruptive than others (e.g., small group science activity) for the class as a whole and the child in particular. Moreover, schools may be especially protective of certain groups of students, such as children with developmental disabilities or behavioral/emotional problems. Conducting research in educational settings consequently requires careful planning and organization to maintain scientific rigor and accommodate the structure of the school. The ensuing discussion is particularly relevant for group studies in which a large sample of participants is required.

Subject recruitment is a necessary first step for most research projects, and schools are a logical source for studies involving children. The choice of school depends on a number of factors, which vary in degree of relevance depending on the focus of the investigation. Among the factors to consider are the socioeconomic

status of the community served by the school and the ethnic distribution of the school's population of students. For investigations involving multiple groups, the number of students enrolled in the school becomes an important consideration, in light of the power needs of the study's design to detect group differences.

Gaining access to research participants involves a stepwise progression to enhance the likelihood of a proposal's acceptance by educational institutions. Prior to contacting schools, it behooves the investigator to obtain endorsements from related agencies and the district superintendent of education—this smoothes the way to the engagement phase of decision making for the school principal. Subsequent contact with the principal should first be handled formally through an individually addressed letter that briefly explains the importance and nature of the study and mentions the endorsements received, followed by a telephone call approximately 1 week later to secure an appointment for a meeting.

Prior to contacting schools, the investigator must prepare to address any questions or concerns that may be forthcoming. Petosa and Goodman (1991) outline five phases of decision making that school officials undergo and discuss them in relation to decisions surrounding research proposals. In the *legitimacy phase,* the credibility of the investigator and the relevance of the study are of central concern. Representatives of the research project who contact the school should be prepared to address questions concerning the expertise and affiliation of the chief investigator, the source of funding for the project, and the project's relevance for the education and welfare of children or the community at large. In the second, *information-seeking phase,* the concerns of school officials involve the impact of the project on the day-to-day operation of the school. The investigator should consider the timing of the data collection with the school's schedule in mind. For example, the weeks preceding the winter holiday break and the end of the school year are often hectic for school personnel. Other times of the year, depending on the school system, are largely devoted to mandated standardized achievement testing. School officials will also wish to know the specific requirements of the project to gauge its impact on the school's mission and function. The investigator should be prepared to address questions pertaining to the measures to be administered, the class time required for participation in the project, the costs of participation accruing to the school (e.g., the extent of involvement of school staff), and the resources that will be available to the school (e.g., teacher training and curricular materials). During this phase, school officials will also be vigilant of any potential controversy that may arise due to the study's methodology, including parent objections to particular rating scales or survey questions and the potential for wasted classroom learning time.

In the ensuing *expression of limitations phase,* education officials are likely to express concerns over idiosyncratic situational factors that may pose obstacles to the school's participation in the project. The resourceful investigator should prepare to address concerns over issues of *fit* between the study's requirements and the school's structure, without undermining the integrity of the research project. If the investigator is successful, all parties enter the *engagement phase* of mutual problem solving, in which various ways of increasing the feasibility of the study are considered, such as finding the optimal dates and times and allocating physical space for data collection.

In the final, *commitment phase,* all parties make specific plans for the school's participation in the research project. Olds and Symons (1990) recommend presenting a *data collection participation form* for the principal's signature to secure the school's commitment. The initial start date for the study should be made far enough in advance to allow the school time to block the dates set aside for the data collection in its master calendar and to disseminate the information to parents and school staff. At least 1 week prior to the agreed-upon date for data collection, a follow-up reminder to the school should be made by telephone. At that time, information concerning any changes in the school's normal schedule or routine—such as assemblies or field trips—made subsequent to the meeting with the principal should be solicited.

The foregoing information and sequence of steps pertains primarily to schoolwide and classroom studies but is also applicable to conducting single-case research. Because single-case research impacts a classroom rather than a school, the key contact becomes the teacher and, as a courtesy, the principal, rather than the district superintendent. The decision-making process on the part of school personnel remains the same, however, and the investigator should follow the same suggestions for proactive planning and organization.

Additional considerations are warranted once project personnel are in the school to facilitate school support and cooperation. Gaining and maintaining the support of school personnel depend on the behavior and professionalism of research associates. These individuals need to be visible but unobtrusive, professional in appearance and demeanor, friendly and polite without exception, and respectful of school policies, procedures, and personnel. Research associates must be mindful of the school's mission to educate children in a safe environment and, from the outset, solicit information from school staff about such routine procedures as signing in and out upon entering and leaving the campus and whether identification badges are required. Most important, researchers should adhere to the original contract as agreed to by the school principal and refrain from such deviations as adding measures, collecting data on other dates, and attempting to deliver any intervention not previously agreed upon.

The research design and methodology are of prime concern to the investigator, but respect for the mission and structure of the educational setting is equally important. Demonstrating respect by minimizing disruption of daily operations increases the likelihood that the research proposal will be approved and ensures that the data collection process will run as smoothly as possible. Moreover, following through with promised benefits after the data have been collected represents a sound investment, as the school may be amenable to participation in future research projects. One means of accomplishing this is to volunteer to provide a teacher workshop that includes summarizing the study's results.

Clinic and Laboratory Investigations

Clinic- and laboratory-based investigations provide researchers with unique opportunities to manipulate carefully designed independent variables and observe their effects on dependent variables. These types of studies tend to have unparalleled

heuristic value and frequently test predictions stemming from extant models of child psychopathology. For example, a recent model of ADHD hypothesized a causal relationship between working memory and activity level in children (Rapport, Chung, Shore, & Isaacs, 2001; Rapport, Kofler, Alderson, & Raiker, in press) and challenged the notion that hyperactivity represents a ubiquitous deficit unrelated to task and setting demands (Porrino et al., 1983). Testing predictions stemming from the working memory model would be difficult to accomplish in a typical classroom for a variety of reasons. Measuring classroom activity level by utilizing actigraphs or behavioral observations does not represent a serious obstacle; however, introducing working memory tasks at varying degrees of difficulty in a counterbalanced manner to specific children and appropriate controls might prove impractical, if not unfeasible. Moreover, obtaining appropriate control over ambient noise levels (including ongoing student conversations or comments), lighting, chair type and placement, distractions, movement, and other setting characteristics is necessary to ensure that changes in the dependent variable (e.g., activity level) are due to differences in working memory demands (e.g., manipulating stimulus set size) rather than extraneous factors. The research clinic provides an ideal setting for these types of investigations.

Research clinics also enable investigators to scrutinize a vast array of personal and interpersonal processes by means of analogue investigations. These studies establish pseudo-situations that mimic important aspects of a child's life, either at home (e.g., asking parents to interact with their children following specific guidelines) or at school (e.g., having children attend a simulated classroom). Results of these studies often have important diagnostic and treatment implications. For example, observations of children with autism interacting with their parents in an analogue setting reveals a high degree of diagnostic sensitivity based on the child's reciprocal social interaction, play, stereotypic behaviors, gesturing, and communication (Filipek, Accardo, & Baranek, 1999). Teaching parents more adaptive ways to interact with their children in analogue settings has also shown exceptional promise for reducing Oppositional Defiant Disorder (Reid, Webster-Stratton, & Hammond, 2003).

CLINICAL EFFICACY, CLINICAL EFFECTIVENESS, AND EMPIRICALLY SUPPORTED THERAPIES

Overview

The terms *clinical efficacy* and *clinical effectiveness* emerged in response to calls for increased use of empirically validated treatments for individuals suffering from mental health problems. Debate evolved between research centers demonstrating that particular treatments resulted in significant client improvement following prescribed protocols, and the failure to find similar levels of improvement in applied settings such as community mental health centers. Hypothesized variables contributing to the lack of generalization of treatment effects across settings included differences in therapist training and orientation, use of manualized therapies to minimize therapist influences and maximize treatment integrity, and the extensive

oversight and supervision typically found in research centers. The essential concern was that treatments developed and tested in highly controlled settings were not enjoying the same level of success in applied settings.

