MIAMI UNIVERSITY – THE GRADUATE SCHOOL

CERTIFICATE FOR APPROVING THE DISSERTATION

We hereby approve the Dissertation

of

James A. Lani

Candidate for the Degree:

Doctor of Philosophy

____

William B. Stiles, Director

____

Mia W. Biran, Reader

____

Roger M. Knudson, Reader

____

Paul V. Anderson, Graduate School Representative

Abstract

The assimilation model (Stiles et al., 1990) describes how a client's problem changes over the course of psychotherapy. The assimilation model proposes an eight-stage process by which these changes occur. Markers are easily identifiable events in psychotherapy that recur across sessions and across clients and that indicate psychologically important phenomena. This study examined the extent to which markers of assimilation stages can be reliably identified.

A manual of twenty-six markers was developed by four assimilation researchers. The manual included a description and three illustrations of each marker. Fourteen raters were trained to use the manual, practiced identifying markers in excerpts, and then identified markers in three sets of excerpts unrelated to the manual's illustrations and practice excerpts.

A reliability coefficient was calculated for each marker and each rater on each data set. The reliability varied substantially across markers. Convergent validity was supported, as the raters' marker endorsements converged with independent researchers' assimilation ratings of the same excerpts. Construct validity was supported too, as the incidence of markers associated with higher assimilation stages increased with the increasing session number in successful therapy cases (two of the three sets of excerpts).

Based on the level of agreement in the endorsements of two seven-raters groups in two of the three data sets and the high level of agreement between raters' and independent researchers' endorsements, six markers (Desiring Change, Getting Stuck, Feeling Confused, Feeling Vulnerable, Recurring Problem, and Difficulty Articulating What's Wrong) could be considered successful. The results suggested that markers' reliabilities are affected by noise in the data set.

IMPROVING A MARKER-BASED SYSTEM
TO RATE ASSIMILATION OF PROBLEMATIC EXPERIENCES

A DISSERTATION

Submitted to the Faculty of
Miami University in partial
fulfillment of the requirements
for the degree of
Doctor of Philosophy
Department of Psychology

by

James Anthony Lani
Miami University
Oxford, Ohio
2003

Dissertation Director: William B. Stiles

Table Of Contents

CHAPTER 1: INTRODUCTION. 1

A Description Of The Assimilation Model 1

Use Of Assimilation Ratings. 4

Assigning Assimilation Ratings Using A Qualitative Method. 5

Three Concerns With The Assimilation Analysis 6

Assigning APES Ratings Using A Marker-Based Strategy 6

The Advantages Of Markers In Assigning APES Stages. 7

The Identification Of Markers Process. 8

Empirical Results Of The Honos-Webb Study 8

Phase 1: The Development Of The 1998 Manual. 9

Phase 2: How Passages Were Selected 10

Phase 3: Raters Endorse Markers And Assign APES Ratings To Passages 10

Phase 4: Reliability Of The Markers And Validity Of The Marker-Based APES Rating. 11

Questions Stimulated By The Honos-Webb (1999) Study. 12

Rationale Of Study. 13

CHAPTER 2: METHOD. 15

Phase 1: Creating The 2000 Marker Manual. 15

Investigators 15

Sources Of Passages Used In Constructing The Manual 15

Procedure. 17

Phase 2: Selecting Passages To Test The Marker Rating Strategy In Phase 3. 18

Bill Case 18

Detert Cases 19

Reid Case 19

Phase 3: Raters Endorse Markers In New Material 20

Raters 21
Procedure: Training Raters To Endorse Markers 21
Raters Identifying Markers In Passages. 30

CHAPTER 3: RESULTS 22
Construction Of The Marker Manual. 22
Reliability 27
Aggregating Kappa. 27
Pairwise Reliabilities. 28
Kappa 28
Mean Marker Kappa. 28
Rater Mean Kappa. 28
Groupwise Reliabilities 29
Group Marker Kappa 29
Group Rater Kappa. 29
Reliabilities Results. 29
Rater Kappas. 30
Marker Reliability. 30
Group Reliability Results. 32
Group Marker Kappas. 32
Group Rater Kappas. 32
Confusion Ratio. 33
Results Of The Confusion Ratio Calculations 34
Three Clusters Of Confused Marker Pairs 36
Validity. 37
Associating Raters' Marker Endorsements To APES Stages 37
Determining A Representative APES Rating For Each Excerpts Based On Fourteen Raters' APES Ratings 37
Determining A Means And Modal APES Ratings For Detert's APES Rating. 38

Determining A Means And Consensus APES Ratings For Reid's APES Rating. . 38
Convergent Validity In The Detert Data 39
Assessing Convergent Validity By A Crosstab Between Detert's And Raters' APES Ratings 39
Convergent Validity In The Reid Data 39
Assessing Convergent Validity By A Crosstab Between Detert's And Raters' APES Ratings 39
Construct Validity Of Assimilation Markers 40

CHAPTER 4: DISCUSSION 41

Reliability Of Markers 41
Using Signal Detection Theory To Understand The Reliability Of Markers 42
Raters Reliabilities Did Not Differ Within Data Sets 43
Validity Of Markers 44
Convergent Validity Of Markers 44
Construct Validity Of Markers 45
Qualifications To The Validity Of Markers 45
Evaluation Of Successful, Potentially Successful, And Failed Markers 45
Improving Marker Reliability 46
Improving Marker Reliability By Better Marker Descriptions 46
Training Of Raters May Improve Reliability 47
Examine The Reliability When A Few Markers Are Studied At One Time 47
Use Clinically Sophisticated Raters 48
Conclusion 48

References. 49

Table 1: Assimilation Of Problematic Experiences Scale 55
Table 2: Markers Of Stages Of Assimilation. 56

Table 3: Cross-reference: Present Markers to Other Researchers' markers 57

Table 4: Calculation of a Kappa Statistic for a Marker between Rater 1 and Rater 2 . . . 58

Table 5: Name, Acronym, And Description Of Types Of Kappa Calculations 59

Table 6: Rater Mean Kappa for Bill, Detert, and Reid Cases .60

Table 7: Marker Mean Kappa for Bill, Detert, and Reid Cases 61

Table 8: Group Marker Kappa for Bill, Detert, and Reid Cases 62

Table 9: Observed Confusion Values, Expected Confusion values, and Confusion Ratios by Marker 63

Table 10: Confusion Ratios Greater Than 2.0 64

Table 11: Frequencies of Excerpts Where Detert's Mean APES Ratings and Raters' APES ratings Agreed and Disagreed 65

Table 12: Frequencies of Excerpts Where Detert's Mean APES Ratings and Raters' APES ratings Agreed and Disagreed 66

Figure 1: Assimilation Curve. 67

CHAPTER 1: INTRODUCTION

The assimilation model (Stiles, Elliott, Llewelyn, Firth-Cozens, Margison, Shapiro, & Hardy, 1990) attempts to explicate the process of psychotherapy by describing how a client's problem changes over the course of psychotherapy. The assimilation model proposes an eight-stage process by which these changes occur. This study examined the extent to which *markers* of assimilation can be reliably identified and facilitate the rating of the stages of assimilation. Markers of assimilation are easily identifiable events in psychotherapy that recur across sessions and across clients. Researchers have used markers to identify various processes in therapy (Greenberg & Foerster, 1996; Honos-Webb, 1999; Honos-Webb, Lani, & Stiles, 1999; Honos-Webb, Stiles, & Greenberg, 2003; Honos-Webb, Surko, & Stiles, 1998; Rice & Greenberg, 1984).

In this dissertation, I construct a manual of markers, assessed the reliability of these markers, and validated the use of markers in the assignment of assimilation ratings. First, I reexamine the assimilation model (Honos-Webb & Stiles, 1998; Stiles et al., 1990). Second, I summarize a qualitative method to make assimilation ratings (Stiles, Meshot, Anderson, & Sloan, 1992) and characterize several problems associated with this method. Third, I discuss the process of assigning assimilation stages to clients' verbalizations in therapy using a strategy of markers. Fourth, I present the empirical results from the Honos-Webb study (1999) in which markers were used to identify assimilation stages, and I examine several remaining questions raised by the study. Fifth, I ask two research questions that followed from a rationale that I describe. Sixth, I present the results of this study. Finally, I evaluate and interpret the results with respect to the reliability of the markers and the validity of the marker-based strategy to rate assimilation stages.

A Description of the Assimilation Model

Clients enter therapy with a range of presenting problems—anxiety, depression, grief—that are psychologically painful. The assimilation model proposed by Stiles et al. (1990) describes sequential patterns in the way these problems can change over the course of psychotherapy. Stiles, Morrison, et al. (1991) summarized these patterns in the Assimilation of

Problematic Experiences Scale (APES; Table 1). The APES is a provisional description of how the client's problematic experiences change over the course of therapy. The scale was developed from case studies in which particular problems are followed longitudinally across sessions (e.g., Stiles, Morrison, et al., 1991). The APES describes how people progress in successful therapy. The progression along the APES is not necessarily a smooth, linear sequence. Psychotherapy is considered successful when clients move from the early stages of the APES (Stages 0, 1 or 2) to the later stages (Stages 5, 6, or 7). Reliable and valid APES ratings can be a useful tool to track and assess a client's progress in therapy.

In the assimilation model, internal voices can be formed through different experiences. For example, experiences of being nurtured can give rise to internal voices that one is worthwhile and that the world is a benevolent place. The assimilation model describes clients' internal experience as comprised of a collection of such voices; these voices are referred to in the model as the community of voices. These voices have the ability to express their own thoughts, feelings, and motivations. Theoretically, people without problematic experiences move from one internal voice to another fluidly and painlessly.

However, not all voices are part of the client's community of voices. A client who experiences a trauma may have the traumatic problematic voice instilled in him or her. This painful voice may lead them to think, feel, or act in certain ways, such as avoiding places that would remind her of the traumatic experience.

The assimilation of voices model suggests that the resolution of a client's problem occurs through a dialogue between two active voices: one from the community of voices and the problematic voice. Theoretically, one of the voices is a dominant voice from the community of voices (Honos-Webb & Stiles, 1998) and could be characterized as the top-dog voice. The other voice is an unwanted, problematic voice, and could be considered an underdog voice.

This *intra*personal process of assimilation of voices parallels the *inter*personal process in which two people, who initially oppose each other, begin to communicate, and ultimately, through shared understanding, reconcile their differences (Stiles, 1998). The interaction between the top-dog and underdog is hypothesized to follow a predictable sequence along the APES continuum (Table 1). As the voices change (described below) when moving along the APES continuum, the client's affect changes as well (Figure 1 and Table 1).

In Stage 0 of the 8-stage assimilation model, the Warded Off stage, the underdog voice is not heard, and the top-dog voice may be indistinguishable from the rest of the community of voices. The client's affect at this stage is often neutral (Figure 1). The underdog voice may, however, be represented in the form of physical symptoms. For example, as one therapist talked about a client's feelings of dependency and weakness, the client responded, "I still feel the lump in my throat, and at times it's worse than others..." (Honos-Webb, Surko, Stiles, & Greenberg, 1998).

At Stage 1, the Unwanted Thoughts stage, the underdog voice begins to emerge, as evidenced by speech in the client's narratives in therapy. For example, in the case of John Jones (Stiles, Meshot, Anderson & Sloan, 1992), a client's unwanted thoughts of his homosexuality presented itself as the client stated, "I would be very upset if I discovered this homosexuality were true of me." Meanwhile, the community of voice's top-dog used defenses to avoid the unwanted, underdog voice. Several moments later the client stated, "Maybe I'm defending some of the feelings I have of myself." Clients at this stage prefer not to think of their problem and typically experience minimal negative affect (see figure 1).

In Stage 2, the Vague Awareness/Emergence stage, the underdog voice clearly emerges in the client and is painfully acknowledged by the community of voices. For example, in the case of Jan (Honos-Webb, Surko, Stiles, & Greenberg, 1999) the top-dog struggled to stay in control whereas the underdog felt weak and dependent. In one excerpt, Jan painfully reported a vague awareness of a problem by stating, "You see right now, I don't know why I'm crying. I can't, you know, put my finger on it like. What is it about this that, I'm talking about something that's really not that difficult? Why am I crying?"

At Stage 3, the Problem Statement/Clarification stage, can be characterized by a clear differentiation of the top-dog and underdog voices; the voices can state their individual positions, attitudes, and feelings. The voices are equally salient, and the conflict between them is explicit (Honos-Webb & Stiles, 1998). For example, Lisa struggled with forgiving her mother by stating, "…the voice is saying, 'no, she needs to be punished'…and the other [voice] is like, 'you know, forgive, she's your mother and she's human….'" (Honos-Webb, Stiles, Greenberg, & Goldman, 1998).

In Stage 4, the Understanding/Insight stage, the voices communicate and come to understand each other. A client in this stage has the perspective to speak about the voices without speaking from the voices (Honos-Webb, Surko, et al., 1998). Typically, the client's affect is mixed. In this example, Lisa connected previously unrelated voices and gained perspective on her problems, stating, "now I put it together and it makes sense why I was stuck…my husband being so domineering and [he] wanted to be in control, I just carried it into the marriage…" (Honos-Webb, Stiles, Greenberg, & Goldman, 1998).

Stages 5, 6, and 7, are named the Working Through, Problem Solution, and Mastery stages, respectively. In the Working Through stage, the client attempts to apply learning gained from therapy to her outside life. These attempts are typically met with positive affect (see Figure 1). In the Problem Solution stage, the application of her solution is successful. For example, Jan (Honos-Webb, Surko, et al., 1999), exemplified the Problem Solution stage with a sense of pride, saying, "I'm really proud of myself the way I'm dealing with things with my moth—especially my mother." In the Mastery stage, the client successfully generalizes and applies the voices' joint solution to the outside world. Later in the above session, Jan displayed mastery by stating, "…it's okay, I don't have to be superwoman, and it's okay to ask for help, that if I can't manage or can't do it, people are not going to think any less of me because of that." Ultimately, as the voices move into stage 5 and beyond, the dialog between the voices leads the underdog voice to assimilate even further into the community of voices, while the community accommodates the underdog voice.

The Use of Assimilation Ratings

APES ratings can be important in psychotherapy research and practice in several ways. First, psychotherapy researchers could use APES ratings as an outcome measure, to examine the effectiveness of different treatments in helping clients to assimilate problems. For example, a research project could look at clients whose problems began with the same level of assimilation and see how the problems progress on the APES with different treatments.

Second, therapists who could accurately assign APES to a client's problems could be responsive to the client's stage of assimilation by providing interventions that help the client move to the next stage of assimilation (Stiles, Shapiro, Harper, & Morrison, 1995). For example, a client's problematic experience rated at stage 2, Unwanted Thoughts, may call for therapists interventions aimed at containing affect, whereas clients at stage 5, Working Though, may call for interventions aimed at problem-solving.

Third, assigning APES early in treatment could be made clinically useful by matching clients to therapies that would most benefit them. Two studies (Stiles, Shankland, Wright & Field, 1997; Stiles, Barkham, Shapiro, & Firth-Cozens, 1992) support this notion. Stiles, Barkham, et al., (1992) reported that progress on a particular problem was steadier in a two-treatment sequence when exploratory psychotherapy (EP), which tends to focus on earlier stages of problematic experiences, preceded cognitive-behavior (CB) therapy, which tends to focus on the latter stages of problematic experiences, than when the therapies were reversed. In another study, Stiles et al. (1997) found that clients with well-formulated problems in stages 2.5 (approaching Problem Statement stage) or higher of the APES had better outcomes in Cognitive-Behavioral therapy compared with exploratory therapies. These studies suggest that some treatments may be more effective, depending upon a problem's stage of assimilation.

The preceding discussion suggests that reliable and valid APES ratings in a client's therapy could be important both clinically and scientifically. In the next section I review how portions of clients' therapies have been rated on the APES.

Assigning Assimilation Ratings Using A Qualitative Method

Assimilation analysis is a qualitative method developed to track the progression of problematic experiences (Honos-Webb, 1999; Stiles, Meshot, Anderson, & Sloan, 1992; Stiles, Morrison, Haw, Harper, Shapiro, & Firth-Cozens, 1991). In an assimilation analysis, the process of assigning of one of the eight APES ratings to a client's narrative has followed a four-step process (Honos-Webb, 1999; Stiles, Meshot, et al., 1992).

1. <u>Taking notes to describe different topics in a therapy.</u> The researcher read and re-read the entire transcript, taking notes of the various topics discussed by the client and where in the session they occurred.
2. <u>Selecting a theme to analyze.</u> The purpose of this step was to select a theme for the assimilation analysis. The researcher identified a topic where the client reached new understanding or insight about her problem.
3. <u>Excerpting passages with a similar theme.</u> All passages in the therapy relating to the identified theme were excerpted.
4. <u>Assigning APES to excerpts.</u> Researchers applied the APES to the selected passages. This was done by casting the excerpted passages into top-dog and underdog voices or by

having an understanding of the case and assessing the degree of assimilation. Based the researcher's knowledge of that therapy and use of the assimilation of voices, the researcher assigned an APES rating to the excerpts.

Three Concerns with Qualitative Assimilation Analysis

The assimilation ratings made using the above qualitative method have presented three practical problems for assimilation researchers and clinicians: (1) the therapy had to be complete before the assimilation analysis was conducted, (2) the analysis required a very labor-intensive procedure, and (3) often the analysis did not result in adequate reliability of APES ratings.

First, qualitative assimilation analysis requires that a client's therapy be completed prior to the assignment of APES ