Training and Education in Professional Psychology
2014, Vol. 8, No. 3, 149–157

1931-3918/14/$12.00 http://dx.doi.org/10.1037/tep0000031

A Systematic Review and Reformulation of Outcome Evaluation in Clinical Supervision: Applying the Fidelity Framework

Robert P. Reiser
Reiser Healthcare Consulting, San Anselmo, California

Derek L. Milne
Newcastle University

Although a strikingly diverse range of outcomes have been measured within clinical supervision research, a dominant perspective is that clinical outcomes remain the "acid test" of its effectiveness (Ellis & Ladany, 1997). We question the wisdom of this acid test logic in 2 ways. First, we summarize alternative conceptualizations of outcome from within the supervision field, highlighting several important reasons for considering clinical benefit as but one of several equally valid, stepwise outcomes. The fidelity framework (Borrelli et al., 2005) is drawn upon to show how these complementary outcomes may be logically and advantageously combined. This framework's dimensions are the design, training, delivery, receipt, and enactment of an intervention. Second, a sample of 12 interpretable studies of the clinical outcomes of supervision is evaluated in terms of the studies' attention to these 5 dimensions. From this conceptual and empirical review, it is concluded that an overemphasis on clinical outcomes carries unnecessary risks (e.g., weak causal reasoning and a failure to identify mechanisms of change), while underemphasizing the several benefits of a more inclusive approach (e.g., increasing outcome research and improving supervision).

Keywords: clinical supervision, outcome, fidelity

Supervision has become an internationally accepted element within modern mental health services (Watkins & Milne, in press) a critical component in the training and regulation of therapists (Holloway & Neufeldt, 1995), and a major device for enhancing the quality of care (Falender & Shafranske, 2008). But this endorsement of supervision has taken place despite a weak evidence base (Ellis & Ladany, 1997). This mismatch between practice and research is more unsatisfactory, given the growing emphasis on competency-based practice (Falender et al., 2004; Rodolfa et al., 2013), accountability, and evidence-based practice (American Psychological Association, 2006). In response to this unsatisfactory gulf between research and practice, Watkins (2011) argued that it was imperative to demonstrate that supervision contributed to patient outcome. This has been termed the "acid test" of supervision: improving patient outcomes, usually measured in terms of symptomatic relief (Ellis & Ladany, 1997).

In the past decade, there have been at least eight reviews dealing more or less directly with the effect of supervision on patient outcomes, including Watkins (2011), Ellis, Ladany, Krengel, and Schult (1996), Freitas (2002), Holloway and Neufeldt (1995), Milne, Aylott, Fitzpatrick, and Ellis (2008), Roth, Pilling, and Turner (2010), Tsui (1997), and Wheeler and Richards (2007). These reviews all critiqued a carefully selected sample of empirical studies, concluding that there were myriad methodological weaknesses within this research. For example, Ellis et al. (1996) highlighted ambiguous hypotheses, the use of psychometrically weak instruments, and small sample sizes. They also noted a shift to pragmatic field studies that may have contributed to these weaknesses. For such methodological reasons, reviewers have concluded that there was insufficient evidence to infer a clear effect of supervision on patient outcomes, bearing out the interpretation offered by Watkins (2011).

Beyond the Acid Test

One response to this lack of proof of supervision's effectiveness, as well as to the lack of an established methodology for obtaining such proof, is to focus on more proximal and research-amenable outcomes, as argued by others (e.g., Holloway & Neufeldt, 1995). Although we do not question that clinical outcomes are a necessary element within a comprehensive outcome evaluation, there are several reasons for questioning that it represents a definitive demonstration of the effectiveness of supervision. One reason to question that clinical outcomes are the penultimate acid test of supervision is that the highest duty of supervision is surely to ensure safe practice (client protection), to oversee the therapy that is conducted by supervisees to try and ensure

Robert P. Reiser, PhD, a Fellow of the Academy of Cognitive Therapy, supervises cases and treats individuals and families with complex and serious mental illnesses. Since 2006, he has collaborated with Derek Milne on a series of research projects involving the development of an instrument (SAGE) to assess competence in supervision. He has provided numerous workshops and Institutes at the Association for Behavioral and Cognitive Therapies (ABCT) focused on improving supervision and training using empirically supported practices. He is a consulting supervisor for the CBT-D national training program with the Veterans Administration and has 8 years' experience running a doctoral level training clinic supervising empirically supported treatments.

Derek L. Milne, PhD, a Fellow of the British Psychological Society, has extensive experience as a supervisor, supervisor trainer, and supervision researcher (inc. authoring Evidence-Based Clinical Supervision, in 2009). This research program has included extensive collaboration with Robert Reiser, involving the development of an instrument (SAGE) to assess competence in supervision and related efforts to foster an evidence-based practice. He has recently retired as Director of The Doctorate in Clinical Psychology, within The School of Psychology, at Newcastle University, England.

Correspondence concerning this article should be addressed to Robert P. Reiser, Reiser Healthcare Consulting, 945 Butterfield Road, San Anselmo, CA 94960. E-mail: robert.reiser@gmail.com

that it causes no harm (the Hippocratic oath). From his review, Milne (2009) concluded that the purpose of supervision was to ensure safe and effective practice. For therapy to be safe and effective, supervision needs to prioritize changes in the supervisee (e.g., in terms of attitudes or competencies). To illustrate, Freitas (2002, p. 363) considered it "axiomatic that clinical supervision is conducted for the benefit of . . . trainees" (i.e., supervisees). Then there is a methodological reason to go beyond the acid test. Knowledge of outcomes in the absence of knowledge of the explanatory processes does little to advance interventions like supervision (Donabedian, 1988; Rossi, Freeman, & Lipsey, 2003).

