The Variables Problem and Progress in Psychotherapy Research

William B. Stiles
Glendale Springs, North Carolina

In this journal’s first article, Strupp (1963) pointed to problems specifying independent and dependent variables as a source of slow progress in psychotherapy outcome research. This commentary agrees, shows how the concept of variable loses its meaning in psychotherapy research because of participants’ responsiveness, and notes an alternative research strategy that does not depend on variables.

Keywords: psychotherapy, responsiveness, variables problem, theory-building qualitative research

Hans Strupp (1963) began this journal’s first article by pointing to “a realization on the part of researchers that a new approach to the [outcome] issue must be found and that more pressing matters must be dealt with first before we can address ourselves meaningfully to the question of the effectiveness of psychotherapy” (p. 1). Research sophistication, instrumentation, and the psychotherapy research literature have grown enormously in the intervening 50 years, but I think this remains an apt characterization.

Strupp devoted the first half of his article to analyzing the problem of variables in psychotherapy outcome research: specifying what is meant by treatment and outcome, the independent variable and the dependent variable, respectively. The logic of the experiment is that if all previous conditions except one (the independent variable) are held constant (controlled), any differences in the outcome (the dependent variable) must have been caused by the one condition that varied. For example, if one client is given psychoanalysis and another identical client is not, but is treated identically in all other respects, any differences in their outcomes must have been caused by the psychoanalysis. Difficulties arise because no two people are identical and because it is impossible to treat two people identically in all respects except one.

Randomized controlled trials (RCTs) are an adaptation of the experimental method that attempts to address these uncontrolled variations statistically. Rather than comparing individual patients, investigators using RCTs randomly assign patients to groups that receive the different levels of the variable—different treatments—on the assumption that any previous differences that might affect the outcomes will be more or less evenly distributed across the groups. Although individuals’ outcomes might vary within groups (because clients are not identical), any mean differences between groups beyond those due to chance should be attributable to the different treatments. Insofar as the controlled experiment is the closest science has come to a means for demonstrating causality, the effort put into RCTs seems worthwhile despite the difficulties.

In his reply to Strupp (1963), Eysenck (1964) asserted, as he had in his earlier review (Eysenck, 1952), “that no data existed disproving the null hypothesis [that psychotherapy fails to facilitate recovery from neurotic disorder] scientifically. There is only one way to answer such an argument and that is to point to an experimental study or investigation conclusively disproving the null hypothesis. Strupp . . . fails to adduce a single study disproving my original conclusion” (Eysenck, 1964, p. 97). Strupp did not adduce such a study in his rejoinder either (Strupp, 1964).

Fifty years on, there have been enough supportive studies, reviews, and meta-analyses that psychotherapy’s effectiveness is widely, if not universally, accepted (American Psychological Association, 2012). Perhaps even Eysenck would have accepted some of the published RCTs as experimental studies disproving his 1952 conclusion. In my opinion, however, the problems of the independent and dependent variables that Strupp (1963) identified continue to interfere hugely with scientific progress in psychotherapy research.

By scientific progress, I mean improving explanatory theories of how psychotherapy works, making them more general, more precise, and more realistic. I think this is what Strupp had in mind in writing “more pressing matters must be dealt with first before we can address ourselves meaningfully to the question of the effectiveness of psychotherapy.” In the following pages, I (a) describe the phenomenon of responsiveness and how it contributes to the variables problem, (b) review some results of research in which I have been involved and which, I think, illustrates some manifestations of the variables problem while nevertheless supporting the effectiveness of psychotherapy, and (c) point to an alternative approach to psychotherapy research that could, I think, facilitate scientific progress.

Responsiveness and Independent Variables

Responsiveness refers to behavior being influenced by emerging context (Stiles, Honos-Webb, & Surko, 1998). People respond to what happens around them, such as what other people do. Human behavior is responsive on time scales that range from months to milliseconds. Paying attention and being polite are responsive. If I answer your questions or repeat something when you look puzzled, I am being responsive. Psychotherapists’ skills and goals involve responsiveness. What therapists do depends on what clients do. Examples include assigning clients to treatments based on their presenting problems, planning their treatment based on how they are progressing.

I thank Tim Carey, Mikael Leiman, and Katerine Osatuke for comments on a draft of this article.

Correspondence concerning this article should be addressed to William B. Stiles, P.O. Box 27, Glendale Springs, NC 28629. E-mail: stileswb@muohio.edu
active listening, timing, staying on topic, turn-taking, attentiveness, adjusting interventions already in progress, and so forth.

Responsiveness is a neutral term, and responsiveness is not necessarily benign. However, therapists and clients generally have benign goals, and they respond to advance those goals in ways consistent with their theoretical and personal principles. This can be called appropriate responsiveness. So, appropriate responsiveness means do the right thing. In general, therapists and clients try to do the right thing.

Strupp (1963) wrote, “One of the major difficulties in psychotherapy research is that of adequately specifying the independent variable—the psychotherapeutic methods—to which therapeutic changes are being attributed” (p. 1). To illustrate, he drew from a list of potentially impactful characteristics by Knight (1941), which included the sort of influence attempted (e.g., suggestion, persuasion, exhortation, intimidation, counseling, interpretation, reeducation, retraining), the general aim (e.g., supportive, suppressive, expressive, cathartic, ventilative), the depth of the work, treatment duration, the theoretical approach, and the variant within that approach.

Such lists highlight but nevertheless understate the problem. The pervasiveness of human responsiveness implies that clients in the same experimental condition of an outcome study each receives a different individually tailored treatment. Such variability impairs the general aim (e.g., supportive, suppressive, expressive, cathartic, ventilative), the depth of the work, treatment duration, the theoretical approach, and the variant within that approach.

In the 1980s, in response to concerns about this lack of standardization, outcome researchers began using manuals to describe treatments, an innovation described as “a small revolution” (Luborsky & DeRubeis, 1984). However, manuals do not provide rigid instructions. Instead, they describe repertoires of interventions and the sorts of situations in which the interventions might be used. They emphasize building rapport, appropriate clinical judgment, timing, tact, and adapting the approach to how clients present. In other words, they prescribe appropriate therapist responsiveness within the theoretical approach. Likewise, clients are not passive recipients but active participants charged with making sense of psychotherapeutic activities and adapting what they gain to the context of their own lives. I see great value in the sort of fine-grained description of a treatment approach manuals provide, but they do not overcome responsive individualization of treatments. The standardization implied by citing manuals to define levels of an experimental independent variable is at best illusory and at worst deceptive.

Of course, no two things in the universe are exactly alike; even two tablets of an antidepressant medication have differences. But the dissimilarity of two sessions of emotion-focused therapy is, I suggest, orders of magnitude greater than the dissimilarity of two fluoxetine tablets. Moreover, whereas the differences among tablets are more or less randomly distributed, the tailoring is responsive to the requirements of the client and the context. We deceive ourselves when we suppose that two courses of manualized emotion-focused therapy, interpersonal therapy, or cognitive therapy are alike in the same way as two tablets of fluoxetine, buproprion, or sertraline.

The individualization achieved through responsiveness is clinically and humanly appropriate (a truly standardized script for interventions would be clinically absurd and probably unethical), but it is disastrous for the logic of experiments. Responsiveness implies that the independent variable in outcome research is causally entangled with client, therapist, and context variables (Elkin, 1999; Krause & Lutz, 2009), so its meaning changes systematically across clients and contexts. For example, clients with overinvolved attachment styles are treated systematically differently than clients with underinvolved attachment styles, even when treated by the same therapists in the same experimental condition (Hardy, Stiles, Barkham, & Startup, 1998). Even worse, the independent variable is causally entangled with dependent variables. Therapists responsively adjust their treatment in light of client progress or lack of progress. Such causal entanglements are more serious than uncontrolled variability in extraneous conditions because they cannot be overcome by randomization (Stiles, 2009b).

Evaluation and Dependent Variables

To introduce the problems with dependent variables, Strupp (1963) cited “an insightful and lucid article” by Holt (1958) regarding a “hidden trick” in global judgments of outcome. Such judgments may seem satisfying, but they are scientifically empty. “As long as one relies on global clinical judgments, like outcome,” Strupp summarized, “one substitutes something for real information” (p. 5). Global judgments are problematic when they are drawn from symptom intensity inventories as well as when they are drawn from clinicians’ impressions. Global indexes of severity or distress, total scores, summary scores, and the like similarly substitute evaluation for specific information.

