Clinicians' prediction and recall of therapeutic interventions: practice research network study

Louis G. Castonguay, Rebecca A. Janis, Soo Jeong Youn, Henry Xiao, Andrew A. McAleavey, James F. Boswell, Dever M. Carney, Mary A. Boutselis, Melora Braver, Nancy R. Chiswick, Neal A. Hemmelstein, Jeffrey S. Jackson, Richard A. Lytle, Marolyn E. Morford, Heather S. Scott, Catherine S. Spayd & Mary O'Leary Wiley

Routledge
Taylor & Francis Group

Clinicians' prediction and recall of therapeutic interventions: practice research network study

Louis G. Castonguay[a], Rebecca A. Janis[a], Soo Jeong Youn[a], Henry Xiao[a], Andrew A. McAleavey[b], James F. Boswell[c], Dever M. Carney[a], Mary A. Boutselis[d], Melora Braver[d], Nancy R. Chiswick[d], Neal A. Hemmelstein[d], Jeffrey S. Jackson[d], Richard A. Lytle[d], Marolyn E. Morford[d], Heather S. Scott[d], Catherine S. Spayd[e] and Mary O'Leary Wiley[f]

[a]Department of Psychology, Pennsylvania State University, University Park, PA, USA; [b]Department of Psychiatry, Weill Cornell Medical College, New York, NY, USA; [c]Department of Psychology, University at Albany, State University of New York, Albany, NY, USA; [d]Independent Practice, State College, PA, USA; [e]Independent Practice, Duncansville, PA, USA; [f]Independent Practice, Altoona, PA, USA

ABSTRACT

Conducted within a practice-research network in private practice, this exploratory study was aimed at examining whether clinicians can accurately predict and recall profiles of therapeutic interventions they used during an entire treatment for a given client. Based on a small sample (7 clinicians and 30 clients), the results tentatively suggest that the predictions that therapists made after session 3 regarding which types of techniques they would use, as well as the retrospective assessment of typical techniques they reported using in therapy, were accurate and generally discriminative. Clinical implications in line with deliberate practice are suggested and future research on complex questions related to clinical prediction is proposed.

It has been argued that one way to foster practitioners' interest in empirical findings is to facilitate their engagement in research (Elliott & Morrow-Bradley, 1994). For busy clinicians, however, conducting research is difficult. Even for many of those who choose to go to graduate school in part because they wanted to learn how to conduct scientific studies, they have quickly realized that the demands of their professional career rob them of at least two things that are necessary to maintain (or reconnect) with their interest in research: an up-to-date knowledge of methodological and statistical advances in the field, and time. As part of practice-oriented research (POR, Castonguay, Barkham, Lutz, & McAleavey, 2013), the development of practice-research networks (PRN) has been identified as a potential conduit to address these obstacles – put within the context of the current series of papers, PRN can be viewed as a venue permitting individuals to be engaged in both the science and practice of psychotherapy.

PRN infrastructures are based on an active collaboration of diverse stakeholders in mental health care, including clinicians and researchers. These partnerships are built on mutual respect, as well as on a combination of complementary expertise. As such, the clinicians'

---

engagement in PRN is a way for these partnerships to recognize and rely on the unique knowledge that comes from day-to-day practice. Reciprocally, and related to the first obstacle mentioned above, the participation of researchers allows these partnerships to have and maintain close links with research innovations and expertise. With regard to the issue of time, PRN partnerships are also mindful to design studies that are minimally burdensome or disruptive of day-to-day practice (Koerner & Castonguay, 2015). In contrast to investigations conducted in controlled settings, PRN studies aim to be conducted in the practitioners' own clinical milieu, and as part of clinical routine. Referred to as clinically syntonic research (Castonguay, 2011), an optimal goal of such studies is for clinicians to seamlessly integrate clinical and research activities.

A commitment to research, of course, is most likely if it is perceived as being relevant and helpful. One characteristic of POR, and thus of any PRN study, is that it indeed seeks to address questions that are of interest to clinicians (Castonguay et al., 2013). For the present study, this interest revolved around examining whether clinicians could accurately predict how their therapy might unfold for any one of their clients, a clinical skill assumed to be part and parcel of their expertise.

Since the seminal work that Meehl (1954) published more than half a century ago, the accuracy of clinical prediction has been regarded, at least in some academic milieus, with skepticism. Over the years, this skepticism has been fueled by a number of studies which demonstrated that for assessment and diagnostic purposes, actuarial and statistical predictions are more valid than clinical ones (Ægisdóttir et al., 2006; Garb & Boyle, 2015; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Grove & Meehl, 1996; Meehl, 1973; Sarbin, 1943).

This skepticism has been extended so that researchers have questioned the assumption that the ability to make accurate clinical decisions and predictions improves with experience. For example, a review of the literature on outcome assessment, therapist expertise, and clinical decision-making led to the conclusion that experienced therapists, when compared with novice therapists, are not more accurate at generating case conceptualizations or evaluating their own clinical competence (Tracey, Wampold, Lichtenberg, & Goodyear, 2014). A number of other studies have similarly demonstrated that when presented with the same clinical information, experienced therapists make diagnostic judgments that are no more valid or accurate when compared with less experienced clinicians (Dawes, 1994; Garb, 1998, 2005; Tracey et al., 2014; Wiggins, 1973). A meta-analysis conducted by Spengler and colleagues (2009) found that the specific clinical training and experiences had no effect on the validity of clinical diagnoses, psychological assessments, or outcome predictions.

Psychotherapists' ability to predict deterioration has also been found to be poor: practitioners tend to underestimate the number of their clients who will deteriorate, and they do not tend to correctly predict who, among their clients, will do worse during therapy, and who will not. In a study of 40 trainees and licensed professionals and 550 patients, clinicians predicted that only three (or .01%) of these clients would leave treatment in a deteriorated state. In reality, 40 patients were worse off by the end of treatment, and only one of the initial three predictions was correct (Hannan et al., 2005). Another study estimated therapists' awareness of negative changes occurring during therapy, as indicated by their session notes. By examining case files, researchers were able to determine that therapists only identified deterioration in 21% of clients whose outcome scores had worsened. The most severe cases that saw an extreme increase in symptoms and distress were only identified at a 32% rate (Hatfield, McCullough, Frantz, & Krieger, 2010). Another study compared clinical and actuarial

predictions of whether children and adolescents would deteriorate, and the actuarial method was significantly more accurate (Salisbury, 2015). Clinicians in this study did not predict a single client would deteriorate during treatment.

Yet, clinical prediction and judgment are not restricted to the assessment and outcome of their clients. One of the cognitive tasks therapists are expected to perform is the planning of treatment. More or less explicitly and systematically, practitioners are trained to anticipate and plan interventions that they are likely to use during the course of therapy for each of their particular clients. To our knowledge, no study has been conducted to determine if therapists are able to accurately predict what they will and will not do in sessions. This contrasts with the fact that the real-world assessment of psychotherapist intervention use has received significantly more attention in recent years. For example, multiple United States government-sponsored task forces have highlighted the importance of measuring therapists' actions during treatment, and their consistency (or lack thereof) with identified evidence-based practice standards (Brown, Scholle, & Azur, 2014; England, Butler, & Gonzalez, 2015). Given the practical barriers to integrate gold standard observational coding measures in routine psychotherapy practice (Garland & Schoenwald, 2013), task force committee members have recommended increased research attention to the use of participant-perspective self-report measures of treatment process; for example, research on the implementation of therapist self-report tools to monitor intervention use and the predictive validity of such self-assessments. Furthermore, the Committee on Developing Evidence-Based Practice Standards recommended that such monitoring tools assess both theory-specific technical factors and common therapeutic factors (England et al., 2015).

The goal of this paper is to address this gap by reporting results from an exploratory study designed and implemented by clinicians in their routine clinical practice. As part of a larger investigation of the use and impact of evidence-based techniques, therapists were specifically interested in examining a number of research questions, including whether they would be able to delineate the profile of technical interventions that they were likely to use with particular clients, and whether these profiles would accurately predict the way they conducted therapy. They were also interested in testing if what they remembered doing at the end of therapy was an accurate estimation of the interventions they actually used during therapy. This study was planned and conducted by the members of the Pennsylvania Psychological Association Practice-Research Network (PPA-PRN), which is a partnership of private practitioners and academics (faculty members and graduate students). Reflecting the essence of PRN, this partnership is based on an active collaboration of all members in every aspect of a study, from the selection of the topics to be investigated, design of the methods, orchestration of the implementation of these methods, collection and analyses of the data, as well as the dissemination of the results. For the present study, clinicians and academics met regularly for several months to decide what to investigate and how to do so. Then, for up to 24 months, licensed therapists in this partnership recruited clients from their routine private practice. In preparing and implementing this study, the partnership members were guided by core characteristics of POR which, in addition of the investigation of issues relevant to clinicians' practice, involve: the adoption of standardized measures in naturalistic settings and as part of clinical routine; the active participation of clinicians in data collection and the shared ownership of these data; the integration of empirically valid and clinically informative tasks; and, as reflected by the authorship on this paper, the recognition of clinicians' contribution to the accumulation of empirical knowledge (Castonguay et al., 2013).

In addition, the preparation and implementation of the current study has relied on lessons learned from previous research efforts of the PPA-PRN (Borkovec, Echemendia, Ragusea, & Ruiz, 2001; Castonguay et al., 2010). Based on the identification of obstacles faced, strategies to deal with these challenges, and general recommendations about POR from clinicians involved in these efforts (see Castonguay et al., 2010), attempts were made to maximize the clinical feasibility (in terms of the time required and complexity) and potential usefulness of the research protocol. By making use of practitioners' expertise (procedural, conceptual, clinical), this study stands in contrast with the historical problem of "letting the knowledge from practice drip through the holes of a colander" (Kazdin, 2008, p. 155).

Method

Participants

Data from seven therapists in private practice who participated in the design and implementation of this study are included in the present analyses (see "Procedures" section below). Five therapists were female, and all were Caucasian. All participating therapists were doctoral-level psychologists, with a mean of 20.71 years (range = 10–36 years) of post-training experience. Therapists' theoretical orientation was determined by using items from the Development of Psychotherapists Common Core Questionnaire (DPCCQ; Orlinsky et al., 1991). As few practitioners in naturalistic settings identify their theoretical allegiance within single and mutually exclusive categories, we chose these items because they dimensionally assess several approaches. On a scale of 0 (*Not at all*) to 5 (*Very much*), items measure how much a therapist's current practice is guided by each of five theoretical orientations (analytic/psychodynamic, behavioral, cognitive, humanistic, systems theory) and the extent to which a therapist views his/her approach as eclectic/integrative. As shown in Table 1, therapists self-defined most strongly as eclectic/integrative and reported being guided most highly by cognitive-behavioral approaches and, to a lesser extent, by humanistic and analytic/psychodynamic orientations.