Before Methodology

In thinking about outcomes that may complement the acid test and that contribute to the main purpose of supervision—ensuring safe and effective practice—it may be advantageous to give greater attention to how we conceptualize supervision's effectiveness. Specifically, how might we best construe the relationship between patient benefit and its causal factors, such as the ways that patients and supervisees interact (Holloway, 1984)? We wish to argue that the conceptualization of supervision outcomes is as deficient, but as necessary, for the proof of supervision's effectiveness as are the more commonly critiqued methodological considerations. Taken together, the reviews summarized have tended to adopt a methodological review approach, using either established research evaluation criteria in a systematic review approach (e.g., 37 validity threats were applied to the sampled studies within Ellis & Ladany, 1997) or general scientific reasoning within a less structured, narrative review approach (e.g., Freitas, 2002).

This emphasis on methodology contrasts with a very limited emphasis on conceptualization. In addition to definitional imprecision and the lack of consensus over what is deemed an outcome, reviews tended to lack a conceptual framework, adopting a more focused, ad hoc, and atheoretical approach to their analyses, corresponding to the reviewed studies (Tsui, 1997). The same criticism has been made of methodological reviews, that is, that they have tended to evaluate the sampled studies in an unsystematic way (Ellis & Ladany, 1997). We found few examples in which reviewers systematically utilized a conceptual framework to analyze their study samples. Exceptions included Holloway (1984) and Milne et al. (2008). Such explicit and coherent conceptual approaches helped to clarify some dominant themes that might otherwise have been missed, due to this diverse approach to supervision outcomes. As a result, a more precise and testable supervision model was outlined.

Rationale for the Fidelity Framework

A conceptual approach to literature review offers broadly the same advantages as a theoretical model. In particular, such a model helps to highlight the strengths and weaknesses of a literature within a logically coherent analysis, better clarifying what we know (and do not know), thus suggesting practice developments and research implications. An example is the fidelity framework, which refers to "methodological strategies used to monitor and enhance the reliability and validity of behavioral interventions" (Borrelli et al., 2005, p. 852). The value of high levels of treatment fidelity in the implementation of clinical trials has been summarized by Bellg et al. (2004) and others (e.g., Schoenwald et al., 2011). We selected this framework for the present review because it best captured the issue of patient outcomes related to the prior steps within supervision: It embodies a taxonomy, teasing apart the successive steps in getting from supervision to patient outcome. Such a stepwise approach seems especially valuable, given the clear distinctions that are made by researchers between steps like supervisee training, adherence, and transfer (Milne, 2009). In addition, the fidelity framework affords an expanded view of supervision, yet reflects current models of supervision—models that recognize complex processes of reciprocal influence (Holloway, 1984). Third, the fidelity framework was chosen because it uses dimensions and concepts that are implicitly accepted and applied by supervision researchers. An additional reason for adopting the fidelity framework is that it helps to specify the intervention and thus reduce unintended variability. It also increases the statistical power and improves the generalizability of study findings, by offering more readily replicated procedures for disseminating interventions like supervision. By contrast, low fidelity has a significant cost in terms of reducing the power of studies to detect significant effects, and in requiring larger and more expensive sample sizes (Borrelli et al., 2005).

A five-part framework for addressing fidelity in clinical trials—including study design, training, delivery, receipt, and enactment—has been developed by Bellg et al. (2004) and Borrelli et al. (2005). Fidelity in terms of *treatment design* involves detailing information about the dose of the experimental and comparison conditions (number of contacts, length of session, duration of contact over time, training of providers, the credentials of providers providing the treatment) and specifying the underlying theoretical model or clinical guidelines utilized. Fidelity in terms of *training providers* requires that the training of providers (for purposes of our review, the supervisors) is appropriately specified and standardized, and that there are measures of skill acquisition to determine how provider skills are improved and maintained over time. We have supplemented these criteria with those from Roth et al. (2010) requiring that there is a published description of the training procedures (e.g., a manual; see Table 1). This was our only addition to the fidelity framework, included to better specify the original meaning of training in Bellg et al. (2004) and Borrelli et al. (2005).

Fidelity in terms of the *delivery of treatment* involves a method to document and specify what has been delivered, including the content, dose, and a mechanism to ensure the supervisor actually adhered to the intervention plan; the assessment of nonspecific effects; and the use of an intervention (supervision) manual. Fidelity in terms of the *receipt of treatment* involves an assessment of the supervisees' comprehension of key elements of the intervention (supervision training) and a strategy to improve supervisee performance of the skills addressed in the intervention. Finally, *enactment of treatment skills* refers to assessment of the clinical outcomes of the supervisee's actual performance of such skills (in our case, objective assessments of the client symptoms, problems, functioning, or quality-of-life changes). It is recognized that the term "enactment" has additional meanings within the fidelity framework (e.g., generalization across settings) as well as across different therapies. By adopting treatment fidelity recommendations (Bellg et al. 2004; Borrelli et al., 2005; Moncher & Prinz, 1991) and applying them to the design and implementation of supervision studies, we selected a well-recognized approach to improving implementation and design in clinical trials. The fidelity framework is entirely consistent with, and supports, the purpose of supervision, that is, to ensure safe and effective practice.