Evaluative measures are robust and popular, I think, because evaluation is universal, a common denominator in people’s diverse understandings. Rogers (1959) described this as the organismic valuing process. Therapists, observers, clients, and readers have firm convictions about what they think is good, how they feel, and what they like. Readers want to know how well a therapy works, so research using evaluative measures has a receptive audience. But such evaluations can short-circuit understanding. A focus on evaluating a treatment’s outcome avoids the much more difficult theoretical accounting for descriptive results. Condensing therapeutic effects and therapeutic relationships into global evaluative dimensions does violence to the theoretically specific conceptualizations that scientific research is meant to test (Stiles, 2006, 2009a).

Even measures that nominally assess specific conditions, such as depression, largely reflect global distress and are at least sometimes closely correlated with more global measures (e.g., Leach et al., 2006). Unless all respondents understand and use terms and concepts in the same way, they do not emerge as common components in rating scales. This is particularly salient in the case of client self-report indexes; clients are unlikely to have a common understanding of theoretical concepts, so most of the commonality in their responses depends on the evaluative component in the items’ meaning. As a result, most internally consistent self-report measures of process and outcome are primarily evaluative.

Inventories that code specific symptoms and behaviors, applied by experts trained to distinguish them, can avoid this problem. In principle, such data can be used to investigate specific changes in individuals. Much of the improvement is undone, however, when the ratings are aggregated into global indexes to compare across...
clients or to address questions posed in the evaluative terms of outcome or the value of a treatment.

Most process or relationship variables on the list of “evidence-based” “effective elements of therapeutic relationships” assembled by Norcross (2002, 2011) are evaluative: alliance, group cohesion, empathy, goal consensus and collaboration, positive regard, congruence/genuineness, repair of ruptures, management of countertransference, and quality of relational interpretations. These elements represent achievements or desired results rather than specific conditions or volitional behaviors (Stiles & Wolfe, 2006). Such evaluative process variables incorporate appropriate responsiveness. They reflect whether the behavior is appropriate to the circumstances rather than describing specific behaviors. The actual attitudes and behaviors that yield high ratings on alliance, group cohesion, empathy, and so forth, differ across cases and times, as therapists do the right thing in response to clients’ emerging needs and circumstances. It is plausible that such achievements predict global outcome, but they do not show which specific activities predict which of therapy’s specific effects.

In summary, the parties involved in psychotherapy seek good outcomes, and all treatments trade on their good intentions. Treatments work, at least in part, because the participants do their responsive best to ensure they work. Research that tests the effects of responsive treatments on global evaluations yields conclusions that may be paraphrased: if participants do their best, things usually turn out well. Reassuring, perhaps, but scientifically disappointing.

The limitations of evaluative dependent variables are serious, but not so insuperable as the contamination of independent variables by responsiveness, in my opinion. Even if their potential scientific yield is limited, evaluative dependent variables can serve useful administrative, clinical, and consumeristic purposes (and there are of course many nonevaluative measures available to investigators). The effect of responsiveness on psychotherapy research’s independent variables has been insidious, however, as addressed empirically in the next section.

Treatment Characteristics and Outcome: Some Puzzling Null Findings

In this section, I summarize three puzzling but replicated results relating outcome to treatment characteristics like those listed by Knight (1941) and cited by Strupp (1963): theoretical approach, treatment duration, and verbal technique. First, outcomes were positive but were not statistically related to treatment approach in routine practice (Stiles, Barkham, Mellor-Clark, & Connell, 2008a). This is often called the Dodo verdict, named after an extinct flightless pigeon that lived on the island of Mauritius. Second, outcomes were not statistically related to treatment duration in routine practice (Stiles, Barkham, Connell, & Mellor-Clark, 2008). Third, outcomes were not statistically related to verbal techniques (Stiles & Shapiro, 1994). Others have reported similar null results, but investigators seem not to believe them and keep repeating the studies. I think all three can be understood as manifestations of responsiveness.

The first two studies were based on a samples drawn from large database of clients assessed for therapy in the British National Health Service. In the database of 33,587 adult clients and in the samples, approximately 70% were women, and mean age was approximately 38 years. Outcomes were assessed using Clinical Outcomes in Routine Evaluation—Outcome Measure (CORE-OM; Barkham et al., 2001; Barkham, Gilbert, Connell, Marshall, & Twigg, 2005; Evans et al., 2002). The CORE-OM includes 34 items that are scored on a scale of 0 to 4; scores were calculated as the mean of all items × 10 and so could range from 0 to 40. We also drew on the CORE Assessment form, completed by the therapist after intake, which records referral information, client demographics, and information about the nature, severity, and duration of presenting problems, and the CORE End of Therapy form, completed by therapist at the end of treatment, which records which theoretical approaches were used, the number of sessions attended, and whether end was planned, among other things. These data were gathered for clinical and administrative reasons and later made available for research.

Treatment Approach and Outcome

To compare the outcomes of different treatment approaches (Stiles, Barkham, Mellor-Clark, & Connell, 2008a, replicating Stiles, Barkham, Twigg, Mellor-Clark, & Cooper, 2006), we selected clients who had completed the CORE-OM at beginning and end of treatment. On the CORE End of Treatment form, therapists indicated which of the following treatment approaches they had used: psychodynamic, psychoanalytic, cognitive, behavioral, cognitive/behavioral, structured/brief, person-centered, integrative, systemic, supportive, art, and other. Many therapists indicated more than one approach.

We targeted three families of approaches: (a) CBT, if the therapist indicated any combination of cognitive, behavioral, or cognitive/behavioral; (b) person-centered therapy (PCT), if the therapist indicated person-centered; and (c) psychodynamic therapy (PDT), if the therapist indicated psychodynamic or psychoanalytic. Based on these definitions, we constructed six groups. Three groups received pure treatments, that is, one and only one of the targeted approaches. The other three groups received one of the targeted approaches plus one additional treatment, such as systemic, supportive, or art therapy, abbreviated CBT + 1, PCT + 1, and PDT + 1, respectively. These treatments might be considered as diluted or enhanced. Clients who received any other combination of treatments were not included. Of the 33,587 clients in the database, 5,613 had complete data and met specifications for one of these six groups. The six groups’ pretreatment CORE-OM means were not significantly different from each other, despite the high statistical power.

Figure 1 shows the similar distribution of pre–post change scores in the three groups. A repeated-measures analysis of variance, with treatment approach and degree of purity (pure vs. “+1”), showed the following:

First, all three treatments were effective, shown by large within-clients main effect of treatment (pre–post change; F(1, 5607) = 6805.63, p < .001, partial η² = .548). Most of the clients improved; on average, clients changed by 8.77 points on the CORE-OM. This is comparable with pre–post change in RCTs (Barkham et al., 2008). The statistical effect size was large (1.39).

Second, CBT, PCT, and PDT were equally effective. The differential treatment effect (Treatment × Pre–post interaction) was not significant (F(2, 5607) = 0.81, p = .446, partial η² < .001). That is, there was no significant difference among the three tar-
geted treatment approaches, despite the high statistical power of this test. As the Dodo said, “Everybody has won, and all must have prizes” (Rosenzweig, 1936, p. 412, quoting Carroll, 1865/1946).

As early as 1936, it was apparent to observers: “If such theoretically conflicting procedures ... can lead to success, often in similar cases, then therapeutic result is not a reliable guide to the validity of theory” (Rosenzweig, 1936, p. 412).

Third, dilution of treatments did not impair effectiveness. The mixed therapies did as well as the pure ones. The treatment purity (pure vs. “+1”) was not significant (Purity × Pre–post interaction; \( F(1, 5607) = 3.23, p = .073 \), partial \( \eta^2 = .001 \)).

The three-way interaction was not significant.

Caveats include limited specification of treatments, nonrandom assignment of clients, absence of a control group, incomplete data, restriction to one self-report measure, and investigator allegiance. These objections and others (e.g., Clark, Fairburn, & Wessely, 2008) have been addressed elsewhere (Stiles, Barkham, Mellor-Clark, & Connell, 2008a, 2008b). Caution is warranted; nevertheless, if one of these treatments were really substantially better than the others, I suggest some difference should have emerged in our results.

The similar distributions may reflect responsiveness within levels of the treatment variable. Different treatments may be equivalently effective because therapists are appropriately responsive to client requirements within each framework. All approaches ask therapists to be appropriately responsive as they work toward treatment goals. As a result, each treatment is responsively adjusted to each client’s changing requirements. To the extent this succeeds, clients have optimal outcomes. This responsiveness account presumes that each approach provides ways to respond appropriately to varied client requirements.

