Single-Case Experimental Designs:
A Systematic Review of Published Research and Current Standards

Justin D. Smith
University of Oregon

This article systematically reviews the research design and methodological characteristics of single-case experimental design (SCED) research published in peer-reviewed journals between 2000 and 2010. SCEDs provide researchers with a flexible and viable alternative to group designs with large sample sizes. However, methodological challenges have precluded widespread implementation and acceptance of the SCED as a viable complementary methodology to the predominant group design. This article includes a description of the research design, measurement, and analysis domains distinctive to the SCED; a discussion of the results within the framework of contemporary standards and guidelines in the field; and a presentation of updated benchmarks for key characteristics (e.g., baseline sampling, method of analysis), and overall, it provides researchers and reviewers with a resource for conducting and evaluating SCED research. The results of the systematic review of 409 studies suggest that recently published SCED research is largely in accordance with contemporary criteria for experimental quality. Analytic method emerged as an area of discord. Comparison of the findings of this review with historical estimates of the use of statistical analysis indicates an upward trend, but visual analysis remains the most common analytic method and also garners the most support among those entities providing SCED standards. Although consensus exists along key dimensions of single-case research design, and researchers appear to be practicing within these parameters, there remains a need for further evaluation of assessment and sampling techniques and data analytic methods.

Keywords: daily diary, research standards, single-case experimental design, systematic review, time-series

The single-case experiment has a storied history in psychology dating back to the field’s founders: Fechner (1889); Watson (1925), and Skinner (1938). It has been used to inform and develop theory, examine interpersonal processes, study the behavior of organisms, establish the effectiveness of psychological interventions, and address a host of other research questions (for a review, see Morgan & Morgan, 2001). In recent years the single-case experimental design (SCED) has been represented in the literature more often than in past decades, as is evidenced by recent reviews (Hammond & Gast, 2010; Shadish & Sullivan, 2011), but it still languishes behind the more prominent group design in nearly all subfields of psychology. Group designs are often professed to be superior because they minimize, although do not necessarily eliminate, the major internal validity threats to drawing scientifically valid inferences from the results (Shadish, Cook, & Campbell, 2002). SCEDs provide a rigorous, methodologically sound alternative method of evaluation (e.g., Barlow, Nock, & Hersen, 2008; Horner et al., 2005; Kazdin, 2010; Kratochwill & Levin, 2010; Shadish et al., 2002) but are often overlooked as a true experimental methodology capable of eliciting legitimate inferences (e.g., Barlow et al., 2008; Kazdin, 2010). Despite a shift in the zeitgeist from single-case experiments to group designs more than a half century ago, recent and rapid methodological advancements suggest that SCEDs are poised for resurgence.

Basics of the SCED

Single case refers to the participant or cluster of participants (e.g., a classroom, hospital, or neighborhood) under investigation. In contrast to an experimental group design, in which one group is compared with another, participants in a single-subject experiment research provide their own control data for the purpose of comparison in a within-subject, rather than a between-subjects, design. SCEDs typically involve a comparison between two experimental time periods, known as phases. This approach typically includes collecting a representative baseline phase to serve as a comparison with subsequent phases. In studies examining single subjects that are actually groups (i.e., classroom, school), there are additional threats to internal validity of the results, as noted by Kratochwill and Levin (2010), which include setting or site effects.

The central goal of the SCED is to determine whether a causal or functional relationship exists between a researcher-manipulated independent variable (IV) and a meaningful change in the dependent variable (DV). SCEDs generally involve repeated, systematic assessment of one or more IVs and DVs over time. The DV is
measured repeatedly across and within all conditions or phases of the IV. Experimental control in SCEDs includes replication of the effect either within or between participants (Horner et al., 2005). Randomization is another way in which threats to internal validity can be experimentally controlled. Kratochwill and Levin (2010) recently provided multiple suggestions for adding a randomization component to SCEDs to improve the methodological rigor and internal validity of the findings.

Examination of the effectiveness of interventions is perhaps the area in which SCEDs are most well represented (Morgan & Morgan, 2001). Researchers in behavioral medicine and in clinical, health, educational, school, sport, rehabilitation, and counseling psychology often use SCEDs because they are particularly well suited to examining the processes and outcomes of psychological and behavioral interventions (e.g., Borckardt et al., 2008; Kazdin, 2010; Robey, Schultz, Crawford, & Sinner, 1999). Skepticism about the clinical utility of the randomized controlled trial (e.g., Jacobson & Christensen, 1996; Wachtel, 2010; Westen & Bradley, 2005; Westen, Novotny, & Thompson-Brenner, 2004) has renewed researchers’ interest in SCEDs as a means to assess intervention outcomes (e.g., Borckardt et al., 2008; Dattilio, Edwards, & Fishman, 2010; Horner et al., 2005; Kratochwill, 2007; Kratochwill & Levin, 2010). Although SCEDs are relatively well represented in the intervention literature, it is by no means their sole home: Examples appear in nearly every subfield of psychology (e.g., Bolger, Davis, & Rafaeli, 2003; Pineccki, Hufford, Solham, & Trull, 2007; Reis & Gable, 2000; Shiffman, Stone, & Hufford, 2008; Solida, Moore, & Lande, 2002). Aside from the current preference for group-based research designs, several methodological challenges have repressed the proliferation of the SCED.

Methodological Complexity

SCEDs undeniably present researchers with a complex array of methodological and research design challenges, such as establishing a representative baseline, managing the nonindependence of sequential observations (i.e., autocorrelation, serial dependence), interpreting single-subject effect sizes, analyzing the short data streams seen in many applications, and appropriately addressing the matter of missing observations. In the field of intervention research for example, Hser, Shen, Chou, Messer, and Anglin (2001) noted that studies using SCEDs are “rare” because of the minimum number of observations that are necessary (e.g., three to five data points in each phase) and the complexity of available data analysis approaches. Advances in longitudinal person-based trajectory analysis (e.g., Nagin, 1999), structural equation modeling techniques (e.g., Lubke & Muthén, 2005), time-series forecasting (e.g., autoregressive integrated moving averages; Box & Jenkins, 1970), and statistical programs designed specifically for SCEDs (e.g., simulation modeling analysis; Borckardt, 2006) have provided researchers with robust means of analysis, but they might not be feasible methods for the average psychological scientist.

Application of the SCED has also expanded. Today, researchers use variants of the SCED to examine complex psychological processes and the relationship between daily and momentary events in peoples’ lives and their psychological correlates. Research in nearly all subfields of psychology has begun to use daily diary and ecological momentary assessment (EMA) methods in the context of the SCED, opening the door to understanding increas-ingly complex psychological phenomena (see Bolger et al., 2003; Shiffman et al., 2008). In contrast to the carefully controlled laboratory experiment that dominated research in the first half of the twentieth century (e.g., Skinner, 1938; Watson, 1925), contemporary proponents advocate application of the SCED in naturalistic studies to increase the ecological validity of empirical findings (e.g., Bloom, Fisher, & Orme, 2003; Borckardt et al., 2008; Dattilio et al., 2010; Jacobson & Christensen, 1996; Kazdin, 2008; Morgan & Morgan, 2001; Westen & Bradley, 2005; Westen et al., 2004). Recent advancements and expanded application of SCEDs indicate a need for updated design and reporting standards.

This Review

Many current benchmarks in the literature concerning key parameters of the SCED were established well before current advancements and innovations, such as the suggested minimum number of data points in the baseline phase(s), which remains a disputed area of SCED research (e.g., Center, Skiba, & Casey, 1986; Huijtema, 1985; R. R. Jones, Vaught, & Weinrott, 1977; Sharpley, 1987). This article comprises (a) an examination of contemporary SCED methodological and reporting standards; (b) a systematic review of select design, measurement, and statistical characteristics of published SCED research during the past decade; and (c) a broad discussion of the critical aspects of this research to inform methodological improvements and study reporting standards. The reader will garner a fundamental understanding of what constitutes appropriate methodological soundness in single-case experimental research according to the established standards in the field, which can be used to guide the design of future studies, improve the presentation of publishable empirical findings, and inform the peer-review process. The discussion begins with the basic characteristics of the SCED, including an introduction to time-series, daily diary, and EMA strategies, and describes how current reporting and design standards apply to each of these areas of single-case research. Interwoven within this presentation are the results of a systematic review of SCED research published between 2000 and 2010 in peer-reviewed outlets and a discussion of the way in which these findings support, or differ from, existing design and reporting standards and published SCED benchmarks.

Method

Review of Current SCED Guidelines and Reporting Standards

In contrast to experimental group comparison studies, which conform to generally well agreed upon methodological design and reporting guidelines, such as the CONSORT (Moher, Schulz, Altman, & the CONSORT Group, 2001) and TREND (Des Jarlais, Lyles, & Crepaz, 2004) statements for randomized and nonrandomized trials, respectively, there is comparatively much less consensus when it comes to the SCED. Until fairly recently, design and reporting guidelines for single-case experiments were almost entirely absent in the literature and were typically determined by the preferences of a research subspecialty or a particular journal’s editorial board. Facations still exist within the larger field of psychology, as can be seen in the collection of standards presented in this article, particularly in regard to data analytic methods of
SCEDs, but fortunately there is budding agreement about certain design and measurement characteristics. A number of task forces, professional groups, and independent experts in the field have recently put forth guidelines; each has a relatively distinct purpose, which likely accounts for some of the discrepancies between them. In what is to be a central theme of this article, researchers are ultimately responsible for thoughtfully and synergistically combining research design, measurement, and analysis aspects of a study.

This review presents the more prominent, comprehensive, and recently established SCED standards. Six sources are discussed: (a) Single-Case Design Technical Documentation from the What Works Clearinghouse (WWC; Kratochwill et al., 2010); (b) the American Psychological Association (APA) Division 12 Task Force on Psychological Interventions, with contributions from the Division 12 Task Force on Promotion and Dissemination of Psychological Procedures and the APA Task Force for Psychological Intervention Guidelines (DIV12; presented in Chambless & Holton, 1998; Chambless & Ollendick, 2001), adopted and expanded by APA Division 53, the Society for Clinical Child and Adolescent Psychology (Weisz & Hawley, 1998, 1999); (c) the APA Division 16 Task Force on Evidence-Based Interventions in School Psychology (DIV16; Members of the Task Force on Evidence-Based Interventions in School Psychology, 2003); (d) the National Reading Panel (NRP; National Institute of Child Health and Human Development, 2000); (e) the Single-Case Experimental Design Scale (Tate et al., 2008); and (f) the reporting guidelines for EMA put forth by Stone and Shiffman (2002). Although the specific purposes of each source differ somewhat, the overall aim is to provide researchers and reviewers with agreed-upon criteria to be used in the conduct and evaluation of SCED research. The standards provided by WWC, DIV12, DIV16, and the NRP represent the efforts of task forces. The Tate et al. (2008) scale was selected for inclusion in this review because it represents perhaps the only psychometrically validated tool for assessing the rigor of SCED methodology. Stone and Shiffman’s standards were intended specifically for EMA methods, but many of their criteria also apply to time-series, daily diary, and other repeated-measurement and sampling methods, making them pertinent to this article. The design, measurement, and analysis standards are presented in the later sections of this article and notable concurrences, discrepancies, strengths, and deficiencies are summarized.