The current study involved 30 clients (see "Procedures" section below) with a mean age of 31.22 years (SD = 12.19). Of these clients, 63.0% were female, and from those who reported their ethnicity, they identified as Caucasian. In terms of marital status, 50.0% were single, 36.7% were married, and 3.3% were divorced (10.0% did no report their marital status). Primary DSM-IV/ICD 9 diagnoses, as provided by therapists at session 3, were as follows: anxiety disorders (37.9%), mood disorders (31.0%), adjustment disorders (27.6%), and eating

Table 1. Mean ratings of therapist orientation and correlations between therapist orientation and "session" MULTI scores.

		Correlations with "session" MULTI			
	M (SD)	BC	IO	IPT	CF
Analytic	2.29 (1.60)	.68	.72	.89*	.37
Behavioral	3.86 (.69)	−.37	−.35	.03	−.21
Cognitive	4.29 (.49)	−.42	−.60	.01	−.64
Humanistic	3.00 (.82)	.16	.23	.26	.44
Systemic	1.86 (1.57)	.76*	.81*	.52	.94*
Eclectic	3.86 (1.21)	.64	.56	.77*	.53

Note: N = 7, BC = Behavioral Change, IO = Insight Oriented, IPT = Interpersonal Therapy, CF = Common Factors.
*Significant p < .05.

disorders (3.4%). The mean number of clients per therapist included in these analyses was 4.29 (range = 3 to 7), and the number of sessions per client ranged from 4 to 67 (M = 14.90; SD = 14.50).

Measures

In addition to providing basic demographic information, participants completed the Multitheoretical List of Therapeutic Interventions (MULTI; McCarthy & Barber, 2009). The MULTI assesses the interventions used by a therapist during a session. Its three versions (client, therapist, observer) are composed of 60 items measuring techniques from seven empirically supported approaches (cognitive [CT], behavioral [BT], dialectical-behavioral [DBT], process-experiential [PE], person-centered [PC], psychodynamic [PD], and interpersonal [IPT]), as well as common factors (CF, such as fostering positive expectancies and therapeutic relationship). The MULTI items were intentionally designed to be relatively jargon neutral and behaviorally anchored. Items describe therapist behaviors that may or may not have occurred to some degree during the session (e.g. "I set an agenda or established specific goals for the therapy session"; "I repeated back to my client (paraphrased) the meaning of what he/she was saying"; "I made connections between my client's current situation and his/her past"). Each item is scored using a Likert-scale, ranging from 1 (*Not at all typical of the session*) to 5 (*Very typical of the session*). The MULTI has demonstrated adequate psychometric qualities in terms of face, content, and criterion validity, as well as its overall factor structure (McCarthy & Barber, 2009). For this study, only the therapist self-report version was used. The internal consistency of each of the eight scales was adequate in the present study, with alpha coefficients ranging from .698 to .876. For the sake of reducing statistical comparisons, similar to previous studies (Boswell, Castonguay, & Wasserman, 2010; McAleavey & Castonguay, 2014), six of the scales were combined into two theoretically broader ones: Behavioral change (BC) and Insight Oriented (IO). The BC scale included CT, BT, and DBT sub-scales, whereas the IO scale combined PE, PC, and PD sub-scales. The internal consistency for both aggregated scales was high ($α$ = .943 for BC and $α$ = .904 for IO).

Procedures

In order to reduce the burden on routine clinical practice, the recruitment of clients for each therapist was limited to four clients at any given time during the study. Specifically, at the end of the first session, therapists invited their first four new (non-returning) adult (18 year old and above) individual therapy clients to participate in the study, unless the therapist believed that this would be clinically counter-indicated for the client.[1] After reading and signing an informed consent (approved by the Penn State University Office of Regulatory Compliance), participating clients were randomly assigned to an experimental condition or a control condition. Participating clients who terminated therapy before the end of the study, or who dropped out of the study while remaining in treatment, were replaced by the next eligible client(s) who agreed to participate. These new clients were randomly assigned to either the experimental or control conditions.

The two conditions were identical with the exception of one research manipulation. In the experimental condition (but not in the control condition), clients and therapists filled out a few questionnaires and questions at the end of each session. Only one of those, the

therapist rated MULTI, was the focus of the present paper. Additionally, only clients in the experimental condition were part of the present study.

Three versions of the MULTI were used, each at a different stage of the client's treatment. First, at the end of each session, therapist filled out the standard version of the MULTI, which, as noted above, rates the interventions that were delivered in the session that was just completed. For the rest of the paper, the mean of all of the MULTIs filled out at the end of each session (after session 3) is referred to as the "session MULTI," representing the average profile of interventions that the therapist reported using at each session with that client. Therapists also filled out two modified versions of the MULTI. The first modified version, called the "special" MULTI, was filled out after the third session of therapy and after the therapist re-diagnosed the client (this was made a systematic part of the research protocol because therapist participants were skeptical about the validity of a diagnosis assigned before or during the first session). As stated in its instructions, the "special" MULTI was aimed at assessing the profile of interventions that therapists *expected* to use with each particular client. The second modified version, called "termination" MULTI, was filled out after the last session of therapy to assess (also as stated in its instructions) the profile of interventions that therapists believed they used with each particular client during the entire course of the treatment.

A total of 43 clients were assigned to the experimental condition, but eight participants were removed from the analyses because they did not have a "special" MULTI completed and/or did not have any "session" MULTIs after session 3. Moreover, 4 of the 11 therapists who participated in the study had contributed less than 3 clients each. These therapists were removed from the analyses in order to establish reliable therapist estimates. As a consequence, five additional clients were removed. In addition, three clients had a completed "special" MULTI and "session" MULTIs after session 3, but did not have a completed "termination" MULTI. Consequently, these clients were excluded from analyses conducted with the "termination" MULTI. This left a sample size of 7 therapists and 30 clients for analyses with the "special" MULTI, and 27 clients for analyses with the "termination" MULTI.

Data analyses

The goal of this study was to assess how well therapists were able to predict and recall the profiles of therapeutic interventions they used with a given client during treatment. In line with this goal, two primary sets of models were analyzed: (1) the "special" MULTI subscales (administered at the third session) predicting the aggregate "session" MULTI subscales for prospective association, and (2) the "termination" MULTI subscales predicting the "session" MULTI subscales for retrospective association. Within both sets of models, four models were estimated – one each with the four "session" MULTI subscales as dependent variables being predicted by the "special" or "termination" MULTI subscales.

Because therapists varied on the overall level of different interventions that they reported using, much of the variance in MULTI scores existed between therapists, causing the MULTI scores of clients within a given therapist to be more similar to each other. To quantify this variance, we calculated and report intraclass correlation coefficients (ICCs) for each subscale on the "session," "special," and "termination" MULTIs, which indicates how much of the variance of each subscale exists between therapists relative to within therapists (i.e. how much of the differences in MULTI scores are due to differences between therapists versus differences

between clients within therapists). Additionally, to account for this shared variance, the analyses outlined above were conducted using multilevel models, with clients nested within therapists. Analyses were conducted using the "nlme" package (Pinheiro, Bates, DebRoy, & Sarkar, 2013) in the R programming language (version 3.2.3; R Development Core Team, 2014).

Each model included as fixed effects the four "special" MULTI subscale scores (model set 1) or the four "termination" MULTI subscale scores (model set 2). "Special" and "termination" MULTI scores were centered around the therapist mean, allowing for the associated fixed effects to be interpreted as deviations from the therapist average. This addresses the question of whether therapists are able to accurately predict and recall their intervention with a specific client. Specifically, when a therapist predicts or recalls using more of a specific intervention than average with a client, do they also report using more of that intervention than average throughout treatment? In this framework, large fixed effects between corresponding subscales (e.g. BC "special" MULTI predicting BC "session" MULTI) would indicate therapist ability to predict and recall their intervention use with a specific client.

The models also included a random therapist intercept, modeling random variation between therapists in levels of "session" MULTI subscale scores. The inclusion of a random therapist effect also allows the results of the analyses to be generalized beyond the therapists included in the present study. The general model equation is

$$\text{Level 1: SessionMULTI}_{ij} = \beta_{0j} + \beta_{1j}\left(\text{BC}_{ij}\right) + \beta_{2j}\left(\text{IO}_{ij}\right) + \beta_{3j}\left(\text{IPT}_{ij}\right) + \beta_{4j}\left(\text{CF}_{ij}\right) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}; \beta_{1j} = \gamma_{10}; \beta_{2j} = \gamma_{20}; \beta_{3j} = \gamma_{30}; \beta_{4j} = \gamma_{40}$$

where predictors BC, IO, IPT, and CF are "special" MULTI subscales for model set 1 and "termination" MULTI subscales in model set 2. SessionMULTI$_{ij}$ is the predicted "session" MULTI subscale score for therapist j with client i. β_{0j} is the unadjusted mean "session" MULTI subscale score for therapist j, and r_{ij} represents the difference between therapist j's average client and client i's individual "session" MULTI score. In the level 2 model, γ_{00} represents the mean "session" MULTI score for clients with the mean level of each predictor, while u_{0j} represents therapist j's unique deviation around that mean. For the predictors, β_{1j} indicates the 1 unit increase in "session" MULTI score associated with a 1 unit increase in BC for therapist j, adjusted for therapist differences in BC and controlling for the other predictors. It is equal to the fixed effect for BC (γ_{10}). The other predictors are interpreted similarly.

Analyses were conducted using restricted maximum likelihood estimation, which has been shown to provide less biased variance estimates with small sample sizes (McNeish & Stapleton, 2016). Further, we applied a Kenward–Roger adjustment, which corrects for inflated Type-I error rates due to underestimated standard errors in small sample sizes. This correction leads to less biased significance tests of fixed effects (Kenward & Roger, 1997; McNeish & Stapleton, 2016).

Due to the small sample size of seven therapists, we approached the current study as a pilot study and exploratory investigation of therapists' ability to predict and recall their reported interventions. In small samples such as ours, the standard error, and resultantly the tests of significance, has been shown to be biased. In line with this, we only interpret the fixed effects coefficients from our analyses, not the inferential statistics, i.e. the significance tests. Point estimates for fixed effects remain relatively unbiased even for small samples,

allowing us to interpret the strength of relationships between MULTI subscales independent of statistical significance.