Clinical Efficacy

The term clinical efficacy was coined to reflect the situation in which a particular treatment has proven useful and beneficial in treating psychological problems in children through controlled research. For example, Stein and colleagues (Stein, Jayxoc, & Kataoka, 2003) found that children treated with cognitive-behavior therapy for 10 weeks experienced significantly fewer symptoms associated with Posttraumatic Stress Disorder and depression, as well as improved psychosocial functioning and classroom behavior, relative to a 3-month wait-list control group in a randomized clinical trial.

Additional conditions are required to satisfy the recommended criteria for establishing clinical efficacy. The therapy must be significantly superior to no treatment, placebo, or an alternative treatment, or equivalent to a treatment already established as efficacious. These comparisons must be accomplished in the context of a randomized controlled trial, equivalent time-samples design, or controlled single-subject experimental design, and replicated by an independent research team.

Empirically Supported Treatments

Psychological therapies meeting clinical efficacy criteria are considered empirically supported treatments. Empirically supported treatments are psychological treatments proven effective for specific populations through controlled research (Chambless & Hollon, 1998). Meta-analytic reviews attest to the clinical efficacy of many therapies for children by computing and analyzing effect size estimates across studies. Thus, studies rather than subjects become the focus of analysis, and an overall effect size based on reviewed studies provides compelling evidence of the magnitude of the difference (in standard deviation units) between treatment and nontreatment groups. For those unfamiliar with meta-analysis, effect size estimates are calculated based on recommended formulas (Hedges, 1982) and involve subtracting treatment means from control group means and dividing by the pooled standard deviation of the two groups. The result is a number that reflects the difference between the two groups—for example, before and after therapy—in standard deviation units. Thus, an effect size of 1.2 indicates that the treated children are 1.2 standard deviations higher on a particular measure (e.g., in social competence) relative to the untreated group. Larger effect sizes are usually desirable and indicate greater mean differences in outcome measures between the treated and untreated or comparison groups.

Effect size metrics reported in meta-analyses can also be related back to a central clinical question. For example, Kofler et al. (in press) examined how often children with ADHD were on-task relative to control children in regular classroom environments. They used best case estimation to determine the magnitude of on-task deficits after accounting for significant moderators and obtained an effect size of 1.40. This effect size was translated onto the control group distribution to estimate

actual percentage rates, and revealed that children with ADHD were on-task an average of 74.7% compared to 87.9% for typically developing children.

Some of the published criteria for establishing a therapeutic intervention as an EST are dubious. For example, the superiority of the intervention must be demonstrated in at least two independent research settings *or* with a sample size of three or more subjects for single-subject design experiments also in at least two independent research settings. This means that a treatment can be considered clinically efficacious based on results from only six children—a rather weak case for suggesting that treatment effects are likely to generalize to other children with similar difficulties and in similar settings. The moniker *possibly efficacious* reflects the situation in which a single, well-controlled study supports the treatment's efficacy, highlighting the critical need for replication.

Clinical Effectiveness

In contrast to clinical efficacy, clinical effectiveness addresses whether the treatment is useful in applied clinical settings, that is, its clinical utility. The distinction between the two terms reflects the frequent failure to replicate treatment effects produced in highly controlled research environments—such as university- or hospital-based research clinics—in regular psychological service delivery systems such as community mental health centers and private practice settings. Variables likely to contribute to a treatment's lack of utility in applied settings, despite proving successful in controlled settings, are reviewed elsewhere (Chambless & Hollon, 1998; Kazdin, 1988). In response to these criticisms, several researchers have attempted to establish clinical effectiveness in applied settings; however, traditional design controls (e.g., random assignment to groups, nontherapeutic waiting-list groups) are certain obstacles to these efforts.

Relevant Issues

Determining whether a treatment or intervention is effective is not as straightforward as it may seem. Effectiveness varies significantly based on the type of measure (self-report, standardized rating scale, direct observation, clinical interview, objective measure such as an actigraph), the source of measurement (child self-ratings, teachers, parents, therapists), and the area of functioning assessed (e.g., social deportment versus academic functioning).

Many measures—and particularly rating scales—currently used in research studies have inadequate measurement units, insufficient psychometric properties, and lack age and sex norms. As an example, none of the available child activity level rating scales has a standard unit of measurement—that is, a meaningful unit of activity or movement that is equivalent within and across measures and raters. Accurately defining activity level is severely compromised as a result and leaves us in the uncomfortable position of describing differences and changes in children's activity level by referring back to the scales used to initially quantify the behavior. The moderate correlations reported between activity rating scales only partially mitigates this concern when contrasted with correlations between rating scales and objective measures of children's activity level. For example, correlations between activity

rating scales and actigraphs are generally between .32 and .58. These values indicate that 66% to 91% of the variability in activity rating scale scores is not linearly related to variability in actigraph scores in the same children measured at the same time. This finding probably reflects the fact that children's activity rating scales tend to reflect other aspects of behavior and not just activity level. The wording of most scale items and reliance on factor analytic scale construction methodology contribute to this phenomenon because descriptions of activity level may correlate highly, but usually reflect a broader range of behavior than just movement.

STATISTICALLY AND CLINICALLY SIGNIFICANT CHANGES

Overview

The importance of changes that occur following an intervention can be quantified in three ways. The first and most common approach is statistical significance testing, which examines group differences and estimates the reliability of change. Statistical significance testing calculates how certain we are that changes from pre- to posttest are robust—for example, not due to other factors such as measurement error. It does not tell us, however, about the magnitude of this change or its impact on children's lives. Effect size and correlational approaches can provide meaningful estimates of the magnitude of change, but these estimates do not necessarily have clinical or practical significance. Clinical significance refers to the practicality of an intervention—whether or not it makes a meaningful difference in the daily lives of children or those who interact with the child.

Intuitively, larger effects are more likely to be clinically relevant. The magnitude of effect, however, is not necessarily related to clinical significance. Kazdin (2004) has argued that an intervention can be clinically significant—or insignificant—with large, small, or no changes from pre- to posttest. As examples, he offers the possibility that a large but clinically insignificant effect may be obtained on a cognitive or laboratory task that has little to do with daily living. Likewise, a clinically significant finding of no change may occur when symptom deterioration was expected. The impetus of these examples is to demonstrate that a measure's ecological validity—its relevance to real life—must be considered when selecting measures used to test clinical significance.

Evaluating Clinical Significance

Several methods currently are used to evaluate clinical significance (Kazdin, 2004). Subjective evaluation methods rely on the judgments of the child or relevant others and ask whether they can discern meaningful improvement in their daily lives. Measures of social impact may also inform clinical significance, for example, whether there was a change in arrest record frequency, days worked, or missed school days. Absolute change methods are interested in determining whether a problem or symptom has been eliminated, or whether the child continues to meet diagnostic criteria for a psychiatric disorder.

The most common type of clinical significance testing is the *comparison method,* which asks whether members of the treatment group are distinguishable from

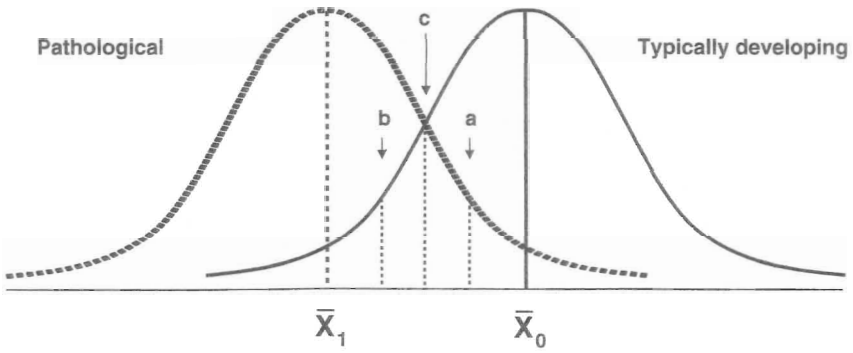

Figure 13.1 Clinical significance cutoff scores based on different criteria.

members of a typically developing (normal control) comparison group after treatment. Comparisons can be made to normative data or to a normal control group. Several commercially and freely available rating scales and direct observation systems provide standard scores based on normative data from nationwide community samples. Clinical significance may be demonstrated when a child changes from the clinical (usually 1.5 or 2 *SD* above the mean) to the normal (usually less than 1.5 *SD* above the mean) range following treatment. Comparison with rating scale norms is potentially problematic, however, due to different assessment conditions (Kazdin, 2004). For example, norms are typically based on a single test administration, whereas treatment outcomes are based on multiple administrations. Comparison with results from a concurrently collected normal control group is preferred.