**Systematic Review Search Procedures and Selection Criteria**

**Search strategy.** A comprehensive search strategy of SCEDs was performed to identify studies published in peer-reviewed journals meeting a priori search and inclusion criteria. First, a computer-based PsycINFO search of articles published between 2000 and 2010 (search conducted in July 2011) was conducted that used the following primary key terms and phrases that appeared anywhere in the article (asterisks denote that any characters/letters can follow the last character of the search term): alternating treatment design, changing criterion design, experimental case*, multiple baseline design, replicated single-case design, simultaneous treatment design, time-series design. The search was limited to studies published in the English language and those appearing in peer-reviewed journals within the specified publication year range. Additional limiters of the type of article were also used in PsycINFO to increase specificity: The search was limited to include methodologies indexed as either quantitative study OR treatment outcome/randomized clinical trial and NOT field study OR interview OR focus group OR literature review OR systematic review OR mathematical model OR qualitative study.

**Study selection.** The author used a three-phase study selection, screening, and coding procedure to select the highest number of applicable studies. Phase 1 consisted of the initial systematic review conducted using PsycINFO, which resulted in 571 articles. In Phase 2, titles and abstracts were screened: Articles appearing to use a SCED were retained (451) for Phase 3, in which the author and a trained research assistant read each full-text article and entered the characteristics of interest into a database. At each phase of the screening process, studies that did not use a SCED or that either self-identified as, or were determined to be, quasiexperimental were dropped. Of the 571 original studies, 82 studies were determined to be quasiexperimental. The definition of a quasiexperimental design used in the screening procedure conforms to the descriptions provided by Kazdin (2010) and Shadish et al. (2002) regarding the necessary components of an experimental design. For example, reversal designs require a minimum of four phases (e.g., ABAB), and multiple baseline designs must demonstrate replication of the effect across at least three conditions (e.g., subjects, settings, behaviors). Sixteen studies were unavailable in full text in English, and five could not be obtained in full text and were thus dropped. The remaining articles that were not retained for review (59) were determined not to be SCED studies meeting our inclusion criteria but had been identified in our PsycINFO search using the specified keyword and methodology terms. For this review, 409 studies were selected. The sources of the 409 reviewed studies are summarized in Table 1. A complete bibliography of the 571 studies appearing in the initial search, with the included studies marked, is available in the Appendix or from the author.

**Coding criteria amplifications.** A comprehensive description of the coding criteria for each category in this review is available from the author by request. The primary coding criteria are described here and in later sections of this article.

- Research design was classified into one of the types discussed later in the section titled Predominant Single-Case Experimental Designs on the basis of the authors’ stated design type. Secondary research designs were then coded when applicable (i.e., mixed designs). Distinctions between primary and secondary research designs were made based on the authors’ description of their study. For example, if an author described the study as a “multiple baseline design with time-series measurement,” the primary research design would be coded as being multiple baseline, and time-series would be coded as the secondary research design.
- Observer ratings were coded as present when observational coding procedures were described and/or the results of a test of interobserver agreement were reported.
- Interrater reliability for observer ratings was coded as present in any case in which percent agreement, alpha, kappa, or another appropriate statistic was reported, regardless of the amount of the total data that were examined for agreement.
- Daily diary, daily self-report, and EMA codes were given when authors explicitly described these procedures in the text by
Table 1
Journal Sources of Studies Included in the Systematic Review (N = 409)

<table>
<thead>
<tr>
<th>Journal title</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Applied Behavior Analysis</td>
<td>45</td>
</tr>
<tr>
<td>Behavioral Interventions</td>
<td>15</td>
</tr>
<tr>
<td>Journal of Positive Behavior Interventions</td>
<td>14</td>
</tr>
<tr>
<td>Research in Autism Spectrum Disorders</td>
<td>14</td>
</tr>
<tr>
<td>Journal of Autism and Developmental Disorders</td>
<td>13</td>
</tr>
<tr>
<td>Education &amp; Treatment of Children</td>
<td>12</td>
</tr>
<tr>
<td>Focus on Autism and Other Developmental Disabilities</td>
<td>12</td>
</tr>
<tr>
<td>Journal of Developmental and Physical Disabilities</td>
<td>10</td>
</tr>
<tr>
<td>Journal of Early Intervention</td>
<td>10</td>
</tr>
<tr>
<td>Aphasiology</td>
<td>9</td>
</tr>
<tr>
<td>Journal of Behavioral Education</td>
<td>9</td>
</tr>
<tr>
<td>Journal of Organizational Behavior Management</td>
<td>9</td>
</tr>
<tr>
<td>Research in Developmental Disabilities</td>
<td>9</td>
</tr>
<tr>
<td>Education and Training in Developmental Disabilities</td>
<td>8</td>
</tr>
<tr>
<td>Behavioral Disorders</td>
<td>8</td>
</tr>
<tr>
<td>Analysis of Verbal Behavior</td>
<td>8</td>
</tr>
<tr>
<td>Child &amp; Family Behavior Therapy</td>
<td>8</td>
</tr>
<tr>
<td>Behavior Modification</td>
<td>6</td>
</tr>
<tr>
<td>Psychology in the Schools</td>
<td>6</td>
</tr>
<tr>
<td>Journal of Emotional and Behavioral Disorders</td>
<td>5</td>
</tr>
<tr>
<td>School Psychology Review</td>
<td>5</td>
</tr>
<tr>
<td>Behavior Therapy</td>
<td>4</td>
</tr>
<tr>
<td>Journal of Applied School Psychology</td>
<td>4</td>
</tr>
<tr>
<td>Topics in Early Childhood Special Education</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. Each of the following journal titles contributed one study unless otherwise noted in parentheses: Augmentative and Alternative Communication; Acta Colombiana de Psicología; Acta Comportamentalka; Adapted Physical Activity Quarterly (2); Addiction Research and Theory; Advances in Speech Language Pathology; American Annals of the Deaf; American Journal of Education; American Journal of Occupational Therapy; American Journal of Speech-Language Pathology; The American Journal on Addictions; American Journal on Mental Retardation; Applied Ergonomics; Applied Psychophysiology and Biofeedback; Australian Journal of Guidance & Counseling; Australian Psychologist; Autism; The Behavior Analyst; The Behavior Analyst Today; Behavior Analysis in Practice (2); Behavior and Social Issues (2); Behaviour Change (2); Behavioural and Cognitive Psychotherapy; Behaviour Research and Therapy (3); Brain and Language (2); Brain Injury (2); Canadian Journal of Occupational Therapy (2); Canadian Journal of School Psychology; Career Development for Exceptional Individuals; Chinese Mental Health Journal; Clinical Linguistics and Phonetics; Clinical Psychology & Psychotherapy; Cognitive and Behavioral Practice; Cognitive Computation; Cognitive Therapy and Research; Communication Disorders Quarterly; Developmental Medicine & Child Neurology (2); Developmental Neurorehabilitation (2); Disability and Rehabilitation: An International, Multidisciplinary Journal (3); Disability and Rehabilitation: Assistive Technology; Down Syndrome: Research & Practice; Drug and Alcohol Dependence (2); Early Childhood Education Journal (2); Early Childhood Services: An Interdisciplinary Journal of Effectiveness; Educational Psychology (2); Education and Training in Autism and Developmental Disabilities; Electronic Journal of Research in Educational Psychology; Environment and Behavior (2); European Eating Disorders Review; European Journal of Sport Science; European Review of Applied Psychology; Exceptional Children; Exceptionality; Experimental and Clinical Psychopharmacology; Family & Community Health; The Journal of Health Promotion & Maintenance: Headache; The Journal of Head and Face Pain; International Journal of Behavioral Consultation and Therapy (2); International Journal of Disability; Development and Education (2); International Journal of Drug Policy; International Journal of Psychology; International Journal of Speech-Language Pathology; International Psychogeriatrics; Japanese Journal of Behavior Analysis (3); Japanese Journal of Special Education; Journal of Applied Research in Intellectual Disabilities (2); Journal of Applied Sport Psychology (3); Journal of Attention Disorders (2); Journal of Behavior Therapy and Experimental Psychiatry; Journal of Child Psychology and Psychiatry; Journal of Clinical Psychology in Medical Settings; Journal of Clinical Sport Psychology; Journal of Cognitive Psychotherapy; Journal of Consulting and Clinical Psychology (2); Journal of Deaf Studies and Deaf Education; Journal of Educational & Psychological Consultation (2); Journal of Evidence-Based Practices for Schools (2); Journal of the Experimental Analysis of Behavior (2); Journal of General Internal Medicine; Journal of Intellectual and Developmental Disabilities; Journal of Intellectual Disability Research (2); Journal of Medical Speech-Language Pathology; Journal of Neurology, Neurosurgery & Psychiatry; Journal of Paediatrics and Child Health; Journal of Prevention and Intervention in the Community; Journal of Safety Research; Journal of School Psychology (3); The Journal of Socio-Economics; The Journal of Special Education; Journal of Speech, Language, and Hearing Research (2); Journal of Sport Behavior; Journal of Substance Abuse Treatment; Journal of the International Neuropsychological Society; Journal of Traumatic Stress; The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences; Language, Speech, and Hearing Services in Schools; Learning Disabilities Research & Practice (2); Learning Disability Quarterly (2); Music: Therapy Perspectives; Neurorehabilitation and Neural Repair; Neuropsychological Rehabilitation (2); Pain; Physical Education and Sport Pedagogy (2); Preventive Medicine; An International Journal Devoted to Practice and Theory; Psychological Assessment; Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences; The Psychological Record; Reading and Writing; Remedial and Special Education (3); Research and Practice for Persons with Severe Disabilities (2); Restorative Neurology and Neuroscience; School Psychology International; Seminars in Speech and Language; Sleep and Hypnosis; School Psychology Quarterly; Social Work in Health Care; The Sport Scientist (3); Therapeutic Recreation Journal (2); The Volta Review; Work: Journal of Prevention, Assessment & Rehabilitation.
name. Coders did not infer the use of these measurement strategies.