Table 1
Fidelity Review Checklist Table (Adapted From Borrelli et al., 2005)

Fidelity framework element	Dimensions (with guiding criteria/examples)
Design of supervision	Dimension I. Criteria/examples: 1. Provided information about supervision dose in the intervention condition: Length of contact session(s) Number of contacts Duration of contact over time 2. Provided information about supervision dose in the comparison condition Length of contact session(s) Number of contacts Duration of contact over time 3. Mention of provider credentials 4. Mention of a theoretical model or clinical guidelines on which the supervision/intervention is based[b] 5. Was the supervision content defined (that is, a manual or treatment descriptions in books or papers)?[a,b] If a manual/supervision description was used, was there either: If the manual is in the public domain, the principal citation/ source; OR[a,b] If not in the public domain, a brief description of manual content (for example, whether the manual gives a comprehensive account of the complete intervention)[a,b] OR is an outline guide?[a,b]
Training of supervisors	Dimension II. Criteria/examples: 1. Description of how supervisors were trained 2. Standardized supervisor training 3. A description of training offered to supervisors (for example, outline content, number of sessions individual or group based, length of training)[b] 4. Measured supervisor skill acquisition posttraining 5. Described how supervisor skills maintained over time 6. Information about the supervisor(s) (for example, qualifications and experience)[a] 7. A description of supervision training arrangements (for example, number of sessions, frequency, and duration)[a] 8. Whether the results of integrity checks (adherence/competence/treatment fidelity measures) were used to signal the need for additional supervision/training *if therapists performed below criterion levels*[a]
Delivery of supervision	Dimension III. Criteria/examples: 1. Included method to ensure that the content of the supervision/intervention was being delivered as specified (e.g., supervision manual, checklist, computer program) 2. Included mechanism to assess if the supervisor actually adhered to the intervention (e.g., audiotape, observation, self-report of provider, exit interview with participant) 3. Assessed nonspecific supervision effects 4. Used supervision manual (see design of supervision element public domain, comprehensive account or outline only?[a] 5. A description of supervision arrangements (for example, number of sessions, frequency, and duration)[a] 6. Whether the results of integrity checks (adherence/competence/treatment fidelity measures) were used to signal the need for additional supervision/training *if therapists performed below criterion levels*[a]
Receipt of supervision	Dimension IV. Criteria/examples: 1. Assessed trainee comprehension of supervision during the intervention period 2. Included a strategy to improve trainee comprehension of supervision above and beyond what is included in the intervention 3. Assessed trainee's ability to perform the skills acquired during the intervention period 4. Included a strategy to improve trainee performance of intervention skills during the intervention period
Enactment of supervision	Dimension V. Criteria/examples 1. Assessed client outcome in terms of changes in problems, symptoms, quality of life (satisfaction and working alliance not considered full clinical outcome) 2. Assessed strategy to improve client performance of the acquired clinical skills in real-world settings

Note. From "A New Tool to Assess Treatment Fidelity and Evaluation of Treatment Fidelity Across 10 Years of Health Behavior Research," by B. Borrelli et al., 2005, *Journal of Consulting and Clinical Psychology*, 73, p. 855. Copyright 2005 APA. Adapted with permission.
[a] Adapted from Roth, Pilling, and Turner (2010). [b] Required field for meeting minimum dimension criteria.

Review Objectives

In this review we reexamine studies that have addressed the outcomes of clinical supervision utilizing the fidelity framework with the following objectives in mind:

1. Profile the extent to which the sampled studies address the fidelity framework dimensions, summarizing strengths and weaknesses;
2. Draw out the theoretical, methodological, and practice implications of our findings; and
3. Recommend future research directions and priorities.

Method

Inclusion Criteria

Past reviews have illustrated the dangers of the overinclusion of studies in reviews of supervision outcome, resulting in many studies in which client outcome was not directly addressed as a dependent variable (Watkins 2011). Hence, the following specific criteria determined study inclusion. Articles were only included if

they involved clinical supervision that was consistent with Milne's (2007) elaboration of the Bernard and Goodyear (2004) definition:

> The formal provision (i.e., sanctioned by relevant organization/s); by senior/qualified health practitioners (or similarly experienced staff) of an intensive education (general problem solving capacity; developing capability) and/or training (competence enhancement) that is case-focused and which supports, directs and guides (including also "restorative" and/or " normative" topics, addressed by means of professional methods, including objective monitoring, feedback and evaluation) the work of junior colleagues (supervisees). (Milne, 2007, p. 3).

For the purposes of this review, we included supervision in any format (e.g., both individual and group supervision), with "senior/qualified health practitioners" broadly defined as professionals including psychiatrists, psychologists, social workers, nurses with mental health specialization, other licensed or master's-prepared mental health counselors, mental health trainees (e.g., psychology practicum students, psychiatric residents, or other mental health professional trainees), mental health workers, or case managers.

Additionally, studies from the clinical supervision literature were included if they satisfied the following criteria: (a) focused on case-based clinical supervision (studies involving primarily one-time training workshops or staff development activities were excluded, as they are typically short-term and not case-focused); (b) were published in a peer-reviewed scientific journal in English over the past 30 years; (c) involved a direct mental health measure of client outcome, including an assessment of symptoms, clinical problems, psychosocial functioning, or quality of life, as determined by self-report or objective assessments (studies that solely examined client satisfaction or the working alliance were excluded); (d) involved a mental health service (we therefore excluded a number of studies in the area of developmental and learning disabilities from past reviews; Milne & James, 2000); (e) included only studies in which inference was possible as to the positive effect of clinical supervision on client outcomes (i.e., studies reporting primarily quantitative data and including experimental designs, although we allowed for certain other designs, including mixed-effect regression models); and (f) were directly relevant to real-world clinical practice (thus, we excluded volunteers, pseudoclients, simulated role-plays, and so forth).

Sample of Studies and Search Criteria

Our search strategy involved sifting through previous major reviews of clinical supervision in handbooks or journal articles up through the Watkins (2011) review. We then conducted our own literature search using the following databases: PsycINFO, PsycARTICLES, PsycBOOKS, Psychology and Behavioral Sciences Collection, Academic Search Premier, and MEDLINE. We used the keywords "psychotherapy supervision outcomes," "clinical supervision outcomes," and "supervision outcomes" to look for any additional references to studies. We then contacted each of the lead researchers, asking them to provide any additional references.

The final selection of studies was determined by balancing the need for a sufficiently broad review sample in order to maintain generalizability against the danger of overinclusion of studies that might conflate findings with very divergent populations and methods of supervision or its evaluation. By applying these exclusion criteria to the initial search sample of 48 studies, and taking account of duplicate studies in the sample, our final sample consisted of 12 studies. This narrowing of selection is consistent with Watkins (2011), who noted in an similar review that 30% of studies selected for inclusion in previous reviews of supervision and client outcomes did not actually directly assess client outcomes as a dependent variable. Our search terms were applied to psychotherapy supervision in general and did not restrict the type of study by treatment type or orientation. However, the final sample included a disproportionate number of studies involving supervision of cognitive–behavioral therapy (CBT). It is likely that our inclusion criteria—and especially our requirement that studies include an outcome measure directly related to symptoms, functioning, and quality of life—may have biased the sample toward studies with a CBT focus. In addition, CBT supervision lends itself to a manualized approach and evidence-based practice.

Review and Coding Procedures Based on the Fidelity Framework

In the next step, we reviewed our final sample of 12 studies using the fidelity framework criteria drawn directly from Borrelli et al. (2005) and supplemented by selected criteria from Roth et al. (2010), as noted above. Our plan was to review the 12 studies against a well-defined and conceptually coherent set of minimum standards, comprising the five elements of the fidelity framework, in a reliable and replicable manner, as per a checklist or audit approach.

Development of Criteria for Fidelity Framework Checklist

The first author initially reviewed the 12 studies against the checklist standards, recording compliance with the standards set out in Table 1. The authors then independently coded four of the 12 studies (33%), well above the recommended proportion of 10% (Borrelli et al., 2005). Two of the randomly selected studies were the first ones coded by the first author, and the other two were the ones last coded by him. This was to ensure reliable coding and to address potential reviewer drift, respectively. We obtained exact percent agreement between the two coders of 90% and 80% for these two assessment stages, respectively. Disagreements were reviewed for the purpose of improving interrater reliability and identifying any areas in the criteria table that needed to be more clearly defined. These assessments indicated that the fidelity checklist (see Table 1) could be applied reliably over the review period.

Results

Reporting of Treatment Fidelity Strategies

The main results of our review are presented in Table 2, which summarizes findings in terms of study fidelity for each of the five dimensions of the fidelity framework. Because we defined enactment as including a direct assessment of client outcomes, a key study inclusion criterion, all studies by definition met this requirement.

Design. Apart from enactment, the highest level of adherence to the fidelity framework was in the design category, in which 83%

Table 2
Studies of Clinical Supervision and Client Outcome, Reviewed Using the Fidelity Framework (Borrelli et al., 2005)

Study	Date	Design	Training	Delivery	Receipt	Enactment	Percent
Dodenhoff (1981)	1981	1[a]	0[b]	1	0	1	60%
Couchon & Bernard (1984)	1984	0	0	1	0	1	40%
Steinhelber, Patterson, Cliffe, & LeGoullon (1984)	1984	1	0	0	0	1	40%
Harkness & Hensley (1991)	1991	0	1	1	0	1	60%
Triantafillou (1997)	1997	1	1	1	1	1	100%
Bambling, King, Raue, Schweitzer, & Lambert (2006)	2006	1	1	1	1	1	100%
Bradshaw, Butterworth, & Mairs (2007)	2007	1	1	1	0	1	80%
Grey, Salkovskis, Quigley, Clark, & Ehlers (2008)	2008	1	0	0	1	1	60%
Callahan, Almstrom, Swift, Borja, & Heath (2009)	2009	1	0	0	0	1	40%
Reese et al. (2009)	2009	1	0	0	0	1	40%
Schoenwald, Sheidow, & Chapman (2009)	2009	1	1	1	1	1	100%
White & Winstanley (2010)	2010	1	1	1	0	1	80%
Percent of total studies meeting fidelity checklist criteria		83%	50%	67%	33%	100%	67%

[a] Study met fidelity checklist criteria. [b] Study did not meet fidelity checklist criteria.

of the studies included in our sample met criteria for adherence. For the purposes of our review, a key criterion for design fidelity was reference to a theoretical model or clinical guidelines (Borrelli et al., 2005, p. 858, Table 1). Borrowing from the Roth et al. (2010, p. 297) recommendations, we required that a supervision manual or guideline be either in the public domain or outlined within the text. We did not consider specification of contact time, length, and duration of supervision alone to be sufficient to operationalize the dependent variable in these studies. This level of adherence to the design aspect of the fidelity framework was roughly comparable with the mean of .80 reported by Borrelli et al. (2005) in their review of 342 articles. In the area of adherence to design, there appeared to be a positive trend over time, with more articles in the past 10 years demonstrating good adherence to design.

Training. Lower levels of adherence were identified in relation to the training of supervisors, with only 50% of studies adequately specifying elements of training. This level of adherence compared favorably with the findings in the Borrelli et al. (2005) study that reported a mean adherence level of .22 in the category of training.

Delivery. In terms of the delivery of supervision, 67% of studies met minimum standards of adherence. This compares quite favorably with the findings in Borrelli et al. (2005), which indicated a mean adherence level of .35 in this category. There was a slight trend toward improvement over time, but it should be noted that three out of seven studies in the past 10 years (43%) failed to meet minimum criteria in our fidelity checklist review approach.