**Treatment Duration and Outcome**

The dose–effect model (e.g., Howard, Kopta, Krause, & Orlinsky, 1986; Kopta, Howard, Lowry, & Beutler, 1994) tries to ascertain the optimum dose of therapy, where dose is measured as the number of sessions. More sessions represent a stronger dose. We investigated the relation of treatment dose to outcome in routine practice (Stiles, Barkham, Connell, & Mellor-Clark, 2008, replicating Barkham et al., 2006).

Of 33,587 clients in the database, 9,703 had complete data, planned endings as reported by the therapist, completed 20 or fewer sessions, and began with a CORE-OM score of 10 or higher, the clinical cutoff on this instrument (Connell et al., 2007). To assess the effect of different doses, we used two indexes: rate of reliable improvement and rate of reliable and clinically significant improvement (RCSI) among clients who received each dose (\( N = 0–20 \) sessions). Following Jacobson and Truax (1991), we considered clients as reliably improved if they improved to a degree that was probably not due to chance, based on a 95% confidence interval. For this study, we considered changes larger than 4.5 points as reliable (the reliable change index). Jacobson and Truax argued further that a change could be considered as clinically significant if the client entered treatment in a dysfunctional state.

Figure 1. Distributions of Clinical Outcomes in Routine Evaluation—Outcome Measure (CORE-OM) pre–post change score in six treatment groups (from Stiles, Barkham, Mellor-Clark, & Connell, 2008a). In these notched box plots, the dot in the middle represents the median. The notch represents the 95% confidence interval. The box represents the middle 50% of the distribution. The whiskers represent the range. The marks above and below the whiskers represent outliers. CBT = cognitive-behavioral therapy; PCT = person-centered therapy; PDT = psychodynamic therapy. CBT + 1 = CBT combined with one other therapy; PCT + 1 = PCT combined with one other therapy; PDT + 1 = PDT combined with one other therapy. Adapted from “Effectiveness of Cognitive-Behavioural, Person-Centred, and Psychodynamic Therapies in UK Primary Care Routine Practice: Replication in a Larger Sample,” by W. B. Stiles, M. Barkham, J. Mellor-Clark, & J. Connell, 2008, Psychological Medicine, 38, p. 681. Cambridge University Press.
and left in a normal state, interpreted as moving from above to below the cutoff between the dysfunctional and normal populations. Thus, we considered clients as having achieved RCSI if their CORE-OM score changed by at least 4.5 points and went from 10 or above to below 10.

Figure 2 shows the percentage of clients who made reliable improvement as a function of how many sessions they received. That is, they changed by 4.5 points or more on the CORE-OM. Approximately 80% of clients improved reliably regardless of how many sessions they had; outcomes did not improve with larger doses. The percentages were more variable for the longer treatments because fewer clients had longer treatments, and the rates were less stable. Figure 3 shows that the RCSI rate was slightly negatively correlated with number of sessions. This is partly because clients who received longer treatments had slightly higher initial scores, so although they averaged as much improvement as clients with fewer sessions, more ended above the cutoff.

Finding improvement unrelated to treatment dose may seem paradoxical and surprising if treatment is considered as an experimental manipulation, but it is clinically sensible if clients and therapists are considered as responsively ending treatment when a good-enough level has been reached (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009; Barkham et al., 2006; Stiles, Barkham, Connell, & Mellor-Clark, 2008). Clients change at different rates and achieve a good-enough level of gains at different treatment durations. When they have had enough, they stop. That is, participants respond to treatment in a way that is subjectively satisfying.

Verbal Technique and Outcome

The drug metaphor approach to assessing process–outcome relationships (Stiles & Shapiro, 1989) suggests that the verbal and nonverbal components of psychotherapy can be treated like the ingredients of pharmacological agents in evaluating their strength, integrity, and effectiveness (cf. Yeatn & Sechrest, 1981). If a process component is an active ingredient, then administering a high level of it should yield a positive outcome. If it has no effect, the process component is presumed to be inert. Experimental studies of individual process ingredients would be prohibitively expensive, so most process–outcome studies have assessed naturally occurring covariation of process and outcome variables (Orlinsky, Grawe, & Parks, 1994). If a process component is an active ingredient, clients who receive relatively more of it should tend to improve more, so measures of this component should be positively correlated with measures of outcome across clients.