- The number of baseline observations was either taken directly from the figures provided in text or was simply counted in graphical displays of the data when this was determined to be a reliable approach. In some cases, it was not possible to reliably determine the number of baseline data points from the graphical display of data, in which case, the “unavailable” code was assigned. Similarly, the “unavailable” code was assigned when the number of observations was either unreported or ambiguous, or only a range was provided and thus no mean could be determined. Similarly, the mean number of baseline observations was calculated for each study prior to further descriptive statistical analyses because a number of studies reported means only.

- The coding of the analytic method used in the reviewed studies is discussed later in the section titled Discussion of Review Results and Coding of Analytic Methods.

Results of the Systematic Review

Descriptive statistics of the design, measurement, and analysis characteristics of the reviewed studies are presented in Table 2. The results and their implications are discussed in the relevant sections throughout the remainder of the article.

Discussion of the Systematic Review Results in Context

The SCED is a very flexible methodology and has many variants. Those mentioned here are the building blocks from which other designs are then derived. For those readers interested in the nuances of each design, Barlow et al. (2008); Franklin, Allison, and Gorman (1997); Kazdin (2010); and Kratochwill and Levin (1992), among others, provide cogent, in-depth discussions. Identifying the appropriate SCED depends upon many factors, including the specifics of the IV, the setting in which the study will be conducted, participant characteristics, the desired or hypothesized outcomes, and the research question(s). Similarly, the researcher’s selection of measurement and analysis techniques is determined by these factors.

Predominant Single-Case Experimental Designs

Alternating/simultaneous designs (6%; primary design of the studies reviewed). Alternating and simultaneous designs involve an iterative manipulation of the IV(s) across different phases to show that changes in the DV vary systematically as a function of manipulating the IV(s). In these multielement designs, the researcher has the option to alternate the introduction of two or more IVs or present two or more IVs at the same time. In the alternating variation, the researcher is able to determine the relative impact of two different IVs on the DV, when all other conditions are held constant. Another variation of this design is to alternate IVs across various conditions that could be related to the DV (e.g., class period, interventionist). Similarly, the simultaneous design would occur when the IVs were presented at the same time within the same phase of the study.

Changing criterion design (4%). Changing criterion designs are used to demonstrate a gradual change in the DV over the course of the phase involving the active manipulation of the IV. Criteria indicating that a change has occurred happen in a stepwise manner, in which the criterion shifts as the participant responds to the presence of the manipulated IV. The changing criterion design is particularly useful in applied intervention research for a number of reasons. The IV is continuous and never withdrawn, unlike the strategy used in a reversal design. This is particularly important in situations where removal of a psychological intervention would be either detrimental or dangerous to the participant, or would be otherwise unfeasible or unethical. The multiple baseline design also does not withdraw intervention, but it requires replicating the effects of the intervention across participants, settings, or situations. A changing criterion design can be accomplished with one participant in one setting without withholding or withdrawing treatment.

Multiple baseline/combined series design (69%). The multiple baseline or combined series design can be used to test within-subject change across conditions and often involves multiple participants in a replication context. The multiple baseline design is quite simple in many ways, essentially consisting of a number of repeated, miniature AB experiments or variations thereof. Introduction of the IV is staggered temporally across multiple participants or across multiple within-subject conditions, which allows the researcher to demonstrate that changes in the DV reliably occur only when the IV is introduced, thus controlling for the effects of extraneous factors. Multiple baseline designs can be used both within and across units (i.e., persons or groups of persons). When the baseline phase of each subject begins simultaneously, it is called a concurrent multiple baseline design. In a nonconcurrent variation, baseline periods across subjects begin at different points in time. The multiple baseline design is useful in many settings in which withdrawal of the IV would not be appropriate or when introduction of the IV is hypothesized to result in permanent change that would not reverse when the IV is withdrawn. The major drawback of this design is that the IV must be initially withheld for a period of time to ensure different starting points across the different units in the baseline phase. Depending upon the nature of the research questions, withholding an IV, such as a treatment, could be potentially detrimental to participants.

Reversal designs (17%). Reversal designs are also known as introduction and withdrawal and are denoted as ABAB designs in their simplest form. As the name suggests, the reversal design involves collecting a baseline measure of the DV (the first A phase), introducing the IV (the first B phase), removing the IV while continuing to assess the DV (the second A phase), and then reintroducing the IV (the second B phase). This pattern can be repeated as many times as is necessary to demonstrate an effect or otherwise address the research question. Reversal designs are useful when the manipulation is hypothesized to result in changes in the DV that are expected to reverse or discontinue when the manipulation is not present. Maintenance of an effect is often necessary to uphold the findings of reversal designs. The demonstration of an effect is evident in reversal designs when improvement occurs during the first manipulation phase, compared with the first baseline phase, then reverts to or approaches original baseline levels during the second baseline phase when the manipulation has been withdrawn, and then improves again when the manipulation is then reinstated. This pattern of reversal, when the manipulation is introduced and then withdrawn, is essential to
<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Range</th>
<th>%</th>
<th>IRR</th>
<th>%</th>
<th>M</th>
<th>SD</th>
<th>Range</th>
<th>Method of analysis (%)&lt;sup&gt;g&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research design&lt;sup&gt;a&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternating condition</td>
<td>26</td>
<td>4.77</td>
<td>3.34</td>
<td>1–17</td>
<td>84.6</td>
<td>95.5</td>
<td>3.8</td>
<td>8.44</td>
<td>9.50</td>
<td>2–39</td>
<td>23.1</td>
</tr>
<tr>
<td>Changing/shifting criterion</td>
<td>18</td>
<td>1.94</td>
<td>1.06</td>
<td>1–4</td>
<td>77.8</td>
<td>85.7</td>
<td>0.0</td>
<td>5.29</td>
<td>2.93</td>
<td>2–10</td>
<td>27.8</td>
</tr>
<tr>
<td>Multiple baseline/combined series</td>
<td>283</td>
<td>7.29</td>
<td>18.08</td>
<td>1–200</td>
<td>75.6</td>
<td>98.1</td>
<td>7.1</td>
<td>10.40</td>
<td>8.84</td>
<td>2–89</td>
<td>21.6</td>
</tr>
<tr>
<td>Reversal</td>
<td>70</td>
<td>6.64</td>
<td>10.64</td>
<td>1–75</td>
<td>78.6</td>
<td>100.0</td>
<td>4.3</td>
<td>11.69</td>
<td>13.78</td>
<td>1–72</td>
<td>17.1</td>
</tr>
<tr>
<td>Simultaneous condition</td>
<td>2&lt;sup&gt;d&lt;/sup&gt;</td>
<td>8</td>
<td>50.0</td>
<td>100.0</td>
<td>0.0</td>
<td>2.00</td>
<td>0.0</td>
<td>50.0</td>
<td>50.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Time-series</td>
<td>10&lt;sup&gt;e&lt;/sup&gt;</td>
<td>26.78</td>
<td>35.43</td>
<td>2–114</td>
<td>50.0</td>
<td>40.0</td>
<td>10.0</td>
<td>6.21</td>
<td>2.59</td>
<td>3–10</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Mixed designs&lt;sup&gt;b&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple baseline with reversal</td>
<td>12</td>
<td>6.89</td>
<td>8.24</td>
<td>1–32</td>
<td>92.9</td>
<td>100.0</td>
<td>7.1</td>
<td>13.01</td>
<td>9.59</td>
<td>3–33</td>
<td>14.3</td>
</tr>
<tr>
<td>Multiple baseline with changing criterion</td>
<td>6</td>
<td>3.17</td>
<td>1.33</td>
<td>1–5</td>
<td>83.3</td>
<td>80.0</td>
<td>16.7</td>
<td>11.00</td>
<td>9.61</td>
<td>5–30</td>
<td>16.7</td>
</tr>
<tr>
<td>Multiple baseline with time-series</td>
<td>6</td>
<td>5.00</td>
<td>1.79</td>
<td>3–8</td>
<td>16.7</td>
<td>100.0</td>
<td>50.0</td>
<td>17.30</td>
<td>15.68</td>
<td>4–42</td>
<td>16.7</td>
</tr>
<tr>
<td>Total of reviewed studies</td>
<td>409</td>
<td>6.63</td>
<td>14.61</td>
<td>1–200</td>
<td>76.0</td>
<td>97.1</td>
<td>6.1</td>
<td>10.22</td>
<td>9.59</td>
<td>1–89</td>
<td>20.8</td>
</tr>
</tbody>
</table>

**Note.** SCED = single-case experimental design; EMA = ecological momentary assessment; IRR = interrater reliability; % = the proportion of reviewed studies that satisfied criteria for this code; for example, the percent of studies reporting observer ratings.

<sup>a</sup> The categories in the "Research design" subsection are the primary designs identified by the authors.  
<sup>b</sup> Categories in the "Mixed designs" subsection are included in the "Research design" subsection. Only the three most prevalent mixed designs are reported.  
<sup>c</sup> One study of 624 subjects was excluded from the calculation of the number of subjects because it was a significant outlier.  
<sup>d,e</sup> Similarly, one study with 500 subjects and one study with 950 subjects were excluded from the number of subject analyses for the simultaneous condition and time-series designs, respectively. This resulted in only one simultaneous condition study, which is why no standard deviation or range is reported.  
<sup>f</sup> Because of reporting inconsistencies in the reviewed articles, the mean number of baseline observations for each study was first calculated and then combined and reported in this table.  
<sup>g</sup> In contrast to the results reported in text, the findings here are based on the total number of studies and are not divided into those that reported an analysis and those that did not. Visual and statistical analyses are not applicable to most studies using changing criterion designs. However, some authors reported using visual analysis methods.
attributing changes in the DV to the IV. However, maintenance of the effects in a reversal design, in which the DV is hypothesized to reverse when the IV is withdrawn, is not incompatible (Kazdin, 2010). Maintenance is demonstrated by repeating introduction—withdrawal segments until improvement in the DV becomes permanent even when the IV is withdrawn. There is not always a need to demonstrate maintenance in all applications, nor is it always possible or desirable, but it is paramount in the learning and intervention research contexts.

Mixed designs (10%). Mixed designs include a combination of more than one SCED (e.g., a reversal design embedded within a multiple baseline) or a SCED embedded within a group design (i.e., a randomized controlled trial comparing two groups of multiple baseline experiments). Mixed designs afford the researcher even greater flexibility in designing a study to address complex psychological hypotheses but also capitalize on the strengths of the various designs. See Kazdin (2010) for a discussion of the variations and utility of mixed designs.

Related Nonexperimental Designs

Quasiexperimental designs. In contrast to the designs previously described, all of which constitute “true experiments” (Kazdin, 2010; Shadish et al., 2002), in quasiexperimental designs the conditions of a true experiment (e.g., active manipulation of the IV, replication of the effect) are approximated and are not readily under the control of the researcher. Because the focus of this article is on experimental designs, quasiexperiments are not discussed in detail; instead, the reader is referred to Kazdin (2010) and Shadish et al. (2002).

Ecological and naturalistic single-case designs. For a single-case design to be experimental, there must be active manipulation of the IV, but in some applications, such as those that might be used in social and personality psychology, the researcher might be interested in measuring naturally occurring phenomena and examining their temporal relationships. Thus, the researcher will not use a manipulation. An example of this type of research might be a study about the temporal relationship between alcohol consumption and depressed mood, which can be measured reliably using EMA methods. Psychotherapy process researchers also use this type of design to assess dyadic relationship dynamics between therapists and clients (e.g., Tschacher & Ramseyer, 2009).

Research Design Standards

Each of the reviewed standards provides some degree of direction regarding acceptable research designs. The WWC provides the most detailed and specific requirements regarding design characteristics. Those guidelines presented in Tables 3, 4, and 5 are consistent with the methodological rigor necessary to meet the WWC distinction “meets standards.” The WWC also provides less-stringent standards for a “meets standards with reservations” distinction. When minimum criteria in the design, measurement, or analysis sections of a study are not met, it is rated “does not meet standards” (Kratochwill et al., 2010). Many SCEDs are acceptable within the standards of DIV12, DIV16, NRP, and in the Tate et al. (2008) SCED scale. DIV12 specifies that replication occurs across a minimum of three successive cases, which differs from the WWC specifications, which allow for three replications within a single-subject design but does not necessarily need to be across multiple subjects. DIV16 does not require, but seems to prefer, a multiple baseline design with a between-subjects replication. Tate et al. (2008) stated that the “design allows for the examination of cause and effect relationships to demonstrate efficacy” (p. 400). Determining whether a design meets this requirement is left up to the evaluator, who might then refer to one of the other standards or another source for direction.

The Stone and Shiffman (2002) standards for EMA are concerned almost entirely with the reporting of measurement characteristics and less so with research design. One way in which these standards differ from those of other sources is in the active manipulation of the IV. Many research questions in EMA, daily diary, and time-series designs are concerned with naturally occurring phenomena, and a researcher manipulation would run counter to this aim. The EMA standards become important when selecting an appropriate measurement strategy within the SCED. In EMA applications, as is also true in some other time-series and daily diary designs, researcher manipulation occurs as a function of the sampling interval in which DVs of interest are measured according to fixed time schedules (e.g., reporting occurs at the end of each day), random time schedules (e.g., the data-collection device prompts the participant to respond at random intervals throughout the day), or on an event-based schedule (e.g., reporting occurs after a specified event takes place).

Measurement

The basic measurement requirement of the SCED is a repeated assessment of the DV across each phase of the design in order to draw valid inferences regarding the effect of the IV on the DV. In other applications, such as those used by personality and social psychology researchers to study various human phenomena (Bolger et al., 2003; Reis & Gable, 2000), sampling strategies vary widely depending on the topic area under investigation. Regardless of the research area, SCEDs are most typically concerned with within-person change and processes and involve a time-based strategy, most commonly to assess global daily averages or peak daily levels of the DV. Many sampling strategies, such as time-series, in which reporting occurs at uniform intervals or on event-based, fixed, or variable schedules, are also appropriate measurement methods and are common in psychological research (see Bolger et al., 2003).

Repeated-measurement methods permit the natural, even spontaneous, reporting of information (Reis, 1994), which reduces the biases of retrospection by minimizing the amount of time elapsed between an experience and the account of this experience (Bolger et al., 2003). Shiffman et al. (2008) aptly noted that the majority of research in the field of psychology relies heavily on retrospective assessment measures, even though retrospective reports have been found to be susceptible to state-congruent recall (e.g., Bower, 1981) and a tendency to report peak levels of the experience instead of giving credence to temporal fluctuations (Redelmeier & Kahneman, 1996; Stone, Broderick, Kaell, Deles-Paul, & Porter, 2000). Furthermore, Shiffman et al. (1997) demonstrated that subjective aggregate accounts were a poor fit to daily reported experiences, which can be attributed to reductions in measurement error resulting in increased validity and reliability of the daily reports.
### Table 3

**Research Design Standards and Guidelines**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Experimental manipulation (independent variable [IV])</strong></td>
<td>The IV (i.e., the intervention) must be systematically manipulated as determined by the researcher</td>
<td>Need a well-defined and replicable intervention for a specific disorder, problem behavior, or condition</td>
<td>Specified intervention according to the Institute of Medicine’s (1994) classification system</td>
<td>Specified intervention</td>
<td>Scale was designed to assess the quality of interventions; thus, an intervention is required</td>
<td>Manipulation in EMA is concerned with the sampling procedure of the study</td>
</tr>
<tr>
<td><strong>2. Research designs</strong></td>
<td>At least 3 attempts to demonstrate an effect at 3 different points in time or with 3 different phase repetitions</td>
<td>Many research designs are acceptable beyond those mentioned</td>
<td>The stage of the intervention program must be specified (see Rossi &amp; Freeman, 1993)</td>
<td>The design allows for the examination of cause and effect to demonstrate efficacy</td>
<td>EMA is almost entirely concerned with measurement of variables of interest; thus, the design of the study is determined solely by the research question(s)</td>
<td></td>
</tr>
<tr>
<td>Reversal (e.g., ABAB)</td>
<td>Minimum of 4 A and B phases</td>
<td>Mentioned as acceptable</td>
<td>Mentioned as acceptable Both within and between subjects Considered the strongest because replication occurs across individuals</td>
<td>Single-subject or aggregated subjects</td>
<td>Mentioned as acceptable</td>
<td></td>
</tr>
<tr>
<td>Multiple baseline/combined series</td>
<td>At least 3 baseline conditions</td>
<td>At least 3 different, successive subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternating treatment</td>
<td>At least 3 alternating treatments compared with a baseline condition or two alternating treatments compared with each other</td>
<td>Mentioned as acceptable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simultaneous treatment</td>
<td>Same as for alternating treatment designs</td>
<td>Mentioned as acceptable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changing/shifting criterion Mixed designs</td>
<td>At least 3 different criteria</td>
<td>Mentioned as acceptable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quasi-experimental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3. Baseline</strong></td>
<td>Minimum of 3 data points</td>
<td>Minimum of 3 data points</td>
<td>Minimum of 3 data points, although more observations are preferred</td>
<td>No minimum specified</td>
<td>No minimum (“sufficient sampling of behavior occurred pretreatment”)</td>
<td></td>
</tr>
<tr>
<td><strong>4. Randomization specifications provided</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* APA = American Psychological Association.
The necessity of measuring at least one DV repeatedly means that the selected assessment method, instrument, and/or construct must be sensitive to change over time and be capable of reliably and validly capturing change. Horner et al. (2005) discussed the important features of outcome measures selected for use in these types of designs. Kazdin (2010) suggested that measures be dimensional, which can more readily detect effects than categorical and binary measures. Although using an established measure or scale, such as the Outcome Questionnaire System (M. J. Lambert, Hansen, & Harmon, 2010), provides empirically validated items for assessing various outcomes, most measure-validation studies conducted on this type of instrument involve between-subjects designs, which is no guarantee that these measures are reliable and valid for assessing within-person variability. Borsboom, Mellenbergh, and van Heerden (2003) suggested that researchers adapting validated measures should consider whether the items they propose using have a factor structure within subject similar to that obtained between subjects. This is one of the reasons that SCEDs often use observational assessments from multiple sources and report the interrater reliability of the measure. Self-report measures are acceptable practice in some circles, but generally additional assessment methods or informants are necessary to uphold the highest methodological standards. The results of this review indicate that the majority of studies include observational measurement (76.0%). Within those studies, nearly all (97.1%) reported interrater reliability procedures and results. The results within each design were similar, with the exception of time-series designs, which used observer ratings in only half of the reviewed studies.

Time-series. Time-series designs are defined by repeated measurement of variables of interest over a period of time (Box & Jenkins, 1970). Time-series measurement most often occurs in uniform intervals; however, this is no longer a constraint of time-series designs (see Harvey, 2001). Although uniform interval reporting is not necessary in SCED research, repeated measures often occur at uniform intervals, such as once each day or each week, which constitutes a time-series design. The time-series design has been used in various basic science applications (Scol lon, Kim-Pietro, & Diener, 2003) across nearly all subspecialties in psychology (e.g., Bolger et al., 2003; Piasecki et al., 2007; for a review, see Reis & Gable, 2000; Soliday et al., 2002). The basic time-series formula for a two-phase (AB) data stream is presented in Equation 1. In this formula \( \alpha \) represents the step function of the data stream; \( S \) represents the change between the first and second phases, which is also the intercept in a two-phase data stream and a step function being 0 at times \( i = 1, 2, 3 \ldots n1 \) and 1 at times \( i = n1 + 1, n1 + 2, n1 + 3 \ldots n; n1 \) is the number of observations in the baseline phase; \( n \) is the total number of data points in the data stream; \( i \) represents time; and \( \epsilon_r = \rho \epsilon_{r-1} + \epsilon_r \), which indicates the relationship between the autoregressive function (\( \rho \)) and the distribution of the data in the stream.

\[
y_i = \alpha \times S(n_1) + \epsilon_r
\]

(1)

Time-series formulas become increasingly complex when seasonality and autoregressive processes are modeled in the analytic procedures, but these are rarely of concern for short time-series data streams in SCEDs. For a detailed description of other time-series design and analysis issues, see Borckardt et al. (2008); Box and Jenkins (1970); Crosbie (1993), R. R. Jones et al. (1977), and Velicer and Fava (2003).

Time-series and other repeated-measures methodologies also enable examination of temporal effects. Borckardt et al. (2008) and others have noted that time-series designs have the potential to reveal how change occurs, not simply if it occurs. This distinction is what most interested Skinner (1938), but it often falls below the purview of today’s researchers in favor of group designs, which Skinner felt obscured the process of change. In intervention and psychopathology research, time-series designs can assess mediators of change (Doss & Atkins, 2006), treatment processes (Stout, 2007; Tscherger & Ramseyer, 2009), and the relationship between psychological symptoms (e.g., Alloy, Just, & Panzarella, 1997; Hanson & Chen, 2010; Oslin, Cary, Slaymaker, Colleran, & Blow, 2009) and might be capable of revealing mechanisms of change (Kazdin, 2007, 2009, 2010). Between-subjects and within-subject SCED designs with repeated measurements enable researchers to examine similarities and differences in the course of change, both during and as a result of manipulating an IV. Temporal effects have been largely overlooked in many areas of psychological science (Bolger et al., 2003): Examining temporal relationships is sorely needed to further our understanding of the etiology and amplification of numerous psychological phenomena.