Absolute change and two normal comparison, cutoff criterion methods are shown in Figure 13.1. X_0 represents the mean of the control group; X_1 represents the pathological group pretreatment mean. Three distinct cutoff scores have been proposed for determining clinical significance. The first (a) represents an absolute change method; the other two (b and c) estimate *normalization:* clinical significance in relation to a normal control group. Some argue that clinically significant improvement has occurred if a child's posttest functioning is (a) outside the range of the pathological group, defined as 2 standard deviations above the pathological control group pretreatment mean. Others argue that the child must fall (b) within the range of the normal population as defined by a score within 2 standard deviations of the normal group mean. A third proposal represents a compromise between these two options, defining normalization as occurring when a child is functioning (c) closer to the mean of the normal group than the pathological group.

In practice, children are classified as deteriorated, unchanged, improved, or normalized. The first three are determined by creating a confidence interval around the children's pretreatment score.* Children falling within the confidence interval are considered unchanged; those falling outside of it are labeled deteriorated

*Estimated true score is used if regression to the mean is considered a threat. See Speer (1992) for a discussion and simple method for determining whether regression to the mean is likely affecting posttreatment scores.

or improved. For example, children with a score of X_0 at pretreatment would be considered improved if they scored at or above cutoff score *a* in Figure 13.1. Children who improved and exceeded the predetermined cutoff score (*b* or *c*; see Figure 13.1) are considered normalized.

The difference between statistical and clinical significance is illustrated in the following example. Rapport, Denney, DuPaul, and Gardner (1994) evaluated the effectiveness of methylphenidate (Ritalin) on the classroom performance of children with ADHD. They used three outcome measures: teacher ratings of children's classroom deportment, academic efficiency scores (percentage of assigned class work completed correctly), and observed rates of on-task behavior in the classroom. Traditional significance tests revealed that all three outcome measures were significantly improved as a function of medication. Were these changes meaningful? Rapport and colleagues used the Jacobson and Truax (1991) model, correcting for regression to the mean as recommended by Speer (1992). As depicted in Figure 13.2, they found that between 50% and 78% of children were normalized, meaning that they were functioning within the normal control group range. Interestingly, although statistical significance was demonstrated for all three outcome variables, the rates of improvement or normalization were quite different. Most noticeable is the difference between teacher ratings of behavior, with 94% of ADHD children rated as improved or normalized, and actual classroom academic assignments completed correctly, with only 53% of ADHD children improved or normalized. Their findings also illustrate the importance of selecting relevant outcome

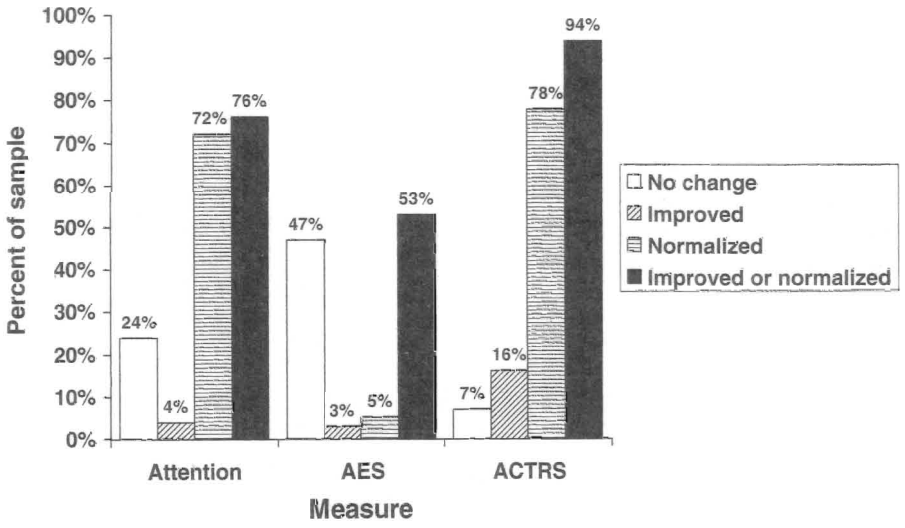

Figure 13.2 Clinical status of the group ($n = 76$) collapsed across methylphenidate dose conditions for three classroom measures. AES = Academic Efficiency Score; ACTRS = Abbreviated Conners Teacher Rating Scale. *Source:* "Attention Deficit Disorder and Methylphenidate: Normalization Rates, Clinical Effectiveness, and Response Prediction in 76 Children," by M. D. Rapport, C. Denney, G. J. DuPaul, and M. J. Gardner, 1994, *Journal of the American Academy of Child and Adolescent Psychiatry, 33*(6), pp. 882–893. Reprinted with permission.

measures. Teachers rated 94% of ADHD children as improving significantly when taking medication. When looking at their schoolwork, however, only 53% made actual improvements compared to their pretreatment performance. In most situations, when relevant outcome measures are used, statistical significance will be a necessary but not sufficient condition for clinical significance. Likewise, large effects may be more likely to be clinically significant effects. The presence of one or both of these conditions, however, does not demonstrate clinical significance.

In summary, clinical significance testing operates at the individual rather than group level. Beyond statistically reliable group differences, it is interested in the practical effects of an intervention—whether treatment makes a noticeable difference in the day-to-day lives of the clients we serve. Clinical significance can be defined subjectively by asking children about changes in their behavior, socially by comparing impact measures (e.g., school absences or test scores), or statistically using absolute and normative comparisons.

RESEARCH METHODS

Research *methodology* refers to the principles, practices, and procedures that direct research, whereas research *design* refers to the organized plan used to examine questions of interest. Both aspects represent important knowledge that guide the planning, implementation, and analysis of a research project.

Selection of Measures

Relationship between and among Variables

Selection of measures will depend on your specific research question as well as the level of understanding you wish to accomplish with respect to the phenomenon studied. For example, research may focus on examining *correlates,* or relationships between or among variables without consideration of time sequence or causality. Other studies may focus on identifying *risk factors,* or characteristics that precede and increase the likelihood of some event, outcome, or behavior pattern that may be modifiable. *Marker variables*—characteristics that precede and increase the likelihood of some event or behavior, but are not modifiable or have no effect on the outcome if modified—may be the focus of other investigations. Other measures will be more appropriate for questions concerning *treatment outcome* or *prevention* in children (for a detailed discussion, see Kazdin, 1999). Establishing causal factors, wherein one variable precedes and influences some other variable, represents the purest form of research if one embraces prediction of human behavior as the sine qua non of research.

Studying variables (*moderators*) that influence the direction, magnitude, and nature of a relationship between or among variables represents important avenues for clinical research, particularly when *protective factors* (a special case of moderator variables) can be identified that reduce the likelihood of an undesirable outcome. Researchers examine moderator variables when they suspect that the relationship among variables may differ because of the influence of some other variable. For example, if the relationship between type of aggressive behavior and a diagnosis of

Conduct Disorder is different for boys and girls, sex may be considered a moderator variable. Several moderating variables may be important to examine in research with children, including age, sex, ethnicity, height, weight, and motivation.

Once a causal relationship is established in the literature, researchers may wish to focus on explaining the process or mechanism by which the process evolves (*mediator variables* such as psychological or biological mechanisms). In this situation, a previously unconsidered variable is hypothesized to intercede between two related variables and reflects an indirect effect of one variable on another. For example, Rapport, Scanlan, and Denney (1999) found that the relationship between children's early attention problems and later scholastic achievement was mediated by differences in working memory.