We (Stiles & Shapiro, 1994) applied this logic to data from the First Sheffield Psychotherapy Project (Shapiro & Firth, 1987). This was a comparative trial of two treatments for depression: psychodynamic–interpersonal (PI) and cognitive–behavioral (CB). The study had an unusual cross-over design in which clients received eight sessions of each treatment with the same therapist throughout. Order of treatments was counterbalanced across clients. After an initial assessment, clients received eight sessions of one treatment, then a mid-treatment assessment, then eight sessions of the other treatment, and then an end-of-treatment assessment. The assessment battery included the Present State Examination (Wing, Cooper, & Sartorius, 1974), which is an interview-based rating schedule done by a trained assessor, and two standard self-report measures: the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and the Symptom Checklist-90 (Derogatis, Lipman, & Covi, 1973). Clients in both groups showed substantial improvement on these measures, and the mean differences between treatments were negligible after the counterbalancing was dealt with statistically (Shapiro & Firth, 1987). So this was another Dodo result, but that is not the point I want to make with this study.

We coded the verbal response mode (Stiles, 1992) of every therapist and client utterance in half of the sessions in the project (eight sessions per client, four from each treatment), totaling more than 350,000 utterances (Stiles, Shapiro, & Firth-Cozens, 1988). Verbal response modes are speech acts; they describe what people do when they speak rather than the content of what they say. For this analysis, we selected four theoretically important categories of therapist interventions and one important client category. The therapist categories were Questions, which are information-seeking utterances; General Advertisements, which are directives guiding client behavior outside sessions (distinguished from Process Advertisements, which direct in-session behaviors); Interpretations, which include any explaining or labeling of client thoughts or actions; and Exploratory Reflections, which express client’s meanings and feelings from the client’s perspective (distinguished from Simple Reflections, which are repetitions). Disclosures, the client category, reveal the client’s private experience (see Stiles, 1992).
Table 1 shows the percentages of therapist or client utterances. These categories accounted for a substantial proportion of what happened in the session, especially insofar as many of the other utterances were "mm-hm" (Stiles et al., 1988). Questions and General Advisements were much more common in the CB than in the PI treatment, whereas Interpretation and Exploratory Reflections were much more common in the PI than in the CB treatment, although the same therapists conducted both treatments. These patterns conform to the theories of how these respective treatments should be conducted. Client Disclosure was a bit more common in PI than in CB treatment, but it constituted a large proportion of client utterances in both treatments.

You might expect that such common theoretically important categories would be correlated with improvement across treatment on the Present State Examination, Beck Depression Inventory, and Symptom Checklist-90. Table 2, however, shows the process-outcome correlations were negligible. These correlations encompassed the whole treatment, but separate analyses examining use of these verbal response modes in PI and CB treatments separately yielded the same null result (Stiles & Shapiro, 1994).

This does not appear to reflect a failure of measurement. The verbal process components were reliably coded. They represented common theoretically important techniques. They discriminated between treatments in sensible expected ways. The outcome measures were standard, and they detected large clinically and statistically significant changes in these treatments. These null correlations replicate generally disappointing inconsistent yield of treatment outcomes; they catch on quickly and apply what they learn. Others may be resistant or slow, requiring many repetitions and explanations. Appropriately responsive therapists adapt their interventions to these variations. They give relatively few interpretations to clients who catch on quickly and relatively more to clients who require repetition or rephrasing. If therapists are good but imperfect, the quicker easier clients may have better outcomes, whereas the slower more difficult ones may have poorer outcomes, despite the therapists’ best efforts. As depicted in Figure 4, this can yield a negative correlation between interpretations and outcome, although, by assumption, interpretations are an important active ingredient in therapy. Such seemingly paradoxical results have been reported in some studies (e.g., Sloane, Staples, Cristol, Yorkston, & Erbaugh, 1961; Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); SCL-90 = Symptom Checklist-90 (Derogatis, Lipman, & Covi, 1973).