Time-series studies were very infrequently found in this literature search (2%). Time-series studies traditionally occur in subfields of psychology in which single-case research is not often used (e.g., personality, physiological/biological). Recent advances in methods for collecting and analyzing time-series data (e.g., Borckardt et al., 2008) could expand the use of time-series methodology in the SCED community. One problem with drawing firm conclusions from this particular review finding is a semantic factor: Time-series is a specific term reserved for measurement occurring at a uniform interval. However, SCED research appears to not yet have adopted this language when referring to data collected in this fashion. When time-series data analytic methods are not used, the matter of measurement interval is of less importance and might not need to be specified or described as a time-series. An interesting extension of this work would be to examine SCED research that used time-series measurement strategies but did not label it as such. This is important because then it could be determined how many SCEDs could be analyzed with time-series statistical methods.

Daily diary and ecological momentary assessment methods. EMA and daily diary approaches represent methodological procedures for collecting repeated measurements in time-series and non-time-series experiments, which are also known as experience sampling. Presenting an in-depth discussion of the nuances of these sampling techniques is well beyond the scope of this article. The reader is referred to the following review articles: daily diary (Bolger et al., 2003; Reis & Gable, 2000; Thiele, Laireiter, & Baumann, 2002), and EMA (Shiffman et al., 2008). Experience sampling in psychology has burgeoned in the past two decades as technological advances have permitted more precise and immediate reporting by participants (e.g., Internet-based, two-way pagers, cellular telephones, handheld computers) than do paper-and-pencil methods (for reviews, see Barrett & Barrett, 2001; Shiffman & Stone, 1998). Both methods have practical limitations and advantages. For example, electronic methods are more costly and may exclude certain subjects from participating in the study, either because they do not have access to the necessary technology or they do not have the familiarity or savvy to successfully complete
reporting. Electronic data-collection methods enable the researcher to prompt responses at random or predetermined intervals and also accurately assess compliance. Paper-and-pencil methods have been criticized for their inability to reliably track respondents’ compliance: Palermo, Valenzuela, and Stork (2004) found better compliance with electronic diaries than with paper and pencil. On the other hand, Green, Rafaei, Bolger, Shrout, and Reis (2006) demonstrated the psychometric data structure equivalence between these two methods, suggesting that the data collected in either method will yield similar statistical results given comparable compliance rates.

Daily diary/daily self-report and EMA measurement were somewhat rarely represented in this review, occurring in only 6.1% of the total studies. EMA methods had been used in only one of the reviewed studies. The recent proliferation of EMA and daily diary studies in psychology reported by others (Bolger et al., 2003; Piasecki et al., 2007; Shiffman et al., 2008) suggests that these methods have not yet reached SCED researchers, which could, in part, have resulted from the long-held supremacy of observational measurement in fields that commonly practice single-case research.

Measurement Standards

As was previously mentioned, measurement in SCEDs requires the reliable assessment of change over time. As illustrated in Table 4, DIV16 and the NRP explicitly require that reliability of all measures be reported. DIV12 provides little direction in the selection of the measurement instrument, except to require that three or more clinically important behaviors with relative independence be assessed. Similarly, the only item concerned with measurement on the Tate et al. (2008) scale specifies assessing behaviors consistent with the target of the intervention. The WWC and the Tate et al. scale require at least two independent assessors of the DV and that interrater reliability meeting minimum established thresholds be reported. Furthermore, WWC requires that interrater reliability be assessed on at least 20% of the data in each phase and in each condition. DIV16 expects that assessment of the outcome measures will be multisource and multimethod, when applicable. The interval of measurement is not specified by any of the reviewed sources. The WWC and the Tate et al. scale require that DVs be measured repeatedly across phases (e.g., baseline and treatment), which is a typical requirement of a SCED. The NRP asks that the time points at which DV measurement occurred be reported.

Baseline. The baseline measurement represents one of the most crucial design elements of the SCED. Because subjects provide their own data for comparison, gathering a representative, stable sampling of behavior before manipulating the IV is essential to accurately inferring an effect. Some researchers have reported the typical length of the baseline period to range from three to 12 observations in intervention research applications (e.g., Center et al., 1986; Huitema, 1985; R. R. Jones et al., 1977; Sharpley, 1987); Huitema’s (1985) review of 881 experiments published in the Journal of Applied Behavior Analysis resulted in a modal number of three to four baseline points. Center et al. (1986) suggested five as the minimum number of baseline measurements needed to accurately estimate autocorrelation. Longer baseline periods suggest a greater likelihood of a representative measurement of the DVs, which has been found to increase the validity of the effects and reduce bias resulting from autocorrelation (Huitema & McKean, 1994). The results of this review are largely consistent with those of previous researchers: The mean number of baseline observations was found to be 10.22 (SD = 9.59), and 6 was the modal number of observations. Baseline data were available in 77.8% of the reviewed studies. Although the baseline assessment has tremendous bearing on the results of a SCED study, it was often difficult to locate the exact number of data points. Similarly, the number of data points assessed across all phases of the study were not easily identified.

The WWC, DIV12, and DIV16 agree that a minimum of three data points during the baseline is necessary. However, to receive the highest rating by the WWC, five data points are necessary in each phase, including the baseline and any subsequent withdrawal baselines as would occur in a reversal design. DIV16 explicitly states that more than three points are preferred and further stipulates that the baseline must demonstrate stability (i.e., limited variability), absence of overlap between the baseline and other phases, absence of a trend, and that the level of the baseline measurement is severe enough to warrant intervention; each of these aspects of the data is important in inferential accuracy. Detrending techniques can be used to address baseline data trend. The integration option in ARIMA-based modeling and the empirical mode decomposition method (Wu, Huang, Long, & Peng, 2007) are two sophisticated detrending techniques. In regression-based analytic methods, detrending can be accomplished by simply regressing each variable in the model on time (i.e., the residuals become the detrended series), which is analogous to adding a linear, exponential, or quadratic term to the regression equation.

NRP does not provide a minimum for data points, nor does the Tate et al. (2008) scale, which requires only a sufficient sampling of baseline behavior. Although the mean and modal number of baseline observations is well within these parameters, seven (1.7%) studies reported mean baselines of less than three data points.

Establishing a uniform minimum number of required baseline observations would provide researchers and reviewers with only a starting guide. The baseline phase is important in SCED research because it establishes a trend that can then be compared with that of subsequent phases. Although a minimum number of observations might be required to meet standards, many more might be necessary to establish a trend when there is variability and trends in the direction of the expected effect. The selected data analytic approach also has some bearing on the number of necessary baseline observations. This is discussed further in the Analysis section.

Reporting of repeated measurements. Stone and Shiffman (2002) provided a comprehensive set of guidelines for the reporting of EMA data, which can also be applied to other repeated-measurement strategies. Because the application of EMA is widespread and not confined to specific research designs, Stone and Shiffman intentionally placed few constraints on researchers regarding selection of the DV and the reporter, which is determined by the research question under investigation. The methods of measurement, however, are specified in detail: Descriptions of prompting, recording of responses, participant-initiated entries, and the data acquisition interface (e.g., paper-and-pencil diary, PDA, cellular telephone) ought to be provided with sufficient detail for replication. Because EMA specifically, and time-series/
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dependent variable (DV)</td>
<td>( \geq 3 ) clinically important behaviors that are relatively independent</td>
<td>Outcome measures that produce reliable scores (validity of measure reported)</td>
<td>Standardized or investigator-constructed outcomes measures (report reliability)</td>
<td>Measure behaviors that are the target of the intervention</td>
<td>Determined by research question(s)</td>
<td></td>
</tr>
<tr>
<td>Selection of DV</td>
<td>More than one (self-report not acceptable)</td>
<td>Multisource (not always applicable)</td>
<td>Independent (implied minimum of 2)</td>
<td>Interrater reliability is reported</td>
<td>Determined by research question(s)</td>
<td></td>
</tr>
<tr>
<td>Assessor(s)/reporter(s)</td>
<td>On at least 20% of the data in each phase and in each condition. Must meet minimal established thresholds</td>
<td>Multimethod (e.g., at least 2 assessment methods to evaluate primary outcomes; not always applicable)</td>
<td>Quantitative or qualitative measure</td>
<td>Description of prompting, recording, participant-initiated entries, data acquisition interface (e.g., diary)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interrater reliability</td>
<td></td>
<td>List time points when dependent measures were assessed</td>
<td>Sampling of the targeted behavior (i.e., DV) occurs during the treatment period</td>
<td>Density and schedule are reported and consistent with addressing research question(s). Define “immediate and timely response”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method(s) of measurement/assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval of assessment</td>
<td>Must be measured repeatedly over time (no minimum specified) within and across different conditions and levels of the independent variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other guidelines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Baseline measurement</td>
<td>Minimum of 3 data points across multiple phases of a reversal or multiple baseline design; 5 data points in each phase for highest rating; 1 or 2 data points can be sufficient in alternating treatment designs</td>
<td>Minimum of 3 data points (more is preferred)</td>
<td>1. Minimum of 3 data points</td>
<td>No minimum specified</td>
<td>No minimum (“sufficient sampling of behavior [i.e., DV] occurred pretreatment”)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Stability (limited variability)</td>
<td>2. Stability (limited variability)</td>
<td>3. Absence of overlap between baseline and other phases</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Absence of overlap between baseline and other phases</td>
<td>3. Absence of overlap between baseline and other phases</td>
<td>4. Level (severe enough to warrant intervention)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Level (severe enough to warrant intervention)</td>
<td>4. Level (severe enough to warrant intervention)</td>
<td>5. Absence of trends</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5. Absence of trends</td>
<td>5. Absence of trends</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Compliance and missing data guidelines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. APA = American Psychological Association.
daily diary methods similarly, are primarily concerned with the interval of assessment, Stone and Shiffman suggested reporting the density and schedule of assessment. The approach is generally determined by the nature of the research question and pragmatic considerations, such as access to electronic data-collection devices at certain times of the day and participant burden. Compliance and missing data concerns are present in any longitudinal research design, but they are of particular importance in repeated-measurement applications with frequent measurement. When the research question pertains to temporal effects, compliance becomes paramount, and timely, immediate responding is necessary. For this reason, compliance decisions, rates of missing data, and missing data management techniques must be reported. The effect of missing data in time-series data streams has been the topic of recent research in the social sciences (e.g., Smith, Borckardt, & Nash, 2012; Velicer & Colby, 2005a, 2005b). The results and implications of these and other missing data studies are discussed in the next section.

Analysis of SCED Data

Visual analysis. Experts in the field generally agree about the majority of critical single-case experiment design and measurement characteristics. Analysis, on the other hand, is an area of significant disagreement, yet it has also received extensive recent attention and advancement. Debate regarding the appropriateness and accuracy of various methods for analyzing SCED data, the interpretation of single-case effect sizes, and other concerns vital to the validity of SCED results has been ongoing for decades, and no clear consensus has been reached. Visual analysis, following systematic procedures such as those provided by Franklin, Gorman, Beasley, and Allison (1997) and Parsonson and Baer (1978), remains the standard by which SCED data are most commonly analyzed (Parker, Cryer, & Byrns, 2006). Visual analysis can arguably be applied to all SCEDs. However, a number of baseline data characteristics must be met for effects obtained through visual analysis to be valid and reliable. The baseline phase must be relatively stable; free of significant trend, particularly in the hypothesized direction of the effect; have minimal overlap of data with subsequent phases; and have a sufficient sampling of behavior to be considered representative (Franklin, Gorman, et al., 1997; Parsonson & Baer, 1978). The effect of baseline trend on visual analysis, and a technique to control baseline trend, are offered by Parker et al. (2006). Kazdin (2010) suggested using statistical analysis when a trend or significant variability appears in the baseline phase, two conditions that ought to preclude the use of visual analysis techniques. Visual analysis methods are especially adept at determining intervention effects and can be of particular relevance in real-world applications (e.g., Kratochwill, Levin, Horner, & Swoboda, 2011).

However, visual analysis has its detractors. It has been shown to be inconsistent, can be affected by autocorrelation, and results in overestimation of effect (e.g., Matyas & Greenwood, 1990). Visual analysis as a means of estimating an effect precludes the results of SCED research from being included in meta-analysis and also makes it very difficult to compare results to the effect sizes generated by other statistical methods. Yet visual analysis proliferates in large part because SCED researchers are familiar with these methods and are not only generally unfamiliar with statistical approaches but lack agreement about their appropriateness. Still, top experts in single-case analysis champion the use of statistical methods alongside visual analysis whenever it is appropriate to do so (Kratochwill et al., 2011).

Statistical analysis. Statistical analysis of SCED data consists generally of an attempt to address one or more of three broad research questions: (a) Does introduction/ manipulation of the IV result in statistically significant change in the level of the DV (level-change or phase-effect analysis)? (b) Does introduction/ manipulation of the IV result in statistically significant change in the slope of the DV over time (slope-change analysis)? (c) Do meaningful relationships exist between the trajectory of the DV and other potential covariates? Level- and slope-change analyses are relevant to intervention effectiveness studies and other research questions in which the IV is expected to result in changes in the DV in a particular direction. Visual analysis methods are most adept at addressing research questions pertaining to changes in level and slope (Questions a and b), most often using some form of graphical representation and standardized computation of a mean level or trend line within and between each phase of interest (e.g., Horner & Spaulding, 2010; Kratochwill et al., 2011; Matyas & Greenwood, 1990). Research questions in other areas of psychological science might address the relationship between DVs or the slopes of DVs (Question c). A number of sophisticated modeling approaches (e.g., cross-lag, multilevel, panel, growth mixture, latent class analysis) may be used for this type of question, and some are discussed in greater detail later in this section. However, a discussion about the nuances of this type of analysis and all their possible methods is well beyond the scope of this article.

The statistical analysis of SCEDs is a contentious issue in the field. Not only is there no agreed-upon statistical method, but the practice of statistical analysis in the context of the SCED is viewed by some as unnecessary (see Shadish, Rindskopf, & Hedges, 2008). Traditional trends in the prevalence of statistical analysis usage by SCED researchers are revealing: Busk and Marascuilo (1992) found that only 10% of the published single-case studies they reviewed used statistical analysis; Brossart, Parker, Olson, and Mahadevan (2006) estimated that this figure had roughly doubled by 2006. A range of concerns regarding single-case effect size calculation and interpretation is discussed in significant detail elsewhere (e.g., Campbell, 2004; Cohen, 1994; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Kirk, 1996; Manolov & Solanas, 2008; Olive & Smith, 2005; Parker & Brossart, 2003; Robey et al., 1999; Smith et al., 2012; Velicer & Fava, 2003). One concern is the lack of a clearly superior method across data sets. Although statistical methods for analyzing SCEDs abound, few studies have examined their comparative performance with the same data set. The most recent studies of this kind, performed by Brossart et al. (2006), Campbell (2004), Parker and Brossart (2003), and Parker and Vannest (2009), found that the more promising available statistical analysis methods yielded moderately different results on the same data series, which led them to conclude that each available method is equipped to adequately address only a relatively narrow spectrum of data. Given these findings, analysts need to select an appropriate model for the research questions and data structure, being mindful of how modeling results can be influenced by extraneous factors.
The current standards unfortunately provide little guidance in the way of statistical analysis options. This article presents an admittedly cursory introduction to available statistical methods; many others are not covered in this review. The following articles provide more in-depth discussion and description of other methods: Barlow et al. (2008); Franklin, Allison, and Gorman (1997); Kazdin (2010); and Kratochwill and Levin (1992, 2010). Shadish et al. (2008) summarized more recently developed methods. Similarly, a special issue of Evidence-Based

**Systematic Review of SCED Research and Standards**

An introduction to autocorrelation and its implications for statistical analysis is necessary before specific analytic methods can be discussed. It is also pertinent at this time to discuss the implications of missing data.

**Autocorrelation.** Many repeated measurements within a single subject or unit create a situation with which most psychological researchers are unaccustomed to dealing: autocorrelated data, which is the nonindependence of sequential observations, also known as *serial dependence*. Basic and advanced discussions of autocorrelation in single-subject data can be found in Borckardt et al. (2008); Huitema (1985), and Marshall (1980), and discussions of autocorrelation in multilevel models can be found in Snijders and Bosker (1999) and Diggle and Liang (2001). Along with trend and seasonal variation, autocorrelation is one example of the internal structure of repeated measurements. In the social sciences, autocorrelated data occur most naturally in the fields of physiological psychology, econometrics, and finance, where each phase of interest has potentially hundreds or even thousands of observations that are tightly packed across time (e.g., electroencephalography actuarial data, financial market indices). Applied SCED research in most areas of psychology is more likely to have measurement intervals of day, week, or hour.

Autocorrelation is a direct result of the repeated-measurement requirements of the SCED, but its effect is most noticeable and problematic when one is attempting to analyze these data. Many commonly used data analytic approaches, such as analysis of variance, assume independence of observations and can produce spurious results when the data are nonindependent. Even statistically insignificant autocorrelation estimates are generally viewed as sufficient to cause inferential bias when conventional statistics are used (e.g., Busk & Marascuilo, 1988; R. R. Jones et al., 1977; Matyas & Greenwood, 1990). The effect of autocorrelation on statistical inference in single-case applications has also been known for quite some time (e.g., R. R. Jones et al., 1977; Kanfer, 1970; Kazdin, 1981; Marshall, 1980). The findings of recent simulation studies of single-subject data streams indicate that autocorrelation is a nontrivial matter. For example, Manolov and Solanas (2008) determined that calculated effect sizes were linearly related to the autocorrelation of the data stream, and Smith et al. (2012) demonstrated that autocorrelation estimates in the vicinity of 0.80 negatively affect the ability to correctly infer a significant level-change effect using a standardized mean differences method. Huitema and colleagues (e.g., Huitema, 1985; Huitema & McKean, 1994) argued that autocorrelation is rarely a concern in applied research. Huitema’s methods and conclusions have been questioned and opposing data have been published (e.g., Allison & Gorman, 1993; Matyas & Greenwood, 1990; Robey et al., 1999), resulting in abandonment of the position that autocorrelation can be conscionably ignored without compromising the validity of the statistical procedures. Procedures for removing autocorrelation in the data stream prior to calculating effect sizes are offered as one option: One of the more promising analysis methods, ARMA/ARIMA (discussed later in this article), was specifically designed to remove the internal structure of time-series data, such as autocorrelation, trend, and seasonality (Box & Jenkins, 1970; Tiao & Box, 1981).

**Missing observations.** Another concern inherent in repeated-measures designs is missing data. Daily diary and EMA methods are intended to reduce the risk of retrospective error by eliciting accurate, real-time information (Bolger et al., 2003). However, these methods are subject to missing data as a result of honest forgetfulness, not possessing the diary collection tool at the specified time of collection, and intentional or systematic noncompliance. With paper-and-pencil diaries and some electronic methods, subjects might be able to complete missed entries retrospectively, defeating the temporal benefits of these assessment strategies (Bolger et al., 2003). Methods of managing noncompliance through the study design and measurement methods include training the subject to use the data-collection device appropriately, using technology to prompt responding and track the time of response, and providing incentives to participants for timely compliance (for additional discussion of this topic, see Bolger et al., 2003; Shiffman & Stone, 1998).

Even when efforts are made to maximize compliance during the conduct of the research, the problem of missing data is often unavoidable. Numerous approaches exist for handling missing observations in group multivariate designs (e.g., Horton & Kleinman, 2007; Ibrahim, Chen, Lipsitz, & Herring, 2005). Raghunathan (2004) and others concluded that full information and raw data maximum likelihood methods are preferable. Velicer and Colby (2005a, 2005b) established the superiority of maximum likelihood methods over listwise deletion, mean of adjacent observations, and series mean substitution in the estimation of various critical time-series data parameters. Smith et al. (2012) extended these findings regarding the effect of missing data on inferential precision. They found that managing missing data with the EM procedure (Dempster, Laird, & Ruben, 1977), a maximum likelihood algorithm, did not affect one’s ability to correctly infer a significant effect. However, lag-1 autocorrelation estimates in the vicinity of 0.80 resulted in insufficient power sensitivity (<0.80), regardless of the proportion of missing data (10%, 20%, 30%, or 40%). Although maximum likelihood methods have garnered some empirical support, methodological strategies that minimize missing data, particularly systematically missing data, are paramount to post hoc statistical remedies.

**Nonnormal distribution of data.** In addition to the autocorrelated nature of SCED data, typical measurement methods also present analytic challenges. Many statistical methods, particularly those involving model finding, assume that the data are normally distributed. This is often not satisfied in SCED research when

---

1 Autocorrelation estimates in this range can be caused by trends in the data streams, which creates complications in terms of detecting level-change effects. The Smith et al. (2012) study used a Monte Carlo simulation to control for trends in the data streams, but trends are likely to exist in real-world data with high lag-1 autocorrelation estimates.
measurements involve count data, observer-rated behaviors, and other, similar metrics that result in skewed distributions. Techniques are available to manage nonnormal distributions in regression-based analysis, such as zero-inflated Poisson regression (D. Lambert, 1992) and negative binomial regression (Gardner, Mulvey, & Shaw, 1995), but many other statistical analysis methods do not include these sophisticated techniques. A skewed data distribution is perhaps one of the reasons Kazdin (2010) suggested not using count, categorical, or ordinal measurement methods.