Receipt. The lowest levels of adherence were in the receipt of supervision, with only 33% of studies meeting the criteria of adequate reporting of any assessment of the trainees' acquisition of skills or improved comprehension based on supervision. In the design element of receipt, our bar was set relatively low and we included any direct assessment of the trainee's knowledge, skills, or attitudes related to supervision training, only excluding assessments of trainee satisfaction or the supervisory alliance. This compares unfavorably with the findings in Borrelli et al. (2005), who reported a mean adherence level of .49 in the category of receipt. On this dimension, there was no clear trend toward improvement over time, and four out of seven studies over the past 10 years (57%) failed to meet the standard specified in our fidelity checklist approach.

Summary of Results

In summary, across all 12 studies of clinical supervision and client outcome reviewed, there was a mean overall adherence level of .67 to the treatment fidelity framework. The weakest areas of adherence were in the elements of receipt (.33) and training (.50), followed by delivery (.67) of supervision. This is consistent with findings in the Roth et al. (2010) review of clinical trials, in which they had difficulty identifying methods of training and supervision in the clinical trials literature. It is also broadly consistent with Borrelli et al. (2005), who reported low levels of adherence in training (.22), delivery (.35), and receipt (.49) (Borrelli et al., 2005, p. 857).

Discussion

We devised a fidelity assessment tool, based on Borrelli et al. (2005), applying it reliably to 12 studies relating clinical supervision to client outcomes, in order to assess treatment fidelity practices in the supervision literature. To our knowledge, this is the first review that applies this treatment fidelity approach specifically to studies of clinical supervision. Our analysis indicated moderately good adherence overall to the treatment fidelity framework (.67). The definition of high treatment fidelity used by Borrelli et al. (2005) was "studies that had a .80 or greater proportion adherence to our checklist across all strategies" (Borrelli et al., 2005, p. 857). The weakest areas of adherence were in the elements of receipt (.33) and training (.50), followed by delivery (.67) of supervision. There was a general improvement over time, but only three out of 12 studies (25%) met 100% of the fidelity framework adherence benchmarks: Triantafillou (1997), Bambling, King, Raue, Schweitzer, and Lambert (2006), and Schoenwald, Sheidow, and Chapman (2009).

Are these findings surprising? In some ways, these results are quite consistent with prior reviews of workshop-based training (Culloty, Milne, & Sheikh, 2010) and supervision in clinical trials (Roth et al.,

2010). Culloty et al. (2010) concluded their review of the CBT training workshop literature by stating that "fidelity is rarely treated comprehensively" (Culloty et al., 2010, p. 135). Similarly, Roth et al. (2010) concluded that "information about training and supervision is presented inconsistently from paper to paper, and sometimes only in outline form" (p. 296). It is disappointing that there appears to be a continuing lack of adequate specification both in defining the elements of training and in including direct assessments of the changes in the trainees' knowledge, skills, and attitudes. It is our view that this gap reflects the fact that we are still in the early stages of disseminating and operationalizing the competency-based model in supervision (Falender & Shafranske, 2010), and it may also reflect the emphasis on the acid test of clinical outcomes. In short, practice is still considering new conceptualizations of outcome and catching up with the higher levels of specification (e.g., Roth & Pilling, 2010). The lower levels of fidelity in training and receipt we identified reflect a disturbing deficiency in specifying, implementing, and documenting training practices. In our view, this, in turn, reflects the lack of progress in manualizing and standardizing supervisor training.

The fact that only 25% of studies met 100% of the fidelity framework criteria suggests a continuing and significant gap in the design of supervision studies. This raises the question as to whether we have set the bar too high by specifying that high-quality studies meet four of five areas of the fidelity checklist. By insisting that studies pay attention to all five criteria in the fidelity model, we are providing the best platform for future research. As noted in Borrelli et al. (2005), there are significant negative consequences associated with low fidelity studies: "Treatment fidelity prevents the premature rejection of treatments that could be effective as well as the acceptance of treatments that are nonreproducible because of low internal validity" (Borrelli et al., 2005, p. 858).

Although our review has focused on the methodological soundness of 12 studies of client outcomes of supervision by reference to the fidelity framework, we would like to briefly consider what these studies tell us about client outcomes. Despite our attempt to include only the most relevant studies, one subgroup of these studies actually offered no direct information on supervision's clinical outcomes as we defined them. For example, Couchon and Bernard (1984) concluded that the timing of supervision sessions was not significantly correlated with improved client outcomes, while White and Winstanley (2009) concluded that supervision did not contribute to the quality of clinical care or client satisfaction. A second subgroup of studies did report significant clinical outcomes, including Dodenhoff (1981), Steinhelber, Patterson, Cliffe, and LeGoullon (1984), Harkness and Hensley (1991), and Triantafillou (1997). However, they were methodologically compromised and were thus hard to interpret (e.g., use of nonstandard measures; confounding of the experimental and control conditions).

This left us with a third subgroup for which relevant clinical outcomes were measured and that were interpretable studies. For example, Grey, Salkovskis, Quigley, Clark, and Ehlers (2008) determined that therapists who received ongoing clinical supervision (compared with treatment as usual) achieved significantly better outcomes for panic disorder in terms of panic attacks and improvements in self-rated anxiety and avoidance. Similarly clear, interpretable outcomes were reported by Bambling et al. (2006), Bradshaw, Butterworth, and Mairs (2007), Callahan, Almstrom, Swift, Borja, and Heath (2009), and Schoenwald et al. (2009). Although these studies do not clarify precise causal mechanisms (i.e., which aspects of supervision actually contributed to client outcomes), there is reason to believe that supervision that improves adherence to an empirically supported protocol appears likely to improve client outcomes. Second, we can conclude that supervision is likely to outperform no supervision in terms of client outcomes. Future research will hopefully illuminate which specific aspects of supervision are associated with the most significant client improvements.