It should be recognized that relationships between subscales of the different versions of the MULTI only assess the therapists' prediction and recall at the general level of interventions, which we referred to above as profiles. Psychometrically, it would not be sound to conduct analyses at an item level, neither would it be fair to assume that, pragmatically, therapists can precisely predict and recall 60 specific types of therapeutic actions.

Results

Table 1 shows the correlations between therapists reported dimensional orientation scores and their "session" MULTI scores across all clients. Only 5 of the 24 correlations were significant, none of them conceptually related. Systemic orientation ratings were significantly correlated with CF, BC, and IO "session" MULTI ratings, and analytic and eclectic orientation ratings were both significantly correlated with IPT "session" MULTI ratings (Table 2). ICCs for all MULTI subscales (Table 3) ranged from .14 to .89, with IPT consistently producing the lowest ICCs and CF the highest. ICCs were generally consistent for corresponding subscales across time points. Overall, the ICCs indicated significant dependence within scores from the same therapist, necessitating the use of multilevel modeling to account for this variance. Despite the large ICCs indicating a strong therapist effect, there was non-trivial variance within therapists which was of interest to our research question: Can therapists predict and recall the interventions they would use with a given client compared to their other clients. As outlined above, centering the "special" and "termination" MULTI scores on the therapist mean allows us to answer this question.

In general, "session" MULTI subscale scores were most strongly predicted by the corresponding "special" or "termination" MULTI subscale scores, as shown by the fixed effects coefficients presented in Tables 4 and 5. The results of the multilevel models for "session" MULTI subscales regressed on "special" MULTI subscales (i.e. therapist prospective prediction) are presented in Table 4. The fixed effects point estimates for each subscale pairing represent the expected change in the outcome, the "session" MULTI subscale, when the "special" MULTI subscale value is increased by one point, adjusting for the therapist's mean and while

Table 2. Intraclass correlation coefficients (ICCs) for MULTI subscales.

	Subscale	ICC
"Special" MULTI	BC	.72
	IO	.72
	IPT	.30
	CF	.76
"Session" MULTI	BC	.71
	IO	.76
	IPT	.34
	CF	.89
"Termination" MULTI	BC	.73
	IO	.64
	IPT	.14
	CF	.75

Therapist $N = 7$; BC = Behavioral Change, IO = Insight Oriented, IPT = Interpersonal Therapy, CF = Common Factors.

Table 3. Mean MULTI subscale scores.

	"Special" MULTI	"Session" MULTI	"Termination" MULTI
Behavioral Change	3.35 (.60)	2.80 (.67)	3.18 (.72)
Insight Oriented	3.18 (.77)	2.67 (.61)	3.04 (.73)
Interpersonal Therapy	3.24 (.99)	2.78 (.78)	3.18 (.82)
Common Factors	4.31 (.74)	4.01 (.77)	4.17 (.80)

Standard deviations in parentheses. "Special" and "session" MULTI *N* = 30; "termination" MULTI *N* = 27.

Table 4. Fixed effects coefficients and *p* values from multilevel models with "special" MULTI subscales centered around the therapist mean predicting "session" MULTI subscales.

"Session" MULTI subscale	"Special" MULTI subscale fixed effect coefficients and *p* values			
	Behavioral Change	Insight Oriented	Interpersonal Therapy	Common Factors
Behavioral Change	**.39, *p* = .22**	.38, *p* = .18	−.15, *p* = .31	.19, *p* = .44
Insight Oriented	−.08, *p* = .68	**.54, *p* = .01**	.01, *p* = .92	.16, *p* = .33
Interpersonal Therapy	−.49, *p* = .30	−.09, *p* = .08	.43, *p* = .08	**.73, *p* = .06**
Common Factors	−.25, *p* = .23	.31, *p* = .12	−.18, *p* = .09	**.56, *p* < .01**

Client *N* = 30, Therapist *N* = 7; *p* values estimated using a Kenward–Roger adjustment; Largest fixed effect estimate in bold for each "special" MULTI subscale model.

Table 5. Fixed effects coefficients and *p* values from multilevel models with "termination" MULTI subscales predicting "session" MULTI subscales.

"Session" MULTI subscale	"Termination" MULTI Subscale Fixed Effect Coefficients and *p* values			
	Behavioral Change	Insight Oriented	Interpersonal Therapy	Common Factors
Behavioral Change	**.21, *p* = .52**	**.21, *p* = .53**	−.19, *p* = .33	.02, *p* = .93
Insight Oriented	−.21, *p* = .39	**.44, *p* = .10**	−.02, *p* = .89	.02, *p* = .93
Interpersonal Therapy	−.10, *p* = .81	.43, *p* = .33	**.44, *p* = .09**	.13, *p* = .73
Common Factors	−.02, *p* = .93	.01, *p* = .97	.08, *p* = .59	**.27, *p* = .22**

Client *N* = 27, Therapist *N* = 7; *p* values estimated using a Kenward–Roger adjustment; Largest fixed effect estimate in bold for each "special" MULTI subscale model.

controlling for the other predictors. For example, the relationship between "special" MULTI BC and "session" MUTLI BC can be interpreted as follows: A therapist who predicts using Behavioral Change techniques one point higher on the "special" MULTI than what they predict for their average client would also be expected to report an overall .56 point increase in their average reported Behavioral Change technique use on the "session" MULTI compared to their average client.