In summary, a correlational relationship represents the most basic level of understanding between or among variables. A deeper level of understanding is achieved if it can be determined that one or more variables preceded the occurrence of other variables in time. To establish a causal linkage, investigators must provide evidence of a correlation and ordered temporal relationship between or among variables, complemented by three additional considerations: Other influences must be ruled out as possible explanations for the findings; the findings must be replicable across different samples from the same population; and a valid explanation concerning the mechanisms and processes through which the causal variables operate and are related to outcome variables must be explicated.

Types of Dependent Measures

The foregoing summary suggests that researchers need to consider the level of understanding they wish to achieve with respect to their research topic before selecting particular measures. Selection of measures will depend on a host of factors (e.g., availability of instruments or instrumentation, time constraints, costs) and can include a wide range of instruments and techniques. Popular exemplars include self-report inventories, behavior rating scales completed by others, and physiological instrumentation. Consultation of specialty texts on child assessment is advised for those interested in reviewing the broad range of instruments and techniques available (e.g., see Mash & Terdal, 1997; Hersen, 2006; and current volume).

Innovative measures can be created to assess discrete incidences (e.g., stereotypic movement, drooling) or categories of behavior (e.g., bizarre speech) but must meet expected standards of psychometric rigor (e.g., interrater reliability, validity). Computerized assessment is commonly used to assess important constructs such as vigilance (e.g., continuous performance tasks) and behavioral inhibition (e.g., stop-signal paradigm) in children. Incorporating current or archival indices of target behaviors or latent constructs reflects an additional method of data collection available to child researchers. For example, Fergusson and Horwood (1995) used arrest records as an index of later delinquency in children with early conduct problems and ADHD. Duration of hospital stay, school attendance, earned grades, completed academic assignments, and retention rates are other common examples of measures with high face validity used by researchers.

As a general rule of thumb, try to incorporate multiple indices of behavior that have strong psychometric properties. That is, select instruments, indices, or observations that provide valid and reliable measures of both the specific behaviors of interest as well as other potentially related areas of functioning. Doing so demonstrates that identified behavior problems or changes in behavior are not confined to a particular type of instrument or recording procedure and permits measurement of corresponding improvement in other domains related to the child's functioning. For example, designing a treatment outcome study to improve a child's academic performance in the classroom may result in corresponding improvement in the child's attention and classroom behavior. In such cases, teacher rating scales coupled with direct observational recording procedures that measure domains other than just academic performance are needed to document the broader spectrum of treatment gains related to the independent variable. Including collateral measures that reflect areas not expected to change (unless the researcher is expecting a generalized treatment effect) or undesirable change is necessary in some research protocols. The former provides evidence for discriminate validity (e.g., only the targeted area changes with the introduction of the independent variable), whereas the latter addresses the possibility of emergent symptoms or undesirable change due to the intervention.

Studies of specific clinical groups (e.g., ADHD, childhood depression), learning problems (e.g., academic deficits), or maladaptive behavior (e.g., peer aggression) may require measures different from those used to evaluate change due to intervention. For example, structured or semi-structured clinical interviews complemented by specific rating scales are traditionally used to define, classify, or describe children with a particular clinical disorder (i.e., serve as grouping variables), whereas other instruments, observations, or ratings may be used to measure change associated with intervention (i.e., serve as dependent variables). Therapy effects are strongest for outcome measures that match the problems targeted in treatment rather than through nonspecific or artifactual effects (Weisz, Weiss, Han, Granger, & Morton, 1995).

RESEARCH DESIGNS

Single-Subject Research Designs

Single-case research designs are valuable methodological tools used to evaluate different types of research questions involving individuals and groups. Consider some of the important scientific discoveries in psychology over the years. Single-case research was the principal paradigm in Wundt's (1832 to 1920) investigations of sensation and perception, Ebbinghaus's (1850 to 1909) studies of human memory, Pavlov's (1849 to 1936) classic experiments in respondent conditioning, and B. F. Skinner's (1904 to 1990) research in operant conditioning. These designs are particularly relevant to understanding children and the environments in which they live, where changes in everyday life may be of greater importance than obtaining a statistically significant change between groups (i.e., greater clinical relevance). Experimentation at the level of the individual case study may also provide greater insights with respect to understanding therapeutic change.

In contrast to the heavy reliance on established psychometric techniques and instruments required by between-group research, single-case methodology begins with identifying the focus of investigation (i.e., designating target behaviors) and proceeds to selecting potential strategies of assessment. When deciding on target behaviors, try to select observable behaviors and environment events as opposed to covert behaviors such as thoughts and ideas or hypothetical constructs such as anxiety and self-concept. Doing so will facilitate objective measurement and agreement between observers. Definitions should be written with absolute clarity to avoid ambiguity, and the boundaries of the defined target behavior must be clearly specified to minimize inference concerning whether a particular behavior qualifies as an occurrence or a nonoccurrence. On the surface, this may appear to be an easy task, but consider selecting classroom attentiveness ("on-task" in educational parlance) as a target behavior. Should off-task or inattentive behavior be recorded if a child is gazing up at the ceiling while working on a math assignment? Perhaps the child was thinking about the problem at hand or performing mental arithmetic. Are children permitted to ask peers for assistance when working on a problem, or allowed to leave their seat to sharpen a pencil during an academic work period? Does a momentary glance away from one's work count as off-task behavior? These and other definitional parameters must be decided and agreed upon, and preferably subjected to extensive pilot testing, prior to beginning formal observation and data collection. Once established, clear definitions of target behaviors will permit the recording of reliable baseline data and, in turn, serve as the traditional yardstick by which change is measured.

Assessing Behavior

Assessment of behavior can be accomplished in many different ways and will depend on what is being assessed and which method of recording best suits the needs of the researcher. The most commonly used methods of assessment include using frequency measures, classifying responses into discrete categories, counting the number of children or events, and measuring behavior based on discrete units of time.

Frequency counts are used when dealing with discrete behaviors that require a relatively constant time interval to perform. The first criterion enables observers to know when a designated behavior begins and ends, whereas the second permits recorded behavior to be treated as similar units for purposes of comparison. Consider the situation in which a researcher records the frequency with which a child speaks out of turn in a classroom. The child may blurt out an answer on one occasion and on another may turn and talk with a peer for 10 minutes or longer. Clearly, the two incidents are not comparable, and an alternative assessment method such as time interval recording should be considered. For situations in which frequency measures are taken for different periods of time (e.g., 20 minutes on day 1 and 40 minutes on day 2), calculate the frequency per minute or response rate by dividing the frequency of responses by the number of minutes observed each day. This metric, frequency per minute or response rate, will yield data that are comparable for different durations of observation.

Discrete categorization is used in situations in which behavior is best defined by categorical assignment. Commonly used categories include appropriate–not appropriate (e.g., social interactions between children), complete–not complete (e.g., classroom assignments), and discrete behaviors that form a functional (e.g., getting dressed in the morning, wherein each article of clothing counts as a discrete behavior) or correlated but unrelated (e.g., performing household chores) response chain.

Counting children is often used to assess the effectiveness of an intervention program. Examples include counting the number of children who perform a designated target behavior while on a school field trip, or calculating the percentage of children who complete their daily academic assignments in a classroom each morning. Contingencies can be introduced subsequently to reinforce daily academic assignment completion rates, for example, by scheduling in-class free time if at least 80% of the class meets the established criterion or for children who independently complete the assigned work.

Interval recording is used in situations in which the researcher wishes to obtain a representative sample of a target behavior during particular times of the day. Common examples of appropriate child behaviors include staying in one's seat during an academic work period, appropriate talk with peers, and paying attention. A block of time (e.g., a 30-minute daily or every other day observation period) is divided into a series of shorter intervals (e.g., 15-second observation blocks followed by 5-second recording blocks), and the target behavior is recorded as occurring or not occurring during each 15-second observation interval. This example would yield three observation blocks per minute and 90 intervals of recorded behavior per day. Variations of the interval recording method are common (e.g., time sampling) and might involve brief observations of a target behavior throughout the day rather than being confined to a single block of time.