Available statistical analysis methods. Following is a basic introduction to the more promising and prevalent analytic methods for SCED research. Because there is little consensus regarding the superiority of any single method, the burden unfortunately falls on the researcher to select a method capable of addressing the research question and handling the data involved in the study. Some indications and contraindications are provided for each method presented here.

Multilevel and structural equation modeling. Multilevel modeling (MLM; e.g., Schmidt, Perels, & Schmitz, 2010) techniques represent the state of the art among parametric approaches to SCED analysis, particularly when synthesizing SCED results (Shadish et al., 2008). MLM and related latent growth curve and factor mixture methods in structural equation modeling (SEM; e.g., Lubke & Muthén, 2005; B. O. Muthén & Curran, 1997) are particularly effective for evaluating trajectories and slopes in longitudinal data and relating changes to potential covariates. MLM and related hierarchical linear models (HLM) can also illuminate the relationship between the trajectories of different variables under investigation and clarify whether these relationships differ amongst the subjects in the study. Time-series and cross-lag analyses can also be used in MLM and SEM (Chow, Ho, Hamaker, & Dolan, 2010; du Toit & Browne, 2007). However, they generally require sophisticated model-fitting techniques, making them difficult for many social scientists to implement. The structure (autocorrelation) and trend of the data can also complicate many MLM methods. The common, short data streams in SCED research and the small number of subjects also present problems to MLM and SEM approaches, which were developed for data with significantly greater numbers of observations when the number of subjects is fewer, and for a greater number of participants for model-fitting purposes, particularly when there are fewer data points. Still, MLM and related techniques arguably represent the most promising analytic methods.

A number of software options exist for SEM. Popular statistical packages in the social sciences provide SEM options, such as PROC CALIS in SAS, the AMOS module (Arbuckle, 2006) of SPSS, and the SEM package for R (R Development Core Team, 2005), the use of which is described by Fox (2006). A number of stand-alone software options are also available for SEM applications, including Mplus (L. K. Muthén & Muthén, 2010) and Stata (StataCorp., 2011). Each of these programs also provides options for estimating multilevel/hierarchical models (for a review of using these programs for MLM analysis, see Albright & Marinova, 2010). Hierarchical linear and nonlinear modeling can also be accomplished using the HLM 7 program (Raudenbush, Bryk, & Congdon, 2011).

Autoregressive moving averages (ARMA; e.g., Browne & Neselroade, 2005; Liu & Hudack, 1995; Tiao & Box, 1981). Two primary points have been raised regarding ARMA modeling: length of the data stream and feasibility of the modeling technique. ARMA models generally require 30–50 observations in each phase when analyzing a single-subject experiment (e.g., Borckardt et al., 2008; Box & Jenkins, 1970), which is often difficult to satisfy in applied psychological research applications. However, ARMA models in an SEM framework, such as those described by du Toit and Browne (2001), are well suited for longitudinal panel data with few observations and many subjects. Autoregressive SEM models are also applicable under similar conditions. Model-fitting options are available in SPSS, R, and SAS via PROC ARMA.

ARMA modeling also requires considerable training in the method and rather advanced knowledge about statistical methods (e.g., Kratochwill & Levin, 1992). However, Brossart et al. (2006) pointed out that ARMA-based approaches can produce excellent results when there is no “model finding” and a simple lag-1 model, when no differencing and no moving average, is used. This approach can be taken for many SCED applications when phase- or slope-change analyses are of interest with a single, or very few, subjects. As already mentioned, this method is particularly useful when one is seeking to account for autocorrelation or other over-time variations that are not directly related to the experimental or intervention effect of interest (i.e., detrending). ARMA and other time-series analysis methods require missing data to be managed prior to analysis by means of options such as full information maximum-likelihood estimation, multiple imputation, or the Kalman filter (see Box & Jenkins, 1970; Hamilton, 1994; Shumway & Stoffer, 1982) because listwise deletion has been shown to result in inaccurate time-series parameter estimates (Velicer & Colby, 2005a).

Standardized mean differences. Standardized mean differences approaches include the common Cohen’s d, Glass’s delta, and Hedge’s g that are used in the analysis of group designs. The computational properties of mean differences approaches to SCEDs are identical to those used for group comparisons, except that the results represent within-case variation instead of the variation between groups, which suggests that the obtained effect sizes are not interpretively equivalent. The advantage of the mean differences approach is its simplicity of calculation and also its familiarity to social scientists. The primary drawback of these approaches is that they were not developed to contend with autocorrelated data. However, Manolov and Solanas (2008) reported that autocorrelation least affected effect sizes calculated using standardized mean differences approaches. To the applied-research scientist this likely represents the most accessible analytic approach, because statistical software is not required to calculate these effect sizes. The resultant effect sizes of single subject standardized mean differences analysis must be interpreted cautiously because their relation to standard effect size benchmarks, such as those provided by Cohen (1988), is unknown. Standardized mean differences approaches are appropriate only when examining significant differences between phases of the study and cannot illuminate trajectories or relationships between variables.

2 The author makes no endorsement regarding the superiority of any statistical program or package over another by their mention or exclusion in this article. The author also has no conflicts of interest in this regard.
Other analytic approaches. Researchers have offered other analytic methods to deal with the characteristics of SCED data. A number of methods for analyzing N-of-1 experiments have been developed. Borckardt’s (2006) Simulation Modeling Analysis (SMA) program provides a method for analyzing level- and slope-change in short (<30 observations per phase; see Borckardt et al., 2008), autocorrelated data streams that is statistically sophisticated yet accessible and freely available to typical psychological scientists and clinicians. A replicated single-case time-series design conducted by Smith, Handler, and Nash (2010) provided an example of SMA application. The Singwin Package, described in Bloom et al. (2003), is another easy-to-use parametric approach for analyzing single-case experiments. A number of nonparametric approaches have also been developed that emerged from the visual analysis tradition: Some examples include percent nonoverlapping data (Scruggs, Mastropieri, & Casto, 1987) and nonoverlap of all pairs (Parker & Vannest, 2009); however, these methods have come under scrutiny, and Wolery, Busick, Reichow, and Barton (2010) have suggested abandoning them altogether. Each of these methods appears to be well suited for managing specific data characteristics, but they should not be used to analyze data streams beyond their intended purpose until additional empirical research is conducted.

Combining SCED Results

Beyond the issue of single-case analysis is the matter of integrating and meta-analyzing the results of single-case experiments. SCEDs have been given short shrift in the majority of meta-analytic literature (Littell, Corcoran, & Pillai, 2008; Shadish et al., 2008), with only a few exceptions (Carr et al., 1999; Horner & Spaulding, 2010). Currently, few proven methods exist for integrating the results of multiple single-case experiments. Allison and Gorman (1993) and Shadish et al. (2008) presented the problems associated with meta-analyzing single-case effect sizes, and W. P. Jones (2003); Manolov and Solanas (2008); Scruggs and Mastropieri (1998), and Shadish et al. (2008) offered four different potential statistical solutions for this problem, none of which appear to have received consensus amongst researchers. The ability to synthesize and compare single-case effect sizes, particularly effect sizes garnered through group design research, is undoubtedly necessary to increase SCED proliferation.

Discussion of Review Results and Coding of Analytic Methods

The coding criteria for this review were quite stringent in terms of what was considered to be either visual or statistical analysis. For visual analysis to be coded as present, it was necessary for the authors to self-identify as having used a visual analysis method. In many cases, it could likely be inferred that visual analysis had been used, but it was often not specified. Similarly, statistical analysis was reserved for analytic methods that produced an effect. Analyses that involved comparing magnitude of change using raw count data or percentages were not considered rigorous enough. These two narrow definitions of visual and statistical analysis contributed to the high rate of unreported analytic method, shown in Table 1 (52.3%). A better representation of the use of visual and statistical analysis would likely be the percentage of studies within those that reported a method of analysis. Under these parameters, 41.5% used visual analysis and 31.3% used statistical analysis. Included in these figures are studies that included both visual and statistical methods (11%). These findings are slightly higher than those estimated by Brossart et al. (2006), who estimated statistical analysis is used in about 20% of SCED studies. Visual analysis continues to be undoubtedly the most prevalent method, but there appears to be a trend for increased use of statistical approaches, which is likely to only gain momentum as innovations continue.

Analysis Standards

The standards selected for inclusion in this review offer minimal direction in the way of analyzing the results of SCED research. Table 5 summarizes analysis-related information provided by the six reviewed sources for SCED standards. Visual analysis is acceptable to DIV12 and DIV16, along with unspecified statistical approaches. In the WWC standards, visual analysis is the acceptable method of determining an intervention effect, with statistical analyses and randomization tests permissible as a complementary or supporting method to the results of visual analysis methods. However, the authors of the WWC standards state, “As the field reaches greater consensus about appropriate statistical analyses and quantitative effect-size measures, new standards for effect demonstration will need to be developed” (Kratochwill et al., 2010, p. 16). The NRP and DIV12 seem to prefer statistical methods when they are warranted. The Tate et al. (2008) scale accepts only statistical analysis with the reporting of an effect size. Only the WWC and DIV16 provide guidance in the use of statistical analysis procedures: The WWC “recommends” nonparametric and parametric approaches, multilevel modeling, and regression when statistical analysis is used. DIV16 refers the reader to Wilkinson and the Task Force on Statistical Inference (1999) for direction in this matter. Statistical analysis of daily diary and EMA methods is similarly unsettled. Stone and Shiffman (2002) asked for a detailed description of the statistical procedures used, in order for the approach to be replicated and evaluated. They provided direction for analyzing aggregated and disaggregated data. They also aptly noted that because many different modes of analysis exist, researchers must carefully match the analytic approach to the hypotheses being pursued.

Limitations and Future Directions

This review has a number of limitations that leave the door open for future study of SCED methodology. Publication bias is a concern in any systematic review. This is particularly true for this review because the search was limited to articles published in peer-reviewed journals. This strategy was chosen in order to inform changes in the practice of reporting and of reviewing, but it also is likely to have inflated the findings regarding the methodological rigor of the reviewed works. Inclusion of book chapters, unpublished studies, and dissertations would likely have yielded somewhat different results.