The regression coefficients on the diagonals represent relationships between corresponding subscales (i.e. BC predicting BC, IO predicting IO, etc.) and were the largest regression coefficients for the BC, IO, and CF subscales, indicating the strongest predictive relationship between corresponding subscales. These relationships provide preliminary evidence for therapists' ability to predict the interventions they will report using with a given client, relative to their other clients. For IPT, however, CF was the strongest predictor, indicating that amount of common factors predicted for a given client may predict the amount of IPT used, relative to other clients. Further, IO was a strong predictor of BC interventions, indicating that a therapist's reported BC intervention use may be related to their predictions of both BC and IO interventions.

Table 5 presents the models for "termination" MULTI subscales predicting "session" MULTI subscales (i.e. therapist recall). The regression coefficients can be interpreted similarly as above. The coefficients on the diagonal representing relationships between corresponding subscales were the largest for all four subscales, providing preliminary evidence for therapists' ability to recall their interventions. For several subscales, there were additional large coefficients. Therapists' use of BC was predicted not only by their recollection of BC, but also by their recollection of IO, a pattern seen in their predictions as well. Further, in recalling their IPT use, their retrospective report of IO use was a strong predictor, in addition to their retrospective report of IPT. Interestingly, for both IO and CF, the coefficient for the corresponding subscale was the only large coefficient, with all others being negative or close to zero.

Discussion

Driven by the interests of practicing clinicians as part of a PRN, the primary goal of this exploratory study was to investigate whether therapists could predict the interventions that they were going to use with a particular client seen in their private practice. Clinicians were also interested in assessing whether they could, at the end of therapy, accurately recall the interventions they used during the treatment of a given client. As an initial test of therapists' ability to predict and recall interventions, the findings should be considered with caution. Indeed, because of the small sample, our interpretations are based on the size of the fixed effects coefficients, and not on inferential statistics. With this caveat in mind, the results, as a whole, provide preliminary support for the therapists' ability to predict and recall interventions used in day-to-day practice with some accuracy and somewhat less so in terms of precision.

In contrast to the documented inaccuracy of clinicians' diagnostic and outcome predictions in other studies, our findings tentatively suggest that, early in treatment (after the completion of the third session), therapists can formulate reliable and generally discriminative expectations about the profiles of interventions they will use during the entire duration of therapy with a particular client. Each of the four derived subscales (Behavioral Change, Insight Oriented, Interpersonal Therapy, and Common Factors) measured by the "special" MULTI (assessing techniques predicted to be used during treatment from the third session onwards) showed medium-size coefficients (between .39 and .56) with its corresponding subscale of the aggregated "session" MULTIs (average of the techniques reportedly used at every session). With the exception of the IPT subscale, these correlations were higher than those with the other non-corresponding subscales. Although the coefficient sizes were lower (between .21 and .44), a similar pattern of results was obtained vis-à-vis therapists' ability to recall interventions used during the treatment of a given client. Other than the BC subscale, the correlations of each of the "session" MULTIs with its corresponding subscale on "termination" MULTIs (assessing techniques recalled to have been used during the course of therapy) was stronger than other correlations.

The results with respect to the prediction of intervention profiles suggest that after developing a sense of what it is like to work with a client (including re-evaluating her/his diagnosis) for the first few sessions, therapists do set up a treatment plan and are able to work consistently with such a plan. These results are particularly interesting in the context of our other findings showing few significant relationships between therapist's theoretical identification

and the intervention profiles they reported during therapy. Consistent with previous results obtained in a naturalistic setting (Boswell et al., 2010), none of the conceptually related correlations that might have been expected between theoretical orientation and reported intervention use was found to be significant – neither the report of IO techniques during treatment and the level of identification to psychodynamic and humanistic orientations, nor the report of BC during sessions and the level of identification to behavioral and cognitive approaches. Although not resting on direct evidence, this may reflect that in their day-to-day clinical work, therapists intervene in response and attuned to clients' individual needs (which it appears they can predict fairly accurately and precisely after three sessions) more than what could be predicted by their theoretical self-identity.

The results regarding the recall of intervention profiles also suggest that therapists are able to take a distance at the end of treatment and capture a reliable and generally discriminative picture of how they intervened during the entire therapy. Taken together, these two sets of findings provide preliminary evidence for the validity of clinically relevant cognitive skills. Considering the literature on therapist prediction and judgment (e.g. Garb, 2005; Grove et al., 2000), these results can be viewed as conceptually and empirically meaningful. Recent task forces (e.g. Brown et al., 2014; England et al., 2015) have called for the development and testing of measurement tools, including participant self-report, for monitoring the delivery of psychotherapy in routine settings. The results highlight both the feasibility and reliability of integrating a multi-dimensional technique measure in routine clinical practice.

In line with this, these findings may also be used to recommend non-invasive and non-time-consuming clinical practices. Therapists, irrespective of their theoretical allegiance, might want to consider filling out the MULTI, or other psychometrically sound measures of therapeutic interventions, to help them build more specific (i.e. technically precise) and comprehensive treatment plans early in therapy. They might also consider filling out such a measure at certain points during therapy in order to reflect on the process and progress of therapy, to check if they are consistently using all interventions that are in line with their treatment plan, and/or to use other techniques that would allow them to be more responsive to the client's needs. Filling out such a measure at the end of therapy might also increase their awareness of how they conduct therapy, and what, in their subjective experience, they might have used frequently and adequately enough (or not) for a given client. This type of information processing and reflection about therapy appears to be consistent with deliberate practice (see Chow et al., 2015), which is one of the very few factors for which research has provided support in explaining why some therapists are more effective than others (Wampold, Baldwin, Grosse Holtforth, & Imel, 2017).