Recording *duration* is another time-based method for observing behavior and is more appropriate for recording behaviors that are continuous rather than discrete acts. Common examples include observing ongoing social interactions between or among children or the total time required to complete an academic assignment. In these cases, the total duration or time interval that the behavior is performed serves as the dependent variable. An interesting but infrequently used variation of the duration method involves recording elapsed time before a particular behavior is performed (i.e., response latency). An example is to record how long a child takes to perform a particular behavior or chain of behaviors following adult instruction. Contingencies might be established subsequently based on the child's compliance within an increasingly shorter time interval over several days or weeks.

Design Types

The essence of single-subject research lies in its ability to demonstrate experimental control of an independent variable (IV) over one or more dependent variables (DV) by means of shrewd design and graphical illustration. Although statistical procedures exist that can be used to assess outcome effects associated with single-subject case studies, the more traditional means of demonstrating experimental control is to

provide a compelling visual (graphical) illustration that even a Doubting Thomas would acknowledge as evidence. To convince the scientific audience, an IV is introduced using a variety of design options such that its effects on the DV are systematically produced, reproduced, and/or eliminated, as in a carefully choreographed dance. Detailed descriptions of specific types of single-subject designs are provided in this volume (see Chapter 11).

Group Experimental Designs

Overview

Despite the numerous benefits associated with single-subject design methodology, the fact remains that an overwhelming majority of studies in psychology involve the comparison of groups, not individuals. Group designs are conventionally classified as true experimental designs, quasi-experimental designs, and observational designs, and may be used in a variety of contexts.

A true experimental design is a design in which the researcher manipulates one or more independent variables and measures one or more dependent variables. The researcher chooses what independent variables to manipulate, how they are manipulated (e.g., which levels to include), and what dependent variables to measure, based on the nature of the research question. For example, a researcher might be interested in determining what type of therapy works best for children with school phobia. Therefore, the research question has determined that the independent variable is *type of therapy* and the dependent variable is *school phobia*. Now, the researcher must decide whether any other independent or dependent variables should be included in the study and operationally define the independent variable(s) and the dependent variable(s).

Operationally defining independent variables refers to deciding what the levels should consist of, whereas operationally defining dependent variables refers to deciding exactly how to measure them. For example, the researcher needs to decide what levels of therapy to include (e.g., behavior therapy, cognitive therapy, cognitive-behavior therapy), whether or not to include a control group (condition), and how to measure school phobia. In the context of a between-subject design, a control group is a randomly assigned group that receives either no experimental treatment or a substitute for the experimental treatment. In the context of a within-subject design, a control condition is a condition (administered to all participants) in which either no experimental treatment or a substitute for the experimental treatment is administered. Control conditions (groups) are sometimes needed to control for nonspecific treatment effects (e.g., some children might believe that simply taking a pill helps because it's "medicine").

A quasi-experimental design is a design that is set up to emulate a true experimental design but includes one or more independent variables that cannot be manipulated by the researcher (often referred to as quasi-independent variables). These include variables such as sex, ethnicity, height, and weight, or clinical diagnoses such as ADHD or Reading Disorder. Another reason for the inability to manipulate certain variables may be due to ethical reasons such as substance abuse, smoking, exposure to harmful toxins, and cancer. The limitation to quasi-experimental

designs is the inability to express a causal relationship for quasi-independent variables.

Nonspecific Treatment Effects

Nonspecific treatment effects are any effects brought about by the experiment besides the treatment, such as being aware of what the experiment is about, contact with the experimenter, and discussing the experiment with other people. Nonspecific treatment effects are basically extraneous (nuisance) variables related to participants' perceptions concerning the experiment. Related to nonspecific treatment effects are placebo effects, in which participants improve simply because they believe that they are receiving treatment. If a control condition (group) is included, it is important that there are no other differences between the control condition and the experimental conditions except for an absence or substitute for the experimental treatment. Any other differences would introduce extraneous (nuisance) variables. Control conditions are necessary only when there is concern that nonspecific treatment effects may have an effect on the results. Nonspecific treatment effects are possible only if the participants' perceptions about the experiment can have an effect on the results. There are many situations in which participants' perceptions cannot have an effect on the results. For example, if a researcher is interested in studying what type of teaching method works best for teaching mathematics to third-grade children, a control condition consisting of teaching no mathematics would be benighted. Conversely, a child with social phobia may believe that taking a pill provides him with the necessary confidence to complete an in vivo exposure protocol.

There are three main classes of experimental designs: the between-subject design, the within-subject design, and the mixed-subject design, each with multiple variations possible.

Between-Subject Designs

The between-subject design is a design in which participants are randomly assigned to different treatment groups (levels of the independent variable) and each treatment group receives a different experimental condition. For example, in a single-factor experiment (an experiment with only one IV) examining two methods for teaching science to third-grade children (Technique 1 and Technique 2), half of the students would be randomly assigned to receive Technique 1 and half to receive Technique 2.

In a factorial experiment—an experiment with more than one independent variable—examining different methods for teaching science to third-grade children (Technique 1 and Technique 2) and the mode of presentation (teacher versus computer), one fourth of the students would be randomly assigned to receive Technique 1 via computer, one fourth to receive Technique 1 via teacher, one fourth to receive Technique 2 via computer, and one fourth to receive Technique 2 via teacher.

Advantages of the between-subject design as compared to the within-subject design include no carryover effects and no order or sequence effects. Disadvantages

of the between-subject design include the fact that many more participants are required than for the within-subject design, and the between-subject design yields less power than the within-subject design.

Ideally, there are an equal number of participants in each of the different treatment groups and assignment to the different treatment groups is random. Random assignment to treatment conditions simply means that each participant has an equal probability of being assigned to any given treatment group. Both ideals, equal number of participants per treatment group and random assignment to treatment groups, may not be possible for a given study. In the case of an equal number of participants per treatment group, the total number of participants may not be equally divisible among the various treatment groups or some participants may drop out of the study or miss the day that data are collected. In the case of random assignment of participants to the different treatment groups, this may not be possible. As in the example concerning methods of teaching science, it might be difficult to randomly assign half of each class to receive a different teaching method. The teacher cannot very well teach half of her class at a time, and even if he or she could, this would introduce the confound of which method is taught first. Both problems (unequal sample size and nonrandom assignment) can frequently be handled statistically. It is important, however, to take into account both unequal sample size and nonrandom assignment when analyzing data (i.e., one should not ignore them and analyze the data as if there were equal sample size and random assignment).

Another important issue is the loss of participants from the experiment, referred to as *attrition*. If the loss of participants is random (e.g., roughly an equal number of participants dropped out of each of the experimental groups), there is no problem and certain statistical techniques may be used to correct the situation. If, however, the loss of participants is not random but due to some aspect of particular treatment conditions (e.g., almost every participant who dropped out was in one particular treatment condition), then little can be done to salvage the experiment other than conducting a post-hoc test to determine whether attrition significantly biased one group relative to the other. For example, one may need to demonstrate that children with the most deviant scores did not drop out of one of the two groups receiving treatment, or that noncompleters were not significantly different on particular measures relative to completers.

As an additional note, the problems associated with randomly assigning some children within a given classroom to receive one level of the independent variable (treatment) and other children from the same classroom to receive other levels of the independent variable is not uncommon to research conducted in school settings (or even clinic or hospital settings). For example, it may not be possible for the teacher to teach some students using one teaching method and the other students using different teaching methods. When it is not possible to randomly assign children from the same class to different levels of the independent variable, an alternative is to randomly assign different classrooms to the different levels of the independent variable. This type of design is called a *hierarchical design* and requires special analysis. In a hierarchical design, the levels of at least one independent variable are nested under the levels of another independent variable, and the remaining independent variables are fully crossed. For example, if each level of IV 2 (e.g.,

Table 13.4 Hierarchical Design

	Technique 1		Technique 2
Classroom 1 Classroom 2		Classroom 3 Classroom 4	

classrooms) appears with only one level of IV 1 (e.g., teaching method), then IV 2 (classrooms) is said to be nested under IV 1 (teaching method). Examining two different methods for teaching science to third-grade children, two third-grade classes could be randomly assigned to receive Technique 1 and two third-grade classes could be randomly assigned to receive Technique 2. The design for this model is diagramed in Table 13.4.

The separate columns for Technique 1 and Technique 2 indicate that the model is not fully crossed, as classrooms (1, 2, 3, 4) are nested under techniques (1, 2). In this experiment, the independent variable *classroom* (classrooms 1, 2, 3, 4) is an extraneous (nuisance) variable. It is included in the design and analysis because it might have an effect on the dependent variable, and including it allows its effects to be isolated. If a hierarchical design is used, it is incorrect to analyze the data as if students were randomly assigned to each level of the independent variable (e.g., teaching method). There are additional complications that may arise when employing a hierarchical design, for example, when the levels of the nested variable (classroom) cannot be randomly assigned to the levels of the variable it is nested under (teaching method). Hierarchical designs are considered balanced if they satisfy two criteria. First, there must be an equal number of participants in each treatment combination (e.g., students in each class for each teaching method). Second, there must be an equal number of levels of the nested variable under each level of the other independent variable (e.g., two classes under each level of treatment method). If these criteria are not satisfied, the model is considered unbalanced and more complicated to analyze than balanced designs.

Within-Subject Designs

Every participant receives every treatment condition (levels of the independent variables) in a within-subject design. In the single-factor experiment on teaching techniques, every participant would receive Technique A and Technique B. In the factorial experiment involving teaching technique and mode of presentation, every participant would receive Technique A via computer, Technique A via teacher, Technique B via computer, and Technique B via teacher. A within-subject design may have children participating in the different levels of the independent variable simultaneously, or children may complete one level of the independent variable before participating in the next level. When participants must complete one level of the independent variable before participating in the next level, this is commonly referred to as a *crossover design*.

Within-subject designs present some special problems, including carryover effects, order effects, and sequence effects. Because participants receive every treatment condition, the effects of one treatment condition may carry over to the next

treatment condition. For example, once children are exposed to a particular reading method, they may be irreversibly changed. The investigator cannot simply cross them over into an alternative method and assume that nothing has been gained during the earlier experience. The order in which the treatment conditions are presented to the participants may also have an effect on the results. Finally, the sequence in which the treatment conditions are presented may have an effect. For example, in a study of perceived heaviness of objects, whether someone lifted a 10 pound object and then a 20 pound object or lifted a 20 pound object and then a 10 pound object would have an effect on the perceived heaviness of each object. Order and sequence effects may sound similar; however, order effects refer to the ordinal position in which the condition is presented in (e.g., first, second, third). Sequence effects, in contrast, have to do with which treatment follows which other treatment. As an example, consider the two sequences A—B—C and C—B—A. In both sequences, B has the same ordinal position (second). However, in the first case B follows A and in the second case B follows C; therefore, the sequence is different. Counterbalancing procedures are frequently used to control for problems related to carryover effects, order effects, and sequence effects.

Counterbalancing means that participants are exposed to the different treatment conditions in different orders. Ideally, one or more participants could be exposed to every possible order of treatment conditions. This is not possible with more than a few treatment conditions, as the number of possible orders increases rapidly (number of possible orders equals $n!$, such that three treatment conditions produces six possible orders: $3 \times 2 \times 1$). When the number of possible orders is too great, only a subset may be used. One method of obtaining a subset of the possible orders of treatment conditions is to simply randomly select a subset of the possible orders. This may pose a problem, however, because if the order or the sequence of treatment conditions has an effect on the dependent variable, random selection of treatment orders will not control for these effects.

Another method that not only controls for the order effect of treatment conditions but also allows for the order effect to be analyzed separately is the Latin square design. The Latin square design has as many orders represented as there are treatment conditions. Therefore, if there are four treatment conditions (Rx = treatment), four orders (indicated by uppercase letters A, B, C, D) would be represented as depicted in Table 13.5. Next, a random selection procedure is invoked to determine which of the four treatments corresponds to which letter (i.e., the order treatments will be administered). For example, if it was randomly determined that A = Treatment 3, B = Treatment 1, C = Treatment 4, and D = Treatment 2, then the matrix is designed as in Table 13.5.

Table 13.5 Simple Latin Square Design

A = Rx 3	B = Rx 1	C = Rx 4	D = Rx 2
B = Rx 1	C = Rx 4	D = Rx 2	A = Rx 3
C = Rx 4	D = Rx 2	A = Rx 3	B = Rx 1
D = Rx 2	A = Rx 3	B = Rx 1	C = Rx 4

Table 13.6 Balanced Latin Square Design

A = Rx 3	B = Rx 1	C = Rx 4	D = Rx 2
B = Rx 1	D = Rx 2	A = Rx 3	C = Rx 4
C = Rx 4	A = Rx 3	D = Rx 2	B = Rx 1
D = Rx 2	C = Rx 4	B = Rx 1	A = Rx 3

The Latin square design ensures that every treatment condition appears in every possible order (first, second, third, etc.). Therefore, if the order in which participants experience the different treatment conditions has an effect on the dependent variable, the Latin square design will cancel out that effect. The Latin square design also allows for the comparison of the different treatment orders (by including order as another independent variable) to determine whether there is a significant difference between the different orders. As mentioned earlier, in some situations both *order* and *sequence* may have an effect on the dependent variables. The balanced Latin square design controls for the effects of order and sequence effects and allows these differences to be compared. It ensures that treatment condition appears in every possible order (first, second, third, etc.) and that each treatment condition is preceded and followed by every other treatment condition exactly once (e.g., Rx 1 is preceded once by Rx 2, Rx 3, and Rx 4), as depicted in Table 13.6.

As a result, if order or sequence of treatment conditions affects the dependent variable, the balanced Latin square design will cancel out these effects. This design also allows for the comparison of the different treatment orders and sequences (taken together, not separately) to determine whether there is a significant difference between them. In Table 13.6, the rows indicate the order in which treatments will be assigned, and the columns indicate the sequence in which treatments will be received. As in the simple Latin square design, a random selection procedure is invoked to determine which treatment is assigned to which letter.

Mixed-Subject Designs

The mixed-subject design is a design in which one or more independent variables are between-subject variables (e.g., different participants randomly assigned to different levels of the independent variables) and one or more independent variables are within-subject variables (e.g., all participants receive all levels of the independent variables). For example, in the experiment involving teaching technique and mode of presentation, every participant would receive Technique A and Technique B (teaching technique is a within-subject variable), and half of the subjects would be taught by a computer and half would be taught by a teacher (mode of presentation is a between-subject variable). All of the issues discussed concerning between-subject designs and within-subject designs apply to mixed-subject designs.

Extraneous Variables

A major advantage of true experimental designs is that they are the only method that allow causal relationships among variables to be proven, which includes ruling out outside or extraneous influences. Thus, it is important to use a design in which

nothing else differs between the experimental groups except for the experimental conditions (levels of the independent variable). Any other differences besides the treatment conditions (called extraneous or nuisance variables) can call into question the causal inference.

Extraneous variables can sometimes affect all treatment conditions equally, either weakening or strengthening their effects. For example, in the study comparing teaching technique and mode of presentation for schoolchildren, if the treatment conditions were administered immediately following recess, it is possible that participants would be too wound up to pay attention, thus weakening the effects of all of the treatment conditions. When an extraneous variable varies systematically with the different treatment conditions, it is referred to as a *confounding variable* and poses an even greater danger to the interpretation of the results. When a confounding variable is present, any changes in the dependent variable cannot be attributed to the treatment condition with absolute certainty. Using our teacher study example, if everyone receiving Technique A via teacher had one teacher and everyone receiving Technique B via teacher had a different teacher, it could not be determined whether the different technique or the different teacher was responsible for any differences found in the dependent variable. Thus, it is important to keep everything constrained equally across the different treatment levels except for the treatment itself. This is not always possible, especially when conducting research in school and clinic settings. It does not necessarily invalidate the study to have extraneous or confounding variables present; however, it does limit the degree of causality that you can attribute to your treatment conditions.