However, it should be noted that it was often very difficult to locate an actual effect size reported in studies that used statistical analysis. Although this issue would likely have added little to this review, it does inhibit the inclusion of the results in meta-analysis.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Visual analysis</td>
<td>4-step, 6-variable procedure (based on Parsonson &amp; Baer, 1978)</td>
<td>Acceptable (no specific guidelines or procedures offered)</td>
<td>1. Change in level</td>
<td></td>
<td>Not acceptable (“use statistical analyses or describe effect sizes”; p. 389)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Minimal score overlap</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Change in trend</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Adequate length (≥3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5. Stable data (Franklin, Gorman, Beasley, &amp; Allison, 1997; Parsonson &amp; Baer, 1992)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Statistical analysis procedures</td>
<td>Estimating effect sizes: nonparametric and parametric approaches, multilevel modeling, and regression (recommended)</td>
<td>Preferred when the number of data points warrants statistical procedures (no specific guidelines or procedures offered)</td>
<td>Rely on the guidelines presented by Wilkinson and the Task Force on Statistical Inference (1999)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Demonstrating an effect</td>
<td>1. Documented consistency of level, trend, and variability within each phase 2. Documented immediacy of the effect, the proportion of overlap, the consistency of the data across phases 3. Identify for whom the intervention is and is not effective, if available 4. Examine external factors and anomalies 5. Follow-up of original study participants and multiple intervals with same outcome measures</td>
<td>ABAB—stable baseline established during first A period, data must show improvement during the first B period, reversal or leveling of improvement during the second A period, and resumed improvement in the second B period (no other guidelines offered)</td>
<td>1. 0.05 alpha levels 2. Nonsignificant or negative outcomes noted 3. Type of effect size, type of data on which effect size is based, effect size statistic 4. Clinical/educational significance (e.g., social comparison)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(table continues)
A second concern is the stringent coding criteria in regard to the analytic methods and the broad categorization into visual and statistical analytic approaches. The selection of an appropriate method for analyzing SCED data is perhaps the murkiest area of this type of research. Future reviews that evaluate the appropriateness of selected analytic strategies and provide specific decision-making guidelines for researchers would be a very useful contribution to the literature. Although six sources of standards apply to SCED research reviewed in this article, five of them were developed almost exclusively to inform psychological and behavioral intervention research. The principles of SCED research remain the same in different contexts, but there is a need for nonintervention scientists to weigh in on these standards.

Finally, this article provides a first step in the synthesis of the available SCED reporting guidelines. However, it does not resolve disagreements, nor does it purport to be a definitive source. In the future, an entity with the authority to construct such a document ought to convene and establish a foundational, adaptable, and agreed-upon set of guidelines that cuts across subspecialties but is applicable to many, if not all, areas of psychological research, which is perhaps an idealistic goal. Certain preferences will undoubtedly continue to dictate what constitutes acceptable practice in each subspecialty of psychology, but uniformity along critical dimensions will help advance SCED research.

Conclusions

The first decade of the 21st century has seen an upwelling of SCED research across nearly all areas of psychology. This article contributes updated benchmarks in terms of the frequency with which SCED design and methodology characteristics are used, including the number of baseline observations, assessment and measurement practices, and data analytic approaches, most of which are largely consistent with previously reported benchmarks. However, this review is much broader than those of previous research teams and also breaks down the characteristics of single-case research by the predominant design. With the recent SCED proliferation came a number of standards for the conduct and reporting of such research. This article also provides a much-needed synthesis of recent SCED standards that can inform the work of researchers, reviewers, and funding agencies conducting and evaluating single-case research, which reveals many areas of consensus as well as areas of significant disagreement. It appears that the question of where to go next is very relevant at this point in time. The majority of the research design and measurement characteristics of the SCED are reasonably well established, and the results of this review suggest general practice that is in accord with existing standards and guidelines, at least in regard to published peer-reviewed works. In general, the published literature appears to be meeting the basic design and measurement requirements to ensure adequate internal validity of SCED studies.

Consensus regarding the superiority of any one analytic method stands out as an area of divergence. Judging by the current literature and lack of consensus, researchers will need to carefully select a method that matches the research design, hypotheses, and intended conclusions of the study, while also considering the most up-to-date empirical support for the chosen analytic method, whether it be visual or statistical. In some cases the number of observations and subjects in the study will dictate which analytic

### Table 5 (continued)

<table>
<thead>
<tr>
<th>Variable</th>
<th>What Works Clearinghouse</th>
<th>APA Division 12 Task Force on Psychological Interventions</th>
<th>APA Division 16 Task Force on Evidence-Based Interventions in School Psychology</th>
<th>National Reading Panel</th>
<th>Tate et al., 2008</th>
<th>Stone &amp; Shiffman, 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication</td>
<td>4. Replication</td>
<td>1. Minimum of 5 studies</td>
<td>1. 3 replications of ≥5 subjects each</td>
<td>1. Same intervention, treatment protocol and duration</td>
<td>1. 3 replications of ≥5 subjects each</td>
<td>1. Same intervention, treatment protocol and duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. The studies should be conducted by at least 3 different research teams at 3 different geographical locations</td>
<td>2. Replications conducted by ≥2 independent research groups</td>
<td>2. Same target problem and sample</td>
<td>2. Same target problem and sample</td>
<td>2. Same target problem and sample</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. The combined number of participants across the single-case examples totals at least 20</td>
<td>3. Independent evaluation</td>
<td>3. Independent evaluation</td>
<td>3. Independent evaluation</td>
<td>3. Independent evaluation</td>
</tr>
</tbody>
</table>

Note. APA = American Psychological Association.
methods can and cannot be used. In the case of the true N-of-1 experiment, there are relatively few sound analytic methods, and even fewer that are robust with shorter data streams (see Borckardt et al., 2008). As the number of observations and subjects increases, sophisticated modeling techniques, such as MLM, SEM, and ARMA, become applicable. Trends in the data and autocorrelation further obfuscate the development of a clear statistical analysis selection algorithm, which currently does not exist. Autocorrelation was rarely addressed or discussed in the articles reviewed, except when the selected statistical analysis dictated consideration. Given the empirical evidence regarding the effect of autocorrelation on visual and statistical analysis, researchers need to address this more explicitly. Missing-data considerations are similarly left out when they are unnecessary for analytic purposes. As newly devised statistical analysis approaches mature and are compared with one another for appropriateness in specific SCED applications, guidelines for statistical analysis will necessarily be revised. Similarly, empirically derived guidance, in the form of a decision tree, must be developed to ensure application of appropriate methods based on characteristics of the data and the research questions being addressed. Researchers could also benefit from tutorials and comparative reviews of different software packages: This is a needed area of future research. Powerful and reliable statistical analyses help move the SCED up the ladder of experimental designs and attenuate the view that the method applies primarily to pilot studies and idiosyncratic research questions and situations.

Another potential future advancement of SCED research comes in the area of measurement. Currently, SCED research gives significant weight to observer ratings and seems to discourage other forms of data-collection methods. This is likely due to the origins of the SCED in behavioral assessment and applied behavior analysis, which remains a present-day stronghold. The dearth of EMA and diary-like sampling procedures within the SCED research reviewed, yet their ever-growing prevalence in the larger psychological research arena, highlights an area for potential expansion. Observational measurement, although reliable and valid in many contexts, is time and resource intensive and not feasible in all areas in which psychologists conduct research. It seems that numerous untapped research questions are stifled because of this measurement constraint. SCED researchers developing updated standards in the future should include guidelines for the appropriate measurement requirement of non-observer-reported data. For example, the results of this review indicate that reporting of repeated measurements, particularly the high-density type found in diary and EMA sampling strategies, ought to be more clearly spelled out, with specific attention paid to autocorrelation and trend in the data streams. In the event that SCED researchers adopt self-reported assessment strategies as viable alternatives to observation, a set of standards explicitly identifying the necessary psychometric properties of the measures and specific items used would be in order.

Along similar lines, SCED researchers could take a page from other areas of psychology that champion multimethod and multisource evaluation of primary outcomes. In this way, the long-standing tradition of observational assessment and the cutting-edge technological methods of EMA and daily diary could be married with the goal of strengthening conclusions drawn from SCED research and enhancing the validity of self-reported outcome assessment. The results of this review indicate that they rarely intersect today, and I urge SCED researchers to adopt other methods of assessment informed by time-series, daily diary, and EMA methods. The EMA standards could serve as a jumping-off point for refined measurement and assessment reporting standards in the context of multimethod SCED research.

One limitation of the current SCED standards is their relatively limited scope. To clarify, with the exception of the Stone and Shiffman (2002) EMA reporting guidelines, the other five sources of standards were developed in the context of designing and evaluating intervention research. Although this is likely to remain its patent emphasis, SCEDs are capable of addressing other pertinent research questions in the psychological sciences, and the current standards truly only roughly approximate salient crosscutting SCED characteristics. I propose developing broad SCED guidelines that address the specific design, measurement, and analysis issues in a manner that allows it to be useful across applications, as opposed to focusing solely on intervention effects. To accomplish this task, methodology experts across subspecialties in psychology would need to convene. Admittedly this is no small task.

Perhaps funding agencies will also recognize the fiscal and practical advantages of SCED research in certain areas of psychology. One example is in the field of intervention effectiveness, efficacy, and implementation research. A few exemplary studies using robust forms of SCED methodology are needed in the literature. Case-based methodologies will never supplant the group design as the gold standard in experimental applications, nor should that be the goal. Instead, SCEDs provide a viable and valid alternative experimental methodology that could stimulate new areas of research and answer questions that group designs cannot. With the astonishing number of studies emerging every year that use single-case designs and explore the methodological aspects of the design, we are poised to witness and be a part of an upsurge in the sophisticated application of the SCED. When federal granting agencies and journal editors begin to use formal standards while making funding and publication decisions, the field will benefit.

Last, for the practice of SCED research to continue and mature, graduate training programs must provide students with instruction in all areas of the SCED. This is particularly true of statistical analysis techniques that are not often taught in departments of psychology and education, where the vast majority of SCED studies seem to be conducted. It is quite the conundrum that the best available statistical analytic methods are often cited as being inaccessible to social science researchers who conduct this type of research. This need not be the case. To move the field forward, emerging scientists must be able to apply the most state-of-the-art research designs, measurement techniques, and analytic methods.

References


StataCorp. (2011). Stata statistical software (Release 12). College Station, TX: Author.


Appendix

Results of Systematic Review Search and Studies Included in the Review

* Indicates inclusion in study (N = 409).


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


(Appendix continues)


Appendix continues


(Appendix continues)

(Appendix continues)


(Appendix continues)


Appendix continues


Received March 21, 2011
Revision received March 31, 2012
Accepted April 20, 2012