Irrespective of its conceptual relevance and potential usefulness, this study should be viewed as a building block for future studies. Such studies could be conducted to assess if the same findings would replicate with a larger sample, including the relationship observed between IO and BC in both predictive and recall analyses. Additional studies could also address more complex and nuanced questions: Is a therapist's ability to predict interventions likely to be moderated by client problems (level of severity, types of diagnoses), therapist level of experience, and length of treatment? Could it be moderated by theoretical synchronicity: Do cognitive behavioral therapists show higher predictive ability when the treatment they expect to conduct is more in line with their preferred approach rather than other ones? Does consistency between expected and delivered treatment predict, or even mediate,

therapy outcome? Unfortunately, none of these questions could be directly answered in this study due to the small sample size.

Notable limitations of this study should be mentioned. In addition to our small sample size, an important limitation is methodological. Since the therapists self-completed all three sets of MULTIs, there is a clear possibility that the findings are the result of response biases. Because all therapists participated in the design of the study, they were not blind to its purposes. As such, they may have used (or reported using) given techniques during sessions because they predicted that they would, or they may have reported using specific profiles of interventions at the end of therapy because such profiles fit their recollection of what they filled out after therapy. However, it would seem that for report biases to be problematic, one would have to assume that when, for example, therapists completed the MULTI after each session, they always actively and accurately recall how they filled out the "special" MULTI for each given patient. Considering the depth and breadth of the relevant information retained through their clinical routines and across all of their clients, it seems reasonable to assume that the specificity of the correlations that we observed was due, in large part, to the impact of being guided, consistently and cohesively, by a treatment plan, rather than by memorization of the sheer amount of information contained in previous MULTI responses.

Furthermore, for report bias to be a major threat to the validity of the study, one would have to assume that therapists were invested in being correct in their prediction (as if testing a theory of theirs). This is, of course, possible, but therapists, to our knowledge, did not have a stake in "being right." Rather, as part of their desire to better understand therapy while conducting their routine practice within their PRN, they were simply (but meaningfully) curious about whether or not they could anticipate (and remember) what they do (or did) in treatment.

Another notable limitation was our reliance on therapists' perspectives. It should thus be recognized that the strength of the correlation observed between the three sets of MULTI is likely to be due in part to their common method. Related to the same limitation, this study did not include independent observational ratings of intervention use, which has, historically, been considered a gold standard method often used in highly controlled efficacy trials. Nevertheless, experts acknowledge that observational coding of treatment delivery is unlikely to be feasible in most routine treatment settings (Barkham, Stiles, Lambert, & Mellor-Clark, 2010). Consistent with this, the majority of our PRN clinicians expressed a preference for using self-report methods alone and refraining from recording session activities. Consequently, we elected to use self-report methods, which, as noted, has been identified as an area in need of further research.

Note

1. Twenty-nine clients declined to participate in the study, most frequently reporting concerns regarding confidentiality ($n = 9$), unnecessarily complicating treatment ($n = 7$), and just not being interested ($n = 6$). Fifty clients were not asked to participate because of their clinical state at the first interview. Forty of these un-recruited clients were judged to be cognitively impaired (with thirty-eight of them coming from a single therapist whose substantial subset of her practice involved conducting assessments with a geriatric population), seven were deemed to be in crisis during the first session, two exhibited psychotic/paranoid symptomology, and one client was physically incapable of filling the forms out. Of the clients who started the study, two dropped out of the study (one of them because of visual impairment) and continued treatment.

Acknowledgements

Disclosure statement

Funding

Notes on contributors

Louis G. Castonguay is Professor of psychology at the Pennsylvania State University.

Rebecca A. Janis is a doctoral student in the department of psychology at the Pennsylvania State University.

Soo Jeong Youn is a doctoral student in the department of psychology at the Pennsylvania State University.

Henry Xiao is a doctoral student in the department of psychology at the Pennsylvania State University.

Andrew A. McAleavey is an instructor in the department of psychiatry at the Weill Cornell Medical College, NY.

James F. Boswell is an assistant professor at the University of Albany, State University of New York.

Dever M. Carney is a doctoral student in the department of psychology at the Pennsylvania State University.

Mary A. Boutselis is a clinical psychologist in private practice in State College, PA.

Melora Braver is a clinical psychologist in private practice in State College, PA.

Nancy R. Chiswick is a clinical psychologist in private practice in State College, PA.

Neal A. Hemmelstein is a clinical psychologist in private practice in State College, PA.

Jeffrey S. Jackson is a clinical psychologist in private practice in State College, PA.

Richard A. Lytle is a clinical psychologist in private practice in State College, PA.

Marolyn E. Morford is a clinical psychologist in private practice in State College, PA.

Heather S. Scott is a clinical psychologist in private practice in State College, PA.

Catherine S. Spayd is a clinical psychologist in private practice in Duncansville, PA.

Mary O'Leary Wiley is a clinical psychologist in private practice in Altoona, PA.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R., … Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*, 341–382.

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010). Building a rigorous and relevant knowledge base for psychological therapies. In M. Barkham, G. E. Hardy, & J. Mellor-Clark (Eds.), *Developing and delivering practice-based evidence* (pp. 21–61). Chichester: Wiley.

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruiz, M. (2001). The Pennsylvania practice research network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy research. *Clinical Psychology: Science and Practice, 8*, 155–168.

Boswell, J. F., Castonguay, L. G., & Wasserman, R. H. (2010). Effects of psychotherapy training and intervention use on session outcome. *Journal of Consulting and Clinical Psychology, 78*, 717–723.

Brown, J., Scholle, S. H., & Azur, M. (2014). *Strategies for measuring the quality of psychotherapy: A white paper to inform measure development and implementation*. Report submitted to the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Contract number HHSP23320095642WC and task order number HHSP 23320100019WI. Washington, DC: Mathematica Policy Research. Retrieved from http://aspe.hhs.gov/daltcp/reports/2014/QualPsy.cfm

Castonguay, L. G. (2011). Psychotherapy, psychopathology, research and practice: Pathways of connections and integration. *Psychotherapy Research, 21*, 125–140.

Castonguay, L. G., Boswell, J. F., Zack, S., Baker, S., Boutselis, M., Chiswick, N., … Grosse Holtforth, M. (2010). Helpful and hindering events in psychotherapy: A practice research network study. *Psychotherapy: Theory, Research, Practice, and Training, 47*, 327–344.

Castonguay, L., Nelson, D., Boutselis, M., Chiswick, N., Damer, D., Hemmelstein, N., … Borkovec, T. (2010). Psychotherapists, researchers, or both? A qualitative analysis of psychotherapists' experiences in a practice research network. *Psychotherapy: Theory, Research, Practice, and Training, 47*, 345–354.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice-oriented research: Approaches and application. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 85–133). New York, NY: Wiley.

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy, 52*, 337–345. doi: 10.1037/pst0000015

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York, NY: Free Press.

Elliott, R., & Morrow-Bradley, C. (1994). Developing a working marriage between psychotherapists and psychotherapy researchers: Identifying shared purposes. In P. F. Talley, H. H. Strupp, & S. F. Butler (Eds.), *Psychotherapy research and practice: Bridging the gap* (pp. 124–142). New York, NY: Basic Books.

England, M. J., Butler, A. S., & Gonzalez, M. L. (Eds.). (2015). *Psychosocial interventions for mental and substance abuse disorders: A framework for establishing evidence-based standards*. Washington, DC: National Academies Press.

Garb, H. N. (1998). *Case formulation*. Washington, DC: American Psychological Association.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1*, 67–89.

Garb, H. N., & Boyle, P. A. (2015). Understanding why some clinicians use pseudoscientific methods: Findings from research on clinical judgment. In S. D. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology* (2nd ed., pp. 19–41). New York, NY: Guilford Press.

Garland, A. F., & Schoenwald, S. K. (2013). Use of Effective and Efficient Quality Control Methods to Implement Psychosocial Interventions. *Clinical Psychology Science and Practice, 20*, 33–43.

Grove, W. M. & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of clinical psychology, 61*, 155–163.

Hatfield, D., McCullough, L., Frantz, S. H., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy, 17*, 25–32.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146–159.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.

Koerner, K., & Castonguay, L. G. (2015). Practice-oriented research: What it takes to do collaborative research in private practice. *Psychotherapy Research, 25*, 67–83.

McAleavey, A. A., & Castonguay, L. G. (2014). Insight as a common and specific impact of psychotherapy: Therapist-reported exploratory, directive, and common factor interventions. *Psychotherapy, 51*, 283–294.

McCarthy, K. S., & Barber, J. P. (2009). The multitheoretical list of therapeutic interventions (MULTI): Initial report. *Psychotherapy Research, 19*, 96–113.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295–314. doi: 10.1007/s10648-014-9287-x

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Washington, DC: American Psychological Association.

Meehl, P. E. (1973). *Psychodiagnosis: Selected papers*. Kingsport, TN: Kingsport Press.

Orlinsky, D. E., Gerin, P., Davis, J., Davis, M., Ambuhl, H., Willutzski, U., & Iahns, J. F. (1991). *The international study of the development of psychotherapists: Conceptual organizational of the common core questionnaires* (Unpublished manuscript). Chicago, IL: University of Chicago.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2013). R development core team: Linear and nonlinear mixed effects models. *R package version, 3*, e107.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Salisbury, T. N. (2015). *Predicting youth treatment failure: An investigation of clinical versus actuarial judgment*. Retrieved from PsycINFO. http://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com/docview/1721926585?accountid=13158

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*, 593–602.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Rush, J. D. (2009). The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *Counseling Psychologist, 37*, 350–399.

Tracey, T. J. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: An elusive goal? *American Psychologist, 69*, 218–229.

Wampold, B. E., Baldwin, S. A., Grosse Holtforth, M., & Imel, Z. (2017). There are therapist effects but what characterizes effective therapist? In L. G. Castonguay & C. E. Hill (Eds.), *Therapist effects: Toward understanding how and why some therapists are better than others* (pp. 37–54). Washington, DC: American Psychological Association book.

Wiggins, J. S. (1973). *Personality and Prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.