Correlational Designs

In observational designs, the researcher simply measures variables as they occur naturally in the environment, which limits the degree to which causality can be inferred. Many observational studies, however, are conducted to examine whether a relationship exists between two or more variables or to predict certain variables (criterion variables) from other variables (predictor variables). Observing behavior and recording predictor and criterion variables is another method used in observational studies. For example, a researcher could observe children in the classroom and record the number of times they raise their hands to answer questions and the frequency of praise they receive from the teacher. Alternatively, observational studies may rely on self-report methods such as teacher or parent questionnaires that measure the variables of interest. Observational studies may also be conducted by obtaining ratings about children by people who know them, such as parents, teachers, or peers, or use combinations of these techniques.

Longitudinal and Cross-Sectional Designs

Overview

Longitudinal designs involve measuring participants repeatedly over an extended period of time, perhaps years or even decades. Cross-sectional designs and longitudinal designs are ways of obtaining different types of information. Cross-sectional

designs are best used for answering questions concerning how a treatment works at one point in time. Longitudinal designs are best used for answering questions concerning developmental change. For this reason, they can be especially informative when studying children because of their rapid developmental changes. Cross-sectional designs may be set up to assess different age groups simultaneously (e.g., children at ages 2, 6, 10, 14, 18), which would answer questions concerning possible age differences. However, it is possible that differences may exist between children of different age groups at a single point in time because they have different histories (cohort effects). Therefore, the children who are 2 years of age at the same point in time as other children who are 14 years of age might not display the same characteristics when they become 14 years of age. Longitudinal designs involve examining the long-term effects of some event or intervention. An additional advantage to longitudinal designs is that because information is collected with measures repeated at multiple time periods, error variance is reduced, often allowing for the detection of small behavior changes. Additionally, participants are compared to themselves at different points in time.

Potential Limitations

Disadvantages of longitudinal designs are mostly tied to the length of time required and the inherent cost associated with the study. Because the data are collected over long time periods, procedures and measures may become outdated and new procedures and measures may be developed. This leads to a major quandary: Should the outdated procedures and/or measures be continued so that differences may be compared across time, or should the new procedures and/or measures be adopted because they are better? One possible solution is to continue using the old procedures and/or measures to ensure accurate comparison and to adopt new procedures and/or measures as they are developed. Participant attrition is also a major problem with longitudinal designs because it is quite possible that the group that remains at the end of the study is not representative of those who dropped out along the way. Additionally, participant attrition increases the probability of a Type II error (failing to find a treatment effect when one exists) and decreases the generality of the results. Another problem is the potential confound between the effects of personal age and the effects of historical period (e.g., children growing up during a particular decade may be exposed to environmental or other events such as war that younger cohorts are not exposed to).

PARTICIPANT DEMOGRAPHICS AND SAMPLING

Identifying a Population of Interest

After formulating a research question and establishing a basic design, the next step is to consider how best to define and describe the population of interest. Review recent articles in credible journals to gain an understanding of how others have accomplished this task. For example, most research studies dealing with children contain basic sociodemographic information, such as children's age and grade (mean and *SD*), estimated level of intelligence, family socioeconomic status, sex, and ethnicity.

This information is included for other participants (e.g., parents) when relevant to the study (e.g., when studying parental attitudes toward medication compliance in children).

Detailed information is also included relevant to identifying, categorizing, or describing the population of interest. This may involve (a) a clear description of how a diagnosis is ascertained (e.g., using structured or semi-structured clinical interviews combined with using rating scale cutoff scores), (b) how a group is identified (e.g., a learning disability discrepancy formula), or (c) describing the characteristics of a select group of children (e.g., children attending a special education classroom). The use of diagnostic monikers (e.g., ADHD, Social Phobia) for identifying research samples is conventionally based on instruments with acceptable psychometric properties. This approach usually entails the administration of a structured or semi-structured clinical interview coupled with parent and teacher rating scales. Inclusion criteria for a particular diagnostic group is typically based on meeting *DSM-IV* criteria for the disorder in addition to exceeding an identified cutoff score, such as 2 standard deviations above the mean or the scale's identified range for clinical diagnosis. Exclusion criteria vary according to the research question posed. For example, a researcher studying children with learning disabilities may wish to exclude children with ADHD to determine how well an innovative intervention works with reading disabled children independent of ADHD. Conversely, a different research team investigating whether working memory deficits are unique to language-impaired children as opposed to a more generalized deficiency associated with psychiatric disability (e.g., common to many childhood disorders) may wish to include children with and without comorbid ADHD.

Alternative methods for identifying research samples are available, depending on the nature of the investigation. For example, children with poorly developed social skills who do not meet diagnostic criteria for a clinical disorder may nevertheless benefit from a social skills training program. Specific measures of social skill deficits, such as parent-teacher ratings and even direct observations, can be used if they reliably identify the sample.

Sampling

As a general guideline, researchers must sufficiently describe characteristics of their sample that might affect the generalizability of findings as discussed earlier and determine how best to obtain the research sample from the targeted population. This is particularly important for group designs but also applies to single-subject designs.

There are two classes of samples, *probability* and *nonprobability* samples. Probability samples are samples in which every member of the population has a known probability of being selected for inclusion, whereas nonprobability samples are samples in which the probability of being selected for inclusion is unknown.

Four well-known types of probability samples are *simple random samples, stratified random samples, systematic samples,* and *cluster samples.* A simple random sample is one in which every member of the population has an equal probability of being selected, whereas a stratified random sample is a random sample in which the proportion of certain characteristics in the population (e.g., race, ethnicity, culture,

sex, age, education, income, SES) are matched in the sample. For example, if sex is considered important to a study and the population under study consists of 58% boys and 42% girls, the sample reflects these same proportions. Samples may also be stratified on several different characteristics simultaneously. For the sample to be a stratified random sample, sampling would be random within each characteristic considered important to the study.

A systematic sample is a sample selected in a nonrandom fashion. For example, if 10% of the child population is to be included in the sample, then every 10th child would be chosen for inclusion.

A cluster sample is a sample in which clusters (groups) are randomly selected rather than individuals. For example, if researchers are interested in sampling grade school children, they would start with a list of grade schools and randomly select which schools to include. Within each school, you could include all of the students, randomly select students for inclusion, or randomly select additional clusters (e.g., grade levels).

The most commonly used types of nonprobability samples include the *convenience sample,* the *stratified convenience sample,* and the *snowball sample*. A convenience sample is a sample of participants that is convenient for the researcher to obtain (e.g., college undergraduates, hospital patients, child referrals to a university-based specialty clinic). A stratified convenience sample is the same as a stratified random sample, except that the participants are selected for convenience rather than randomly. A snowball sample is a sample that is created by having the initial participants (e.g., children's parents) suggest additional possible participants; these additional participants suggest additional possible participants, and so on. Because researchers do not typically have access to the entire population they wish to study, most studies use nonprobability samples.

Sample Size and Power

An important issue concerning samples is determining how many participants should be included. The main issue concerning sample size is that you want to recruit a sufficient number of participants to have a powerful test, but not more than you need, as this can be costly and time consuming. *Power* refers to the probability of finding a significant treatment effect when one truly exists (probability of rejecting a false null hypothesis). Power can be increased by setting alpha equal to 0.05 rather than 0.01, increasing the size of the treatment effect (increasing the between-condition or group variation), or reducing the error (reducing the within-condition or group variation). One method of reducing the within-condition variation (error) is to increase the number of participants. As sample size increases, within-group variation (error) decreases.