Recommendations

We next draw on helpful examples from studies that were identified in our search but that were excluded from our final sample. They nonetheless afford outstanding illustrations of how one might address fidelity framework criteria. Following the approach used within Roth et al. (2010) and Ellis and Ladany (1997), we seek to highlight good practice and to suggest options for advancing research. Table 3 again summarizes the fidelity framework, this time set alongside a selection of studies that illustrate both the wide range of outcomes that have been measured and some innovative or best-practice methods for achieving those outcomes.

The recommendations that follow from Table 3 include the need for an explicit conceptualization of supervision as noted in the introduction, providing a falsifiable model of the predicted relationship between supervision and its outcomes. This brings significant benefits, such as detailed guidance on the precise nature of the intervention, together with some prioritization of the expected effects (Chen, 1990). In the illustrative study within Table 3, Weingardt, Cucciare, Bellotti, and Lai (2009) specified a "blended learning" system, including online group supervision sessions. Their precision and use of the latest available technology are features that will help to increase the fidelity with which supervision is implemented.

On the subject of training, the explicit application of the fidelity framework by Culloty et al. (2010) included attention to the "mini-impacts" on the supervisees' experiential learning, a feature that draws attention to the logical possibility of introducing subcriteria within the framework. For example, if the model guiding Weingardt et al.'s (2009) blended learning system includes intermediate steps, these could advantageously be assessed to better pinpoint the strengths and weaknesses of such methods. This is conveniently illustrated in the study in Table 3 by Karlin et al. (2012), in their consideration of how their package had impacted supervisees' self-efficacy and attitudes. These assessments were completed at several intervals: pre- and postworkshop training, after 6 months of consultation, and in a follow-up after the end of the training program, providing detailed stepwise knowledge of delivery and receipt of training at key stages of the program. The illustration from Westbrook, Sedgewick-Taylor, Bennett-Levy, Butler, and McManus (2008) goes into exceptional detail in linking the elements of training to a suitably diverse range of outcome measures. This brings into focus how longitudinal measurement of the key learning opportunities within supervision might enhance outcomes. In this sense, the current attention to outcome monitoring could usefully be extended to any of the other intermediate outcomes, as defined within the study's model (e.g., in Westbrook et al., 2008, the effectiveness of experiential exercises in enhancing supervisees' competence). This kind of nested approach to

Table 3
Applying the Fidelity Framework to the Evaluation of Outcomes From Clinical Supervision, With Illustrative Methods and Studies

Fidelity dimension	Outcomes measured	Illustrative supervision study
1. Design (addresses the question "What is the right thing to do?")	Knowledge gain Self-efficacy Burnout	Weingardt, Cucciare, Bellotti, & Lai (2009): Designed a "blended learning" system, featuring e-learning with a self-paced, online course in CBT plus a series of live, online virtual group supervision sessions, designed to promote adherence.
2. Training (addresses the question "Has the right thing been done?")	Competence of the trainer Satisfaction with training Transfer of training to supervision	Culloty, Milne, & Sheikh (2010): Adherence to supervisor training model was assessed by direct observation, questionnaire, and group interview, capturing the supervisees' perceptions of the training and its transfer. This measured training adherence, the associated mini-impacts on the supervisees' experiential learning, and generalization.
3. Delivery (addresses the question "Has it been done right?")	Satisfaction with training Supervisees' self-efficacy, attitudes, behavioral intentions, competence Therapeutic alliance Quality of life and patient outcomes Maintenance of training/supervision effects	Karlin et al. (2012): In addition to evaluating changes in supervisees' competence and patient outcomes at several intervals within a training plus consultation package, this study considered how the package had impacted supervisees' self-efficacy and attitudes to the therapy (CBT: i.e., "nonspecific supervision effects").
4. Receipt (addresses the question "Did it result in the right outcome?")	Dropout from training Satisfaction with training Satisfaction with supervision Supervisees' competence Patient outcomes	Westbrook, Sedgewick-Taylor, Bennett-Levy, Butler, & McManus (2008): Supervisees' discussed tapes of selected patients' sessions in supervision meetings. This linked training workshops focused on helping trainees acquire clinical skills as well as understanding theory by practicing skills through role-plays and other experiential exercises.
5. Enactment (addresses the question "Did it result in the right impact?")	Supervisees' self-rated learning patient dropouts Clinical outcomes (panic severity ratings; avoidance; agoraphobic cognitions; ratings of general anxiety)	Grey, Salkovskis, Quigley, Clark, & Ehlers (2008): Supervisees' clinical outcomes in treating panic disorder were compared before and after training, including ongoing supervision. Significantly better outcomes achieved after training plus supervision, indicating successful dissemination of an approach used within clinical trials to primary care.

Note. Adapted from Borrelli et al. (2005). CBT = cognitive–behavioral therapy.

monitoring and feedback is illustrated in the study by Grey et al. (2008), as, in monitoring agoraphobic cognitions, they measured at least one of their hypothesized mechanisms of change. This degree of precise conceptualization and monitoring helps to improve outcomes or to indicate that the assumed links are invalid and require improvement (Chen, 1990).

Limitations of the Present Review

Although we made every effort to select a representative sample of studies that met our criteria, by selecting from 48 articles cited in several major past reviews, our exclusion criteria may have been overly restrictive and may thus limit the generalizability of our findings. It is possible that treatment fidelity is misrepresented in our study sample, as we chose studies in a relatively exclusionary fashion, eliminating studies that did not focus on mental health services and studies that did not consider client outcomes in terms of assessing symptoms, problems, functioning, and quality of life as the dependent variable. Similarly, we fully recognize that client outcomes are not the only relevant criterion for determining the sample of studies for review. In theory, one might, for example, start with any of the other outcomes listed within Table 3, which might well produce results more consistent with the fidelity framework. Our inclusion criteria also meant that studies that addressed fidelity (or related issues) using qualitative methods were excluded. We are aware that some highly relevant qualitative studies have been published, and these also merit consideration in taking stock of what we know about supervision's outcomes. For instance, Johnston and Milne (2012) explored the "receipt" dimension of fidelity with grounded theory, highlighting how learning outcomes were enhanced when supervisors encouraged valued processes, such as Socratic information exchange.

