An Application of the Three-Step Test-Interview (TSTI) in the Validation of the Relational Depth Frequency Scale

Journal Item

For guidance on citations see FAQs.

Version: Version of Record

oro.open.ac.uk

Article

An Application of the Three-Step Test-Interview (TSTI) in the Validation of the Relational Depth Frequency Scale

Gina Di Malta[1] **iD, Mick Cooper**[2]**, Joel Vos**[3] **iD, and Kees van der Veer**[4]

Abstract

The main objective of this study is to evaluate the utility of a new qualitative scale development methodology—Three-Step Test-Interview (TSTI)—in its first application in the validation of a psychotherapy scale: The Relational Depth Frequency Scale (RDFS). The TSTI is a cognitive pretesting method designed to uncover potential problems in scale construction. The RDFS is a six-item unidimensional scale of in-depth therapeutic relating, designed for use in large-scale outcome studies. Following the creation of an item pool and "expert ratings," a purposive sample of four therapists and four clients (five females, three males, mean age: 49 years) was recruited to take part in the TSTI with the view to refine the original 36-item RDFS prior to psychometric exploration. Structured observations pointed to problems in test-takers' patterns of responses in relation to theoretical knowledge of the relational depth construct. Issues uncovered and addressed included some misinterpretations of instructions and items, redundant content, double-

[1]The Open University, Milton Keynes, UK
[2]University of Roehampton, London, UK
[3]Metanoia Institute, London, UK
[4]VU Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Gina Di Malta, The Open University, Milton Keynes, MK7 6AA, UK.
Email: gina.dimalta@open.ac.uk

barreled items, and test-takers' reactions to intimate content wording. The method supported the refinement of the RDFS including amendment to its instructions and the removal of problematic items. TSTI results produced knowledge on the scale which could not be captured with statistical methods.

Keywords

Three-Step Test-Interview, validity, scale development, relational depth

Valid and reliable psychological assessment is central in the advancement of psychotherapy research and the delivery of evidence-based treatment. The focus of the present report is to describe and evaluate the utility of Three-Step Test-Interview (TSTI) as a methodological step in the validation of a psychotherapy scale: The Relational Depth Frequency Scale (RDFS). Initial stages of scale development are fundamental in ensuring the content validity of new measures. In scale development studies, *content validity*—an aspect of the broad concept of validity, which refers to how well a test measures what it is supposed to measure, and how well it taps into the various aspects of a construct (Haynes et al., 1995)—is typically initially sought with the creation of an item pool based on researchers' expert knowledge and review of literature that highlights the salient features on the construct (Carpenter, 2018). This step of scale development is followed by *expert ratings* of items—where a set of experts on the construct rate each item for their suitability to a scale—to refine the scale (e.g., Devellis, 2016).

In addition to these two steps, further steps can contribute to the content validity of a scale. This has included, for instance: the piloting of the scale using a "non-applicable" item option with the possibility for participants to leave a comment following each item (e.g., Shin et al., 2016). While this method has the advantage of allowing for large numbers of participants to provide feedback on each scale item, such feedback can be difficult to interpret and analyze. In some instances, a comment may not capture the true meaning of the participant's response in relation to the theoretical underpinning of the scale item. A methodology of scale development has been developed, which takes into account test takers' feedback while also giving the possibility of relying on an expert's interpretation of such feedback as it is being delivered in specialized cognitive interviews.

Three-Step Test-Interview

The TSTI is a cognitive pretesting method, which has been used as a tool in validation studies of self-completion questionnaires. Such a tool is particularly

useful in the development of new instruments to extract and organize in-depth feedback on scales and their items (Carpenter, 2018). More specifically, it enables localizing problems in the response process, their effects as well as their causes (Hak et al., 2006). The novel methodology was piloted in the development of new scales and showed that it could help identify problems resulting from a mismatch between theory and participants' understanding of items (Busse & Ferri, 2003; Hak et al., 2008; Jansen & Hak, 2005).

With regard to validity, the TSTI is suited to test two aspects of content validity of new scales. *Logical validity* is typically based on a rigorous assessment of items in a scale (Rubio et al., 2003). It is associated with a scale having a clear language, and content that is cognitively accessible to test takers (e.g., Devellis, 2016). In addition, a central part of the TSTI method relies on a researcher's observation of participants' reactions to the scale items, as they concurrently think aloud as they respond to items. This aspect informs researchers about the *face validity* of the scale—indicating that the scale and scale items appear to be valid.

The TSTI method consists of a series of three stages aimed at gathering different forms of feedback on a scale. The initial stage is the *concurrent think aloud*. In this stage, participants are asked to say their thoughts out loud while completing the measure that is being developed. This stage is aimed at making the test-takers' thought processes observable to the researcher. In addition, the researcher collects other observational data, including any test-takers' pauses, reactions to scale items, signs of fatigue, or any other relevant observable behavior. The concurrent think aloud is followed by a *focused interview* where the researcher elicits direct feedback on any gap observed in the thought process from the previous stage. This step is aimed at gaining a full account of any thoughts that seemed incomplete in the participant's report of their thoughts. Finally, the last step consists of a *semi-structured interview* to elicit the participant's reflections, opinions, and experiences of completing the scale. The semi-structured interview includes general questions on taking the scale such as: "How was it for you completing the scale?" and "Can you give any feedback or recommendations on how we may improve this scale?" It also includes questions that a specific to the instructions of the scale, such as "Were the instructions clear?" In addition, a full explanation of participants' reactions and thought process to each of the scale items in stage one is elicited. Following the interview, a theory-informed researcher interprets the test-taker's responses and develops an analysis aimed at refining the scale structure and content (Hak et al., 2006).

So far, the TSTI methodology has been applied in the validation of scales assessing social phenomena: for example, the Illegal Alien Scale elicits individuals' beliefs and values about illegal aliens (Hak et al., 2006); the ageing

scale elicits experiences of ageing in patients with rheumatic diseases (Bode & Jansen, 2013); and the cumulative scale on fear-based xenophobia elicits people's attitude toward foreigners, in which the TSTI was used to expand the "qualitative validity" of the already psychometrically tested instrument (van der Veer et al., 2013). The TSTI was also applied to develop scales assessing health phenomena: for example, the Alcohol Consumption Scale elicits factual reports around alcohol consumption (Jansen & Hak, 2005); the St George's Respiratory Questionnaire for chronic obstructive pulmonary disease patients is a self-report measure for symptoms of chronic obstructive pulmonary disease (Paap et al., 2016); and finally it was applied to test the validity of the Tampa Scale for Kinesiophobia—one of the most frequently employed measures for assessing pain-related fear in back pain patients (Pool et al., 2010). In these applications, the TSTI has primarily revealed problems with misunderstanding of instructions, difficult formulations and composite items, redundant items, irrelevant content, or inadequate response option.

The present study is the first to describe and evaluate an application of the TSTI for a psychotherapy measure. A psychotherapy scale is different from aforementioned scales in that its content is likely to be less factual and more theoretical, and elicit more emotional, interpersonal, and subjective content. The RDFS is a self-report measure of in-depth relating, designed to assess the temporal frequency of moments of deep connection over the course of therapy (Di Malta et al., 2020). A moment of *Relational depth* is an intense experience of mutual empathy and congruence that can arise between therapist and client in therapy. It is defined as follows: "a state of profound contact and engagement between two people, in which each person is fully real with the Other, and able to understand and value the other's experiences at a high level" (Mearns & Cooper, 2005, p. xii). Relational depth experiences—including heightened empathy, acceptance, congruence, and a willingness to take risks—were similar in clients and therapists with some differences associated with their respective role (Cooper, 2005; Knox, 2008). The RDFS was developed to assess the frequency of relational depth moments over the course of sessions or a whole therapy. The measure was created to support new research avenues in the field of therapeutic relating. Particularly, the scale was designed to assess the relationship between relational depth and therapeutic outcomes in large-scale outcome research and as a routine process measure in clinical settings. The TSTI follows the item generation and ratings, and precedes the empirical testing stage to test the factor structure, reliability, and construct validity of the scale (Di Malta et al., 2020).

The aim of this article is to describe and reflect on an application of the TSTI in the validation of a psychotherapy scale. In particular, it is looking at

the utility of the TSTI method in assessing the validity of the RDFS and in its application in the refinement of the scale structure and its items.

Method

The research project was submitted for ethics consideration under the reference PSYC 15/ 164 in the Department of Psychology and was approved under the procedures of the University of Roehampton's Ethics Committee on 20.05.15.

Relational Depth Frequency Scale (36 Items)

The 36-item RDFS—in development at the TSTI stage (following the creation of an item pool and expert ratings, and preceding psychometric exploration)—consisted of a short definition, then a set of instructions, and an opening stem as per the box below:

> Relational depth is defined as "a state of profound contact and engagement between two people, in which each person is fully real with the Other, and able to understand and value the other's experiences at a high level".
> Please think of a relationship with a client [or therapist] and select how frequently you have experienced the moments described in each item. Each item follows the statement:
> "Over the course of my therapy with my client [or therapist], there were moments where . . . "

Items included for instance, "I felt a clarity of perception between us" and "I experienced a meeting that was beyond words." Participants were asked to rate each item on a 5-point Likert-type scale, where 1 = *not at all*, 2 = *only occasionally*, 3 = *sometimes*, 4 = *often*, and 5 = *most or all of the time*. This subjective frequency scaling was used to reflect the subjectivity of recall of relational depth experiences. The scaling was aligned with the Likert-type scale of the well-validated Clinical Outcome in Routine Evaluation used to assess psychological distress in mental health services in the United Kingdom (CORE-OM; Evans et al., 2002).

Sample

Inclusion criteria were clients or therapists who were 18 years old and above, and who had a minimum of six sessions in therapy. We endeavored to recruit a sample composed of approximately equal numbers of therapists and clients who reflected the general population's familiarity with relational depth. To

achieve this, a purposive sample was recruited. Four clients and five therapists responded to the advert. Half of the participants—three clients and one therapist—were selected because they were not familiar with the concept of relational depth prior to the study. This approach to sampling was undertaken to limit a possible response bias (associated with attracting respondents who would be particularly "relationally" oriented) and to achieve a sample more representative of current client and therapist populations in the United Kingdom. Priority was then given to participants in the order that they responded to the advert.

Participants

We selected eight participants, three males and five females, aged ranging from 26 to 90 years old. The mean age was 49 and median 42. More than half the participants (five out of eight) were from a White British background, two were from a mixed ethnic background, and one was White European. Therapist participants in this sample were three qualified mental health practitioners and one was still in training. Three of the therapist participants were humanistic in orientation, and one was integrative. Half of the client participants were counselling psychologist trainees, two clients were not affiliated with the mental health professions. Two of the clients were in psychoanalytic psychotherapy, one was in Lacanian therapy and one did not know the orientation of her therapist. All participants were educated to at least a Bachelor's degree. Half of the participants—three clients and one therapist—were not familiar with the concept of relational depth prior to the study.

Procedure

An advert was sent via e-mail to students and staff members from the host University and posted on psychology community social media pages. This advert informed about the development of a new scale on therapeutic relating. All interviews were conducted by the first author. At the start of the interview, the researcher gave an overview of the task and obtained informed consent. Then detailed instructions were given according to Hak et al.'s (2008) standard procedures. In a first instance, participants were invited to practice with an exercise on the "concurrent think aloud," while the researcher provided feedback on the task (e.g., Willis, 2004). When agreed that enough exercise had been done, participants started filling the RDFS while "thinking aloud." This was then followed by the "focused interview"—where participants were prompted to go back over items where there were hesitations, and to fill in the thoughts that appeared not to be fully expressed. The last step

was semi-structured interviews, where participants were asked specific questions addressing their response behaviors and thoughts around the scale items, including understanding and definitions of terms, paraphrasing, and rewording of items. Interviews were audio recorded and systematic notes were gathered during each stage of the interview.

Analysis

Interviews were analyzed using thematic analysis, a method of qualitative analysis used to categorize patterns of responses in the data (Braun & Clarke, 2006). Patterns arising from the data were explored in the light of existing theory on relational depth. The analysis was guided by observations of items that posed the most problems to participants and/or appeared inconsistent with theory on relational depth.

The first author conducted the analysis following Braun and Clarke's (2006) six steps of thematic analysis. Initially, all detailed notes taken during interviews were read several times. This was followed by the coding of problematic areas in each interview report. Codes were then grouped into categories across participants. Categories were then reviewed by the first and second author. Finally, a report was produced.

Results

Interviews lasted between 60 and 90 minutes. All participants understood instructions on the TSTI and completed all tasks. The TSTI methodology supported the identification of four categories of problem areas in the original 36-item RDFS (see Table 1).

Lack of Reference to the Opening Stem

Observations revealed that one recurring problem was participants' lack of reference to the opening statement when reading each item. The latter serves to put each item in a phenomenological context. As a result, participants were sometimes unsure of how to select their answers. Some participants viewed some of the items as reflecting the quality of the whole therapeutic relationship, as opposed to representing discrete moments in the course of their therapy.

8

Table 1. TSTI Problem Areas and Amendments.

Relational Depth Frequency Scale (36 items)	Amendments
Instructions: *Please think of a relationship with a client [or therapist] and select how frequently you have experienced the moments described in each item.* Each item follows the statement: "Over the course of my therapy with my client [or therapist], there were moments where . . ."	Based on clients' responses and to limit possible doubts, confusion or distress in future uses of the scale, instructions were amended to include the sentence: *There are no right or wrong answers, individuals relate differently.* In addition, the opening stem's font was increased in order to make it more visible.
1. I experienced an intense connection with him/her	
2. We felt intensely real with each other (4)[a]	Removed because the word "real" created confusions in most participants.
3. We were connected on a level that I rarely experience (5)[a]	Removed as some participants were confused by the adverb "rarely" being redundant with the frequency Likert-type scale
4. I experienced a very profound engagement with her/him	
5. We were both completely genuine with each other (3)[a]	Removed as most participants perceived "genuineness" as a mundane characteristic of the therapeutic relationship
6. I experienced what felt like true mutuality	
7. I felt completely real in relation to him/her (4)[a]	Removed because the word "real" created confusions in most participants.
8. I felt I was being understood beyond my words (2)[a]	Removed as the phrase "beyond my words" is repeated in Items 16 and 31
9. We were deeply connected to one another	
10. We felt accepting of one another (3)[a]	Removed because most participants answered it easily or commented that being "accepting" was a given of a therapeutic relationship
11. I could see we had mutual understanding (2)[a]	Removed as the word "understanding" is repeated in Item 22

(continued)

Table 1. (continued)

Relational Depth Frequency Scale (36 items)	Amendments
12. I felt a deep empathy between us (4)[a]	Removed as it created confusion in most client participants
13 I felt completely immersed in the relationship (2)[a]	Removed as the word "immersed" is repeated in Item 20
14. I felt a clarity of perception between us	
15. The connection between us was much stronger than in other moments (1)[a]	Removed because the two composite parts created confusion in some participants
16. The level of our connection seemed to go beyond words alone (2)[a]	Removed as the phrase "beyond words" is repeated in Items 8 and 31
17. I felt an overall warmth between us (3)[a]	Removed as the word "warmth" was perceived as a mundane characteristic in therapy by most participants
18. The moment between us felt very meaningful (5)[a]	Removed as a participant noted that the word "moment" is redundant with the instructions
19. I felt intensely present with him/her	
20. We were immersed in the present moment (2)	
21. I felt completely understood (2)[a]	Removed due to the word "understood" being repeated in Item 22
22. There was a deep understanding between us (2)	
23. It felt like a shared experience	
24. I felt fully attuned to him/her (4)[a]	Removed because some participants were hesitant and unsure about the meaning of "attuned"
25. I felt that we fully acknowledged each other (2)[a]	Removed due to the word "acknowledged" being repeated in Item 34
26. I felt we deeply trusted each other (3) [a]	Removed as "trust" was perceived as characteristic of the therapeutic relationship by some participants
27. I experienced a deep intimacy between us (4)[a]	Removed due to confusion with the meaning of the word "intimacy"

(continued)

9

Table 1. (continued)

Relational Depth Frequency Scale (36 items)	Amendments
28. I felt deeply valued by him/her (4)[a]	Removed as the word "valued" created confusion in some participant and did not reflect experiences of relational depth for therapist participants familiar with the construct
29. We connected on a human level	
30. I experienced a deep sense of encounter	
31. I experienced a meeting that was beyond words (2)	
32. I felt like we were totally in-the-moment together (2)	
33. We felt really close to each other	Amended in response to a couple of clients' discomfort with having to "speak for their therapist." Item became: "I felt we were really close to each other"
34. We truly acknowledged each other at a very deep level (2)	Amended in response to a couple of clients' discomfort with having to "speak for their therapist." Item became: "I felt we truly acknowledged each other at a very deep level"
35. I felt our relationship provided a greater depth, different to other relationships, that helped me to grow (1)[a]	Removed because its three composite parts created confusion in some participants.
36. We were completely open with each other	Amended in response to a couple of clients' discomfort with having to "speak for their therapist." Item became: "I felt we were completely open with each other"

Note. (1) = Double-barreled; (2) = Repetitions; (3) = Mundane; (4) = Confusion; (5) = Redundant.
[a]Indicates items that were removed as a result of the analysis.

The Scale Evoked a Small Amount of Distress or Frustration

General observations revealed that the scale evoked a small amount of distress in a couple of client participants. These clients expressed concerns or worry as they felt they had not experienced relational depth in their therapy. This made them question the quality of their therapy.

Similarly, two therapists who were familiar with the construct of relational depth reported some frustration when filling the scale. One humanistic therapist felt the format of the scale did not reflect her idiosyncratic experiences of relational depth. She felt her experience was not fully captured in the scale items, and some items did not reflect her experience. The transpersonal therapist said he felt that relational depth could not be measured nor approached from a positivist standpoint.

Incongruent Responses to Mutual Items

One observation in participants' responses was that clients reacted to the mutual items differently than therapists did. Three client participants reported it was difficult to assert the perceived mutuality of their feelings in their relationship with the therapist. This experience was strongly evident for all items which opened with "We felt . . . " One client argued that it was impossible for her to answer such items, as this would mean reading her therapist's mind or speaking for them. Therapists, on the other hand, appeared more comfortable to answer mutual items for their clients and to infer the mutuality of their experience.

Patterns of Responses Highlighting Problematic Items

Five categories were identified in patterns of responses. The problem areas for each item are summarized in Table 1 and described as follows:

*Double-Barreled (**n** = 2).* The "double-barreled" item category refers to where the wording of the item had more than one part to an item. There were two items highlighted during the think aloud task. Item 35—"I felt our relationship provided a greater depth, different to other relationships, that helped me to grow" appeared to bring confusion to most participants because it had three component parts to one item. Item 15 brought confusion to some participants as it had two component parts and was also redundant with the opening stem.

*Redundant (**n** = 2).* The "redundant" item category is where some of the wording in an item was redundant with the opening statement or the scaling. For instance, in Item 3—"I felt we were connected on a level that I rarely experience," "rarely" is redundant with the frequency Likert-type scaling. Similarly, Item 18 repeats the word "moment" which is part of the opening stem.

*Confusion (**n** = 5).* "Confusion" items refer to the meaning of a word being understood differently by different participants or causing doubt or uncertainty. Items 2 and 7, which contained the word "real," were confusing for most participants. More specifically, six participants questioned the meaning of the word "real" in the thinking aloud task. Participants' definitions were explored in semistructured interviews and participants suggested there may be a lack of substance or clear definition to the word "real."

Item 12—"I felt a deep empathy between us" was confusing mostly for client—participants who did not expect it to be the role of the client to feel empathy for their therapist. For instance, one client in psychoanalytic therapy said, "I had no empathy towards him," and another one said, "empathy was mostly coming from her" (client in psychoanalytic therapy). One humanistic therapist also noted that clients may not have empathy for their therapist: "I cannot answer because of the mutuality of the empathy."

In addition, the think aloud and structured interview highlighted that three client participants offered contradictory definitions to the term "attuned" in Item 24—"I felt fully attuned to him/her." One client in Lacanian therapy linked the item to connection without intimacy: "it means engaged but not related to emotions or intimacy." Another client, in unknown therapy, described it as a form of unrealistic empathy: "it's unrealistic, it means feeling their feelings." A third client in psychoanalytic therapy, on the other hand, saw it as a mundane form of connection: "it's hard to be with someone if not attuned to them."

There was a discrepancy between clients' and therapists' understanding of Item 27—"there was a deep intimacy between us." The item elicited strong reactions in two clients who associated "deep intimacy" with physical closeness, touch, or sex. Three of the therapists, on the other hand, found that a "deep intimacy" reflected well "the aim of therapy, and [was] not necessarily related to touch" (humanistic therapist) although it was "unexpected, rare and precious" (transpersonal therapist).

Item 28—"I felt deeply valued by him/her" led three participants to question the meaning of value. For two therapists who were familiar with the construct of relational depth, the connotation of "value" seemed not to fit

with the essence of relational depth. One humanistic therapist reported for instance: "not a word I would use," another humanistic therapist said, "deeply valued doesn't feel as intimate, it implies a distance." An integrative therapist, on the other hand, appeared to view the item as fitting their experience, they referred to their client: "she said it was a valuable relationship."

*Repetition (**n** = 10).* The "repetition" category referred to items that had the same or very similar wording. This appeared to cause boredom and feelings of fatigue in participants as observed in the think aloud task. Five participants also explicitly said the scale was too long and some pointed out items with similar wording. Items 11, 21, and 22 all used the wording "understanding." Items 25 and 34 were constructed around the verb "acknowledge." Items 8, 16, and 31 all used the phrase "beyond words." Similarly, Items 13 and 20 both contained the word "immersed."

In addition, Item 13—"I felt completely immersed in the relationship" was specifically problematic among interviewees, one client in unknown therapy stated, "immersion is like having a bath together, it's not particularly healthy." Another client, in psychoanalytic therapy, said, "I don't like that phrase, it's like suffocating, losing your own person." One humanistic therapist participant had similar reactions to Item 13, they said, "immersed implies lost."

*Mundane (**n** = 4).* The "mundane" category referred to items that reflected mundane characteristics of the therapeutic relationship, as opposed to more distinct and intense moments of relational depth. Observations revealed that most participants answered these items very readily, sometimes commenting that such experiences were "givens" of a therapeutic relationship. For these items, participants also tended to select the highest frequency labels on the Likert-type scale suggesting that these experiences were common.

Discussion

This research presents the first application of the TSTI in the development of a psychotherapy scale. The TSTI—including the concurrent think aloud, focused interview, and semi-structured interview—revealed a range of observations and responses, which were categorized in four main problem areas around the structure and design of the scale, and five problematic patterns of responses in the scale items. TSTI findings are interpreted in the light of an integration of theory on measurement scales and theory on relational depth, and highlight limitations in the RDFS. The scale and items were amended in the light of the TSTI results as per Table 1.

The observation of the think aloud task pointed to participants' lack of reference to the opening item. Such observation revealed potential problems for the validity of the scale. In effect, the opening statement serves as a reference point to anchor each item in a phenomenological context. Here, findings suggest that test-takers would answer differently depending on whether they rated the depth of their relationship or the frequency of moments of relational depth. This could result in a scale which does not systematically measure what it is supposed to measure (Develing, 2016). Such observation may translate in finding poor reliability in a statistical analysis; however, statistical analysis would not point to the causes of such finding. Thus, the scale structure was amended to emphasize the opening statement by using a larger and bolder font in contrast with instructions and items.

Similarly, observations also revealed that the RDFS evoked a small amount of distress in clients and frustration in therapists. This raised an ethical concern as the scale was intended to be used in clinical settings with vulnerable individuals. Client participants' specific concern was that they were not experiencing relational depth in their therapy. It is possible that the inclusion of clients who had only 6 sessions may not be enough time for relational depth experiences to arise in therapy. Furthermore, research on relational depth suggests that these moments are rare (McMillan & McLeod, 2006). One study defines moments of relational depth as present in up to one-third of significant events in therapy (Wiggins et al., 2012). Here, TSTI observations enabled to detect this experience in test-takers and take steps to normalize it by amending instructions. This was done by including the statement: "There are no right or wrong answers, individuals relate differently."

In addition, TSTI observations also allowed researchers to interpret participants' reaction based on the context in which they emerged. For instance, the frustration experienced by two therapist participants could be interpreted in several ways. While these reactions occurred with humanistic therapists who may be less likely to welcome positivist methods of measurement, they also reflected an existing debate around the value of quantifying relational depth (Cooper, 2013). The argument is that the empirical examination of relational depth is antithetical to a philosophical premise of relational depth based on I–thou relations as opposed to I–it reductionism (Buber, 1947). These questions were explored, it was decided that no amendments would be made to the scale with regards to this point. In this study, reactions around experiences not being reflected in the proposed items pointed to this unique limitation in the scale.

Furthermore, TSTI observations of participants could point to questions around the theoretical definitions of the construct being measured. In this study, observations across clients and therapists showed that each group

responded differently to mutual items. The psychotherapy literature describes small differences between clients' and therapists' experiences of relational depth associated with their respective roles in the relationship (e.g., Knox, 2008). More specifically, clients' accounts differed because they tuned-down the aspect of "mutuality" due to more focus on the self (McMillan & McLeod, 2006). These differences captured in the interviews brought questions around the scale's initial design to measure a single experience for therapists and clients. For the present scale, researchers chose to amend the items starting with "We . . . " to "I felt we . . . " to account for clients' perspective on mutuality. In addition, this would improve homogeneity in therapist's and clients' perceptions of items as per the stated aim of the RDFS to capture a single experience. For the present scale development, researchers chose to amend or remove items that emphasized clients and therapists' differences. As a result, the TSTI highlighted a limitation in the content validity of the RDFS: It does not account for differences in therapists' and clients' experiences as identified in the literature (McMillan & McLeod, 2006).

Hence, the TSTI analysis revealed five clusters of items identified in patterns of responses. Three of the clusters—"double-barreled," "redundant," and "repetitions"—highlighted clear structural issues in the scale and items. These pointed to generic problems in scale development, which were not identified in the prior two stages of item generation and expert ratings.

The two other clusters—"mundane" and "confusion"—raised questions at a conceptual level, soliciting reflection on the theory of relational depth. Items in the confusion cluster tended to reveal misunderstandings around definitions of words. For instance, participants reported different definitions for the word "attuned." Other words such as "intimacy" and "immersed" provoked strong opposing reactions, including negative interpretations. The word "real," often used in phenomenological descriptions of relational depth, could not be defined by several participants (e.g., Knox & Cooper, 2011; Mearns & Cooper, 2005). These terms used in phenomenological descriptions of relational depth appeared to lose their intended meaning when used as part of a measurement scale. These items were removed as per Table 1, as they may compromise the reliability and validity of the scale.

Finally, TSTI findings highlighted another question around the representativeness of relational depth content for items in the "mundane" category. These items were easily answerable items describing mundane experiences that may occur in a therapeutic relationship. The item "I felt we were accepting of one another," for instance, may be a common experience in a therapeutic relationship. In the person-centered approach, acceptance or "unconditional positive regard" is an integral part of the therapeutic approach (Rogers, 1961). The RDFS was designed to assess recollections of the frequency of

intense moments of relational depth rather than capture a gradient of depth for which easily answerable items reflecting the core conditions could be included. In this case, the TSTI findings pointed towards removing mundane items from the scale. All four mundane items were removed.

One limitation of this study was the size and therefore also the homogeneity of the sample. The majority of participants were White Caucasian and educated to at least a Bachelor's degree. This could have led to a possible ethnocentric emphasis at this stage of the scale development or to a possible overcomplexity in the language that was not captured during the TSTI process. This is a broader limitation when using the TSTI method, as it tends to rely on small samples. Such small samples are unlikely to ever be diverse enough or capture a wide range of possible problems in a scale (O'Reilly & Parker, 2013). Another limitation of the TSTI, also mentioned in previous studies, is that it is a time-consuming methodology. In this specific use of the TSTI for a psychotherapy scale, the wealth of interpretations also relied on a researcher's expertise in the construct being measured.

The application of the TSTI to the development of a new psychotherapy scale was overall useful as it provided insights, which could not be captured through other scale development methods (e.g., expert ratings or statistical methods). In terms of its similarities with its use on scales of social and health phenomena, TSTI findings included the detection of misunderstanding of instructions, difficult formulations, composite items, and redundant items. One difference was that the researcher relied on specialized theoretical knowledge of the psychological construct of relational depth. As seen in this study, psychotherapy topics could also be more personal, intimate, and subjective; and the TSTI was a sensitive tool in highlighting possible ethical dilemmas and considerations. Future psychotherapy scale development research may similarly benefit from including a TSTI step. In addition, it may be useful to look at TSTI results in combination with exploratory factor analysis or other psychometric techniques, depending on the theoretical concepts to be measured, to gain insights and better understand the meanings of the results of measurement. The TSTI is a method, which promotes direct patient involvement in the creation of new scales. This is in line with current research practices around knowledge exchange and patient and public involvement in psychotherapy research.

Declaration of Conflicting Interests

Funding

ORCID iDs

Gina Di Malta [iD] https://orcid.org/0000-0002-6254-6522
Joel Vos [iD] https://orcid.org/0000-0002-7961-2354

References

Bode, C., & Jansen, H. (2013). Examining the personal experience of aging scale with the Three Step Test Interview. *Methodology*, *9*(3), 96-10. https://doi.org/10.1027/1614-2241/a000071

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Buber, M. (1947). *Between man and man* [Trans. R. G. Smith]. Macmillan.

Busse, A., & Ferri, R. B. (2003). Methodological reflections on a three-step-design combining observation, stimulated recall and interview. *Zentralblatt für Didaktik der Mathematik*, *35*(6), 257-264. https://link.springer.com/article/10.1007/BF02656690

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, *12*(1), 25-44. https://doi.org/10.1080/19312458.2017.1396583

Cooper, M. (2005). Therapists' experiences of relational depth: A qualitative interview study. *Counselling and Psychotherapy Research*, *5*(2), 87-95. https://doi.org/10.1080/17441690500211130

Cooper, M. (2013). Experiencing relational depth in therapy: What we know so far. *Relational depth: New perspectives and developments*, pp. 62-76.

Devellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage.

Di Malta, G., Evans, C., & Cooper, M. (2020). Development and validation of the Relational Depth Frequency Scale. *Psychotherapy Research*, *30*(2), 213-227. https://doi.org/10.1080/10503307.2019.1585590

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE—OM. *British Journal of Psychiatry*, *180*(1), 51-60. https://doi.org/10.1192/bjp.180.1.51

Hak, T., van der Veer, K., & Ommundsen, R. (2006). An application of the Three-Step Test-Interview (TSTI): A validation study of the Dutch and Norwegian versions of the "Illegal Aliens Scale." *International Journal of Social Research Methodology*, *9*(3), 215-227. https://doi.org/10.1080/13645570600656264

Hak, T., van der Veer, K., & Jansen, H. (2008). The Three-Step Test-Interview (TSTI): An observational instrument for pretesting self-completion questionnaires. *Survey Research Methods*, *2*(3), 143-150. https://doi.org/10.18148/srm/2008.v2i3.1669

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247. https://doi.org/10.1037/1040-3590.7.3.238

Jansen, H., & Hak, T. (2005). The productivity of the three-step test-interview (TSTI) compared to an expert review of a self-administered questionnaire on alcohol

consumption. *Journal of Official Statistics*, *21*(1), 103. https://repub.eur.nl/pub/20960/

Knox, R. (2008). Clients' experiences of relational depth in person-centred counselling. *Counselling & Psychotherapy Research*, *8*(3), 182-188. https://doi.org/10.1080/14733140802035005

Knox, R., & Cooper, M. (2011). A state of readiness: An exploration of the client's role in meeting at relational depth. *Journal of Humanistic Psychology*, *51*(1), 61-81. https://doi.org/10.1177/0022167810361687

McMillan, M., & McLeod, J. (2006). Letting Go: The client's experience of relational depth. *Person-Centered & Experiential Psychotherapies*, *5*(4), 277-292. https://doi.org/10.1080/14779757.2006.9688419

Mearns, D., & Cooper, M. (2005). *Working at relational depth in counselling and psychotherapy*. Sage.

O'Reilly, M., & Parker, N. (2013). "Unsatisfactory Saturation": A critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research*, *13*(2), 190-197. https://doi.org/10.1177/1468794112446106

Paap, M. C., Lenferink, L. I., Herzog, N., Kroeze, K. A., & van der Palen, J. (2016). The COPD-SIB: a newly developed disease-specific item bank to measure health-related quality of life in patients with chronic obstructive pulmonary disease. *Health and quality of life outcomes*, *14*(1), 97.

Pool, J. J., Hiralal, S. R., Ostelo, R. W., van der Veer, K., & de Vet, H. C. (2010). Added value of qualitative studies in the development of health related patient reported outcomes such as the pain coping and cognition list in patients with sub-acute neck pain. *Manual Therapy*, *15*(1), 43-47. https://doi.org/10.1016/j.math.2009.05.010

Rogers, C. R. (1961). *One becoming a person: A therapist's view of psychotherapy*. Houghton-Mifflin.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Shin, R. Q., Ezeofor, I., Smith, L. C., Welch, J. C., & Goodrich, K. M. (2016). The development and validation of the Contemporary Critical Consciousness Measure. *Journal of Counseling Psychology*, *63*(2), 210. https://doi.org/10.1037/cou0000137

van der Veer, K., Ommundsen, R., Yakushko, O., Higler, L., & Hagen, K. A. (2013). Psychometrically and qualitatively validating a cross-national cumulative measure of fear-based xenophobia. *Quality & Quantity*, *47*(3), 1429-1444. https://doi.org/10.1007/s11135-011-9599-6

Wiggins, S., Elliott, R., & Cooper, M. (2012). The prevalence and characteristics of relational depth events in psychotherapy. *Psychotherapy Research*, *22*(2), 139-158. https://doi.org/10.1080/10503307.2011.629635

Willis, G. B. (2004). Cognitive interviewing revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 23-43). Wiley. https://doi.org/10.1002/0471654728.ch2

Author Biographies

Gina Di Malta is a counselling psychologist, lecturer in counselling and psychotherapy at the Open University, UK, and a research associate at the Centre for Research in Social and Psychological Transformation (CREST), University of Roehampton, UK. Gina's research is focused on the psychology of connection and the developments of pluralistic psychotherapy.

Mick Cooper is a professor of counselling psychology at the University of Roehampton, where he is the Director of the Centre for Research in Social and psychological Transformation (CREST). He is a chartered psychologist, a UKCP-registered existential psychotherapist, and a Fellow of the British Association for Counselling and Psychotherapy. He is the author and editor of a range of texts on person-centered, existential, and relational approaches to therapy; including Working at Relational Depth in Counselling and Psychotherapy (2nd ed., with Dave Mearns) and Integrating Counselling and Psychotherapy: Directionality, Synergy, and Social Change. He is the father of four children and lives in Brighton on the south coast of England.

Joel Vos is a psychologist, philosopher, researcher and existential therapist in the UK. He is Senior Researcher at the Metanoia Institute, Research Coordinator at the New School of Psychotherapy and Counselling, and Chair of the IMEC International Meaning Conferences. His research focuses on existential psychology and humanistic psychotherapies. His recent books include 'The psychology of COVID-19' (SAGE), 'The economics of Meaning in Life' (University Professors Press), 'Mental health in crisis' (SAGE) and 'Meaning in Life' (Palgrave McMillan).

Kees van der Veer is professor in psychology, currently connected to TPU and Galtung Institute (Grenzach-Wyhlen). He is also director of the Cultural Studies Foundation in Amsterdam, a former research fellow from the International Peace Research Institute Oslo and emeritus from VU University in Amsterdam. His research interests are within the fields of peace research, relationships and conflicts at the micro level, social research methodology, migration, and xenophobia.