Convention generally holds that .80 is the minimum acceptable level of power. This means that if there truly is a treatment effect, your statistical test has an 80% probability of finding that treatment effect (rejecting the null hypothesis). Before you conduct an experiment, you should do a power analysis to determine how many subjects are needed to achieve (at least) 80% power. Conducting a power analysis requires the researcher to make several educated guesses concerning

the data (e.g., size of treatment effect and the population standard deviation). Specialty texts (see Cohen, 1988) and software (GPower: www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/) for computing a power analysis are readily available. Researchers recognize that a certain degree of controversy exists concerning hypothesis testing, because with a large enough sample, even trivial treatment effects may be statistically significant. For this reason, researchers should specify the minimum interesting treatment effect (i.e., the minimum effect that would be of interest) to be found with 80% power before conducting a power analysis.

SUMMARY

Conducting research with children is a multifaceted enterprise that requires broad knowledge of research methods, research design, psychometric theory, statistics, and child development, coupled with a healthy dose of curiosity and tenacity. Textbooks and coursework are useful resources for novice researchers, but there is no substitution for working in an active research laboratory under the guidance of an accomplished mentor.

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioural and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213–232.

Angold, A., & Costello, E. J. (2000). The Child and Adolescent Psychiatric Assessment (CAPA). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(1), 39–48.

Burns, L. G., & Haynes, S. N. (2006). Clinical psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid & E. S Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 401–418). Washington, DC: American Psychological Association.

Campbell, S. B. (1990). *Behavioral problems in preschoolers: Clinical and developmental issues*. New York: Guilford Press.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7–18.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Erlbaum.

Dishion, T. J., & Patterson, G. R. (2006). The development and ecology of antisocial behavior in children and adolescents. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental psychopathology: Vol. 3. Risk, disorder, and adaptation* (2nd ed., pp. 503–431). Hoboken, NJ: Wiley.

Duhig, A. M., Renk, K., Epstein, M. K. & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7*(4), 435–453.

Fergusson, D. M., & Horwood, L. J. (1995). Early disruptive behavior, IQ, and later school achievement and delinquent behavior. *Journal of Abnormal Child Psychology, 23*(2), 183–199.

Filipek, P. A., Accardo, P. J., & Baranek, G. T. (1999). The screening and diagnosis of autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 29*(6), 439–484.

Harris, F. C., & Lahey, B. B. (1982). Recording system bias in direct observational methodology: A review of critical analysis of factors causing inaccurate coding behavior. *Clinical Psychological Review, 2*(4), 539–556.

Haviland-Jones, J., Gebelt, J. L., & Stapley, J. C. (1997). The questions of development in emotion. In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 233–256). New York: Basic Books.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499.

Hersen, M. (2006). *Clinician's handbook of child behavioral assessment*. San Diego, CA: Academic Press.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kazdin, A. E. (1988). *Clinical psychotherapy: Developing and identifying effective treatments*. New York: Pergamon Press.

Kazdin, A. E. (1999). Current (lack of) status of theory in child and adolescent psychotherapy research. *Journal of Clinical Child Psychology, 28*(4), 533–543.

Kazdin, A. E. (2004). Clinical significance: Measuring whether interventions make a difference. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (3rd ed., pp. 691–710). Washington, DC: American Psychological Association.

Kofler, M. J., Rapport, M. D., & Alderson, R. M. (in press). Classroom observation of ADHD and comparison children: A meta-analytic review. *Journal of Child Psychology and Psychiatry*.

Lapouse, R., & Monk, M. A. (1959). Fears and worries in a representative sample of children. *American Journal of Orthopsychiatry, 29*, 803–818.

MacFarlane, J. W., Allen, L., & Honzik, M. P. (1954). *A developmental study of the behavioral problems of normal children between 21 months and 14 years*. Berkeley: University of California Press.

Mash, E. J., & Terdal, L. G. (1997). *Assessment of childhood disorders* (3rd ed.) New York: Guilford Press.

McClellan, J. M., & Werry, J. S. (2000). Research psychiatric diagnostic interviews for children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(1), 19–27.

Olds, R. S., & Symons, C. W. (1990). Recommendations for obtaining cooperation to conduct school-based research. *Journal of School Health, 60*(3), 96–98.

Ollendick, T. H., & Hersen, M. (1998). *Handbook of child psychopathology* (3rd ed.) New York: Plenum Press.

Orvaschel, H. (2006). Structured and semistructured interviews. In M. S Hersen (Ed.), *Clinician's handbook of child behavioral assessment* (pp. 159-179). San Diego, CA: Academic Press.

Palfrey, J. S., Levine, M. D., Walker, D. K. & Sullivan, M. (1985). The emergence of attention deficits in early childhood: A prospective study. *Behavioral Pediatrics, 6*(6), 339–348.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention-deficit/hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 449–476.

Petosa, R., & Goodman, R. M. (1991). Recruitment and retention of schools participating in school health research. *Journal of School Health, 61*(10), 426–429.

Plake, B. S. Impara, J. C. & Spies, R. A. (Eds.). (2003). *The fifteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Porrino, L. J., Rapoport, J. L., Behar, D., Sceery, W., Ismond, D. R., & Bunney, W. E. (1983). A naturalistic assessment of the motor activity of hyperactive boys: Pt. I. Comparison with normal controls. *Archives of General Psychiatry, 40*(6), 681–687.

Rapoport, J., Abramson, A., Alexander, D., & Lott, I. (1971). Playroom observations of hyperactive children on medication. *Journal of the American Academy of Child and Adolescent Psychiatry, 10*(3), 524–534.

Rapport, M. D., Chung, K., Shore, G., & Isaacs, P. (2001). A conceptual model of child psychopathology: Implications for understanding attention deficit hyperactivity disorder and treatment efficacy. *Journal of Clinical Child Psychology, 30*(1), 48–58.

Rapport, M. D., Denney, C., DuPaul, G. J., & Gardner, M. J. (1994). Attention deficit disorder and methylphenidate: Normalization rates, clinical effectiveness, and response prediction in 76 children. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*(6), 882–893.

Rapport, M. D., Kofler, M., Alderson, M., & Raiker, J.(in press). Attention-deficit/hyperactivity disorder. In M. Hersen, & D. Reitman (Eds.), *Handbook of psychological assessment, case conceptualization and treatment: Vol. 2 Children and adolescents*. Hoboken, NJ: Wiley.

Rapport, M. D., Scanlan, S. W., & Denney, C. B. (1999). Attention-deficit/hyperactivity disorder and scholastic achievement: A model of dual developmental pathways. *Journal of Child Psychology and Psychiatry, 40*(8), 1169–1183.

Rapport, M. D., Timko, T. M., & Wolfe, R. (2006). Attention-deficit/hyperactivity disorder. In M. Hersen (Ed.), *Clinician's handbook of child behavioral assessment* (pp. 401–435). San Diego, CA: Academic Press.

Reid, J. M., Webster-Stratton, C., & Hammond, M. (2003). Follow-up of children who received the Incredible Years intervention for oppositional-defiant disorder: Maintenance and prediction of 2-year outcome. *Behavior Therapy, 34*(4), 471–491.

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60*(3), 402–408.

Stein, B. D., Jayxoc, L. H., & Kataoka, S. H. (2003). A mental health intervention for schoolchildren exposed to violence: A randomized controlled trail. *Journal of the American Medical Association, 290*(5), 603–611.

Stevens, T. M., Kupst, M. J., Suran, B. G., & Schulman, J. L. (1978). Activity level: A comparison between actometer scores and observer ratings. *Journal of Abnormal Child Psychology, 6*(2), 163–173.

Tryon, W. W., & Pinto, L. P. (1994). Comparing activity measurement and ratings. *Behavior Modification, 18*(3), 251–261.

Walker, K. C., & Bracken, B. A. (1996). Inter-parent agreement on four preschool behavior rating scales: Effects of parent and child gender. *Psychology in the Schools, 33*(4), 273–281.

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117*(3), 450–468.