In developing our fidelity checklist review system, we operationalized a well-established framework with only minor changes (e.g., supervisor substituted for therapist; trainee substituted for client). Future reviews might usefully extend our present analysis to include other criteria. For example, evaluation of the training and education of mental health professionals has often drawn on Kirkpatrick's (1967) four consecutive levels of effectiveness: reactions, learning, transfer, and impact. To these might be added the prior need for participation (Belfield, Thomas, Bullock, Eynon, & Wall, 2001). Whatever criteria are used, it is important to check that their application is reliable, and we independently coded over 30% of our final study sample. However, only the two authors acted as raters, and without an independent rater who is blind to the study hypotheses, it is possible that our ratings were subject to bias.

Conclusions

We have questioned the wisdom of treating client outcomes as the acid test of clinical supervision, arguing that there are good grounds for giving equal weight to a number of complementary outcome criteria. These are conveniently and systematically captured within the fidelity framework. Application of this framework to 12 carefully selected studies of clinical supervision highlighted significant infidelity, which hampers progress. To assist future research, we have developed a way of assessing adherence to this

framework, and we have recommended ways to enhance supervision research, consistent with the fidelity framework. This framework appears to offer a useful way to progress toward the long overdue need for proof of supervision's effectiveness.

References

American Psychological Association. (2006). Evidence-based practice in psychology: APA presidential task force on evidence-based practice. *American Psychologist, 61,* 475–476.

Bambling, M., King, R., Raue, P., Schweitzer, R., & Lambert, W. (2006). Clinical supervision: Its influence on client-rated working alliance and client symptom reduction in the brief treatment of depression. *Psychotherapy Research, 16,* 317–331. doi:10.1080/10503300500268524

Belfield, C., Thomas, H., Bullock, A., Eynon, R., & Wall, D. (2001). Measuring effectiveness for best evidence medical education: A discussion. *Medical Teacher, 23,* 164–170. doi:10.1080/0142150020031084

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., . . . Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology, 23,* 443–451. doi:10.1037/0278-6133.23.5.443

Bernard, J. M., Goodyear, R. K., & Bernard, J. M. (1992). *Fundamentals of clinical supervision*. Boston: Allyn and Bacon.

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Greger, R., DeFrancesco, C., . . . Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology, 73,* 852–860. doi:10.1037/0022-006X.73.5.852

Bradshaw, T., Butterworth, A., & Mairs, H. (2007). Does structured clinical supervision during psychosocial intervention education enhance outcome for mental health nurses and the service users they work with? *Journal of Psychiatric and Mental Health Nursing, 14,* 4–12. doi:10.1111/j.1365-2850.2007.01021.x

Callahan, J. L., Almstrom, C. M., Swift, J. K., Borja, S. E., & Heath, C. J. (2009). Exploring the contribution of supervisors to intervention outcomes. *Training and Education in Professional Psychology, 3,* 72–77. doi:10.1037/a0014294

Chen, H. (1990). *Theory-driven evaluation*. Newbury Park, CA: Sage.

Couchon, W. D., & Bernard, J. M. (1984). Effects of timing of supervision on supervisor and counselor performance. *The Clinical Supervisor, 2,* 3–20. doi:10.1300/J001v02n03_02

Culloty, T., Milne, D. L., & Sheikh, A. I. (2010). Evaluating the training of clinical supervisors: A pilot study using the fidelity framework. *The Cognitive Behaviour Therapist, 3,* 132–144. doi:10.1017/S1754470X10000139

Dodenhoff, J. T. (1981). Interpersonal attraction and direct–indirect supervisor influence as predictors of counselor trainee effectiveness. *Journal of Counseling Psychology, 28,* 47–52. doi:10.1037/0022-0167.28.1.47

Donabedian, A. (1988). The quality of care: How can it be assessed? *Journal of the American Medical Association, 260,* 1743–1748. doi:10.1001/jama.1988.03410120089033

Ellis, M., & Ladany, N. (1997). Inferences concerning supervisees and clients in clinical supervision: An integrative review. In C. E. Watkins (Ed.), *The handbook of psychotherapy supervision* (pp. 447–507). Chichester, UK: Wiley.

Ellis, M. V., Ladany, N., Krengel, M., & Schult, D. (1996). Clinical supervision research from 1981–1993: A methodological critic. *Journal of Counseling Psychology, 43,* 35–50. doi:10.1037/0022-0167.43.1.35

Falender, C. A., Cornish, J. A. E., Goodyear, R., Hatcher, R., Kaslow, N. J., Leventhal, G., . . . Grus, C. (2004). Defining competencies in psychology supervision: A consensus statement. *Journal of Clinical Psychology, 60,* 771–785. doi:10.1002/jclp.20013

Falender, C. A., & Shafranske, E. P. (Eds.). (2008). *Casebook for clinical supervision: A competency-based approach*. Washington DC: American Psychological Association. doi:10.1037/11792-000

Falender, C. A., & Shafranske, E. P. (2010). Psychotherapy-based supervision models in an emerging competency-based area: A commentary. *Psychotherapy: Theory, Research, Practice, Training, 47,* 45–50. doi:10.1037/a0018873

Freitas, G. J. (2002). The impact of psychotherapy supervision on client outcome: A critical examination of two decades of research. *Psychotherapy: Theory, Research, Practice, Training, 39,* 354–367. doi:10.1037/0033-3204.39.4.354

Grey, N., Salkovskis, P., Quigley, A., Clark, D. M., & Ehlers, A. (2008). Dissemination of cognitive therapy for panic disorder in primary care. *Behavioural and Cognitive Psychotherapy, 36,* 509–520. doi:10.1017/S1352465808004694

Harkness, D., & Hensley, H. (1991). Changing the focus of social work supervision: Effects on client satisfaction and generalized contentment. *Social Work, 36,* 506–512.

Holloway, E. L. (1984). Outcome evaluation in supervision research. *The Counseling Psychologist, 12,* 167–174. doi:10.1177/0011000084124014

Holloway, E. L., & Neufeldt, S. A. (1995). Supervision: Its contribution to treatment efficacy. *Journal of Consulting and Clinical Psychology, 63,* 207–213. doi:10.1037/0022-006X.63.2.207

Johnston, L. H., & Milne, D. L. (2012). How do supervisees' learn during supervision? A grounded theory study of the perceived developmental process. *The Cognitive Behaviour Therapist, 5,* 1–23. doi:10.1017/S1754470X12000013

Karlin, B. E., Brown, G. K., Trockel, M., Cunning, D., Zeiss, A. M., & Taylor, C. B. (2012). National dissemination of cognitive behavioral therapy for depression in the Department of Veterans Affairs health care system: Therapist and patient-level outcomes. *Journal of Consulting and Clinical Psychology, 80,* 707–718. doi:10.1037/a0029328

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and development handbook* (pp. 87–112). New York, NY: McGraw- Hill.

Milne, D. L. (2007). An empirical definition of clinical supervision. *British Journal of Clinical Psychology, 46,* 437–447. doi:10.1348/014466507X197415

Milne, D. L. (2009). *Evidence-based clinical supervision: Principles and practice*. Chichester, UK: BPS Blackwell.

Milne, D. L. (in press). Beyond the acid test: A conceptual review of outcome evaluation in clinical supervision. *American Journal of Psychotherapy*.

Milne, D. L., Aylott, H., Fitzpatrick, H., & Ellis, M. V. (2008). How does clinical supervision work? Using a best evidence synthesis approach to construct a basic model of supervision. *The Clinical Supervisor, 27,* 170–190. doi:10.1080/07325220802487915

Milne, D. L., & James, I. (2000). A systematic review of effective cognitive behavioral supervision. *British Journal of Clinical Psychology, 39,* 111–127. doi:10.1348/014466500163149

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247–266. doi:10.1016/0272-7358(91)90103-2

Reese, R. J. Usher, E. L., Bowman, D. C., Norsworthy, L. A., Halstead, J. L., Rowlands, S. R., & Chisholm, R. R. (2009). Using client feedback in psychotherapy training: An analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology, 3,* 157–168. doi:10.1037/a0015673

Rodolfa, E., Greenberg, S., Hunsley, J., Smith-Zoeller, M., Cox, D., Sammons, M., . . . Spivak, H. (2013). A competency model for the practice of psychology. *Training and Education in Professional Psychology, 7,* 71–83. doi:10.1037/a0032415

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (2003). *Evaluation: A systematic approach* (7th ed.). Newbury Park, CA: Sage.

Roth, A. D., Pilling, S., & Turner, J. (2010). Therapist training and supervision in clinical trials: Implications for clinical practice. *Behavioural and Cognitive Psychotherapy, 38,* 291–302. doi:10.1017/S1352465810000068

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health, 38,* 32–43. doi:10.1007/s10488-010-0321-0

Schoenwald, S. K., Sheidow, A. J., & Chapman, J. E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology, 77,* 410–421. doi:10.1037/a0013788

Steinhelber, J., Patterson, V., Cliffe, K., & LeGoullon, M. (1984). An investigation of some relationships between psychotherapy supervision and patient change. *Journal of Clinical Psychology, 40,* 1346–1353. doi:10.1002/1097-4679(198411)40:6<1346::AID-JCLP2270400612>3.0.CO;2-L

Triantafillou, N. (1997). A solution-focused approach to mental health supervision. *Journal of Systemic Therapies, 16,* 305–328.

Tsui, M.-S. (1997). Empirical research on social work supervision: The state of the art (1970–1995). *Journal of Social Service Research, 23,* 39–54. doi:10.1300/J079v23n02_03

Watkins, C. E. (2011). Does psychotherapy supervision contribute to patient outcomes? Considering thirty years of research. *The Clinical Supervisor, 30,* 235–256. doi:10.1080/07325223.2011.619417

Watkins, C. E., & Milne, D. L. (in press). *International handbook of clinical supervision*. Chichester, UK: Wiley.

Weingardt, K. R., Cucciare, M. A., Bellotti, C., & Lai, W. P. (2009). A randomized trial comparing two models of web-based training in cognitive–behavioral therapy for substance abuse counselors. *Journal of Substance Abuse Treatment, 37,* 219–227. doi:10.1016/j.jsat.2009.01.002

Westbrook, D., Sedgwick-Taylor, A., Bennett-Levy, J., Butler, G., & McManus, F. (2008). A pilot evaluation of a brief CBT training course: Impact on trainees' satisfaction, clinical skills and patient outcomes. *Behavioural and Cognitive Psychotherapy, 36,* 569–579. doi:10.1017/S1352465808004608

Wheeler, S., & Richards, K. (2007). The impact of clinical supervision on counsellors and therapists, their practice and their clients. A systematic review of the literature. *Counselling & Psychotherapy Research, 7,* 54–65. doi:10.1080/14733140601185274

White, E., & Winstanley, J. (2009). Clinical supervision for nurses working in mental health settings in Queensland, Australia: A randomised controlled trial in progress and emergent challenges. *Journal of Research in Nursing, 14,* 263–276. doi:10.1177/1744987108101612

Received July 1, 2013
Revision received October 14, 2013
Accepted October 22, 2013 ■