When responsiveness is substantial but imperfect, the correlation of crucial process components with outcomes may be misleadingly positive, null, or even negative. Suppose, hypothetically, that interpretations are an important active ingredient in psychotherapy. Clients vary in their requirements for interpretations for many reasons. Some may be motivated and receptive to therapists’ interventions; they catch on quickly and apply what they learn. Others may be resistant or slow, requiring many repetitions and explanations. Appropriately responsive therapists adapt their interventions to these variations. They give relatively few interpretations to clients who catch on quickly and relatively more to clients who require repetition or rephrasing. If therapists are good but imperfect, the quicker easier clients may have better outcomes, whereas the slower more difficult ones may have poorer outcomes, despite the therapists’ best efforts. As depicted in Figure 4, this can yield a negative correlation between interpretations and outcome, although, by assumption, interpretations are an important active ingredient in therapy. Such seemingly paradoxical results have been reported in some studies (e.g., Sloane, Staples, Cristol, Yorkston, & Whipple, 1975). Readers and reviewers could be tempted to conclude incorrectly that interpretations are useless or even harmful.

How Can We Do Psychotherapy Research Without Variables?

Let us back up a step. The central purpose of science, I submit, is to develop a coherent and accurate understanding of how the world works, that is, to construct general, precise, realistic theories. A theory is a semiotic representation of the world—a description in words, numbers, diagrams, or other signs. Among other things, a good theory is a practical guide for decisions about how to act in the world, for

<table>
<thead>
<tr>
<th>therapists</th>
<th>clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>question</td>
<td>general advisement</td>
</tr>
<tr>
<td>(*percentage larger than in the other treatment, p &lt; .001)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Correlations of Verbal Response Mode Percentages With Change From Intake to Termination on Outcome Measures

<table>
<thead>
<tr>
<th>Therapist VRM</th>
<th>PSE</th>
<th>BDI</th>
<th>SCL-90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapist VRM</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.11</td>
</tr>
<tr>
<td>Client VRM</td>
<td>-0.11</td>
<td>-0.08</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

Note: N = 39 clients. VRM = verbal response mode; PSE = Present State Examination (Wing, Cooper, & Sartorius, 1974); BDI = Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); SCL-90 = Symptom Checklist-90 (Derogatis, Lipman, & Covi, 1973).

Table 1: Mean Percentages of Therapist Utterances and Client Utterances in 312 Sessions of Psychodynamic–Interpersonal and Cognitive–Behavioral Therapy (Eight Sessions per Client, Which Was 50% of Each Treatment)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Therapists</th>
<th>Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>psychodynamic–interpersonal</td>
<td>question</td>
<td>5.7</td>
</tr>
<tr>
<td>cognitive–behavioral</td>
<td>general advisement</td>
<td>0.8</td>
</tr>
<tr>
<td>explanation</td>
<td>interpretation</td>
<td>20.3</td>
</tr>
<tr>
<td>disclosure</td>
<td>exploratory reflection</td>
<td>8.3</td>
</tr>
</tbody>
</table>

Note: N = 39 clients. * Percentage larger than in the other treatment, p < .001.
example, how to practice psychotherapy. The purpose of empirical scientific research is quality control—checking that the theory is a good one by comparing it with observations. The observations change the theory by increasing or decreasing confidence or more often by showing where modifications, elaborations, or extensions are needed (Lakatos, 1978; Stiles, 1981, 2009a; cf. Miller, 2009).

One strategy for systematically bringing observations to bear on a theory is hypothesis testing. Its logic is the hypothetico-deductive method. The familiar statistical version of this is to deduce one or a few statements from a theory and compare each such statement with many observations. That is, we test a hypothesis by seeing whether it holds across cases. This may involve experiments (e.g., RCTs) but also naturalistic comparisons and correlational studies. If the observations tend to correspond to the statement, our confidence in that statement is substantially increased. We say the hypothesis was confirmed, or at least the null hypothesis was rejected. This yields a small increment or decrement of confidence in the theory as a whole, or information about where the theory needs modifications.

Theory-building qualitative research offers an alternative strategy (Stiles, 2009a, 2010). In a theory-building psychotherapy case study, investigators compare each of many theoretically based statements with one or a few observations. At issue is the correspondence of theory and observation—how well the theory describes details of the case. Like variables in statistical hypothesis-testing studies, terms in theoretical descriptions have explicit links within the theory, but unlike variables, descriptions need to be applicable to only one or a few cases to be scientifically useful. Each detail may be observed only once, but many details are observed across the course of treatment. For familiar reasons, such as selective sampling, low power, investigator biases, and so forth, the change in confidence in any one statement may be small. However, because many theoretical statements are compared with case observations, the gain in confidence in the theory may be as large as from a statistical hypothesis-testing study. A collection of systematically analyzed cases that match a theory in precise detail may give people a good deal of confidence in the theory as a whole, although each component assertion may remain tentative and uncertain when considered in isolation. The key is making many detailed theory-relevant observations of the case. Campbell (1979) described these multiple observations as analogous to multiple degrees of freedom in a statistical hypothesis-testing study.

Whereas statistical hypothesis testing requires theoretically framed independent and dependent variables that can be observed in every case, theory-building case studies can use any case observations that can be described in theoretical terms. As Strupp (1963) pointed out and I have elaborated here, both independent and dependent variables are highly problematic in psychotherapy research, contaminated by context, responsiveness, and global judgments. In contrast, major theories of psychotherapy—psychodynamic, CB, humanistic—apply across diverse contexts, across diverse client characteristics, and across therapeutic relationships that follow responsive nonlinear courses. By comparing the rich theoretical descriptions with rich case observations, theory-building case studies can empirically assess theoretical precision and realism and accumulate improvements.

The assimilation model—a theory of psychological change that tracks the assimilation of disconnected problematic experiences into the usual self in successful psychotherapy (Stiles, 1999, 2002, 2011)—has developed and continues to develop mainly through theory-building case studies (e.g., Caro Gabalda, & Stiles, in press; Gray, & Stiles, 2011; Meystre, Kramer, De Roten, Despland, & Stiles, 2012; Osatuke, Reid, Stiles, Zisook, & Mohamed, 2011; Ribeiro, Bento, Salgado, Stiles, & Gonçalves, 2011; Schielke et al., 2011; Tikkanen, Stiles, & Leiman, 2011). Assimilation model research has used theory-building case study deliberately, but I think most of our rich explanatory theories of psychotherapy have been developed in substantial part through similar strategies informally.

According to their logic, both hypothesis-testing and theory-building qualitative strategies center on an articulated theory. Hypothesis testing assesses specific consequences of a theory; results inductively yield some small incremental or decrement in confidence about the theory. If there is no link to theory, there is no systematic way to generalize the results, except perhaps in the limited sense of statistical generalization of a single statement to the population from which a particular sample was randomly drawn. The theory must be logically interconnected enough that observations on one part of the theory (the hypothesis) can affect confidence in other parts of the theory. Likewise, theory-building case studies depend on logical coherence of the theory to ensure that the theoretical descriptions of one case link to statements about other cases.

The qualitative theory-building strategy does not directly address the consumeristic global evaluative question of whether psychotherapy is effective—or of which treatment approach is most effective. As Strupp (1963) said, “more pressing matters must be dealt with first before we can address ourselves meaningfully to [that] question” (p. 1). I suggest that a solid empirically supported theoretical account of how people change and how psychotherapy facilitates changes is such a pressing prerequisite.

The distinction between statistical hypothesis testing and theory-building strategies parallels the classic distinction between nomothetic and idiographic research. I have here argued, in effect,
that idiographic observations need not be restricted to the context of discovery but can be systematically brought to bear on theory in the context of justification. In addition, theories built from clinical case studies may be more useful to clinicians than theories built from hypothesis-testing studies because case-tested theories will have had more theoretical features examined (Miller, 2009).

Of course, the theory-building qualitative strategy does not overcome all of the problems of idiographic research. Problems involving the limited view offered by a small selection of cases and potential observer biases, for example, must be addressed by investigators who choose qualitative strategies. However, the theory-building logic does address the problem of generalizing from small numbers of cases. Campbell (1979), who had earlier been one of the most vociferous critics of case study research (e.g., Campbell, 1961; Campbell & Stanley, 1966), explicitly retracted much of his earlier criticism of case studies in light of the degrees of freedom logic. Nomothetic research has methodological problems of its own, and adding the variables problem targeted by Strupp in 1963 makes the idiographic alternative relatively more attractive for research on psychotherapy.

References


http://jrul.libraries.rutgers.edu/index.php/pcsp/article/view/973/2384

Received August 29, 2012
Accepted September 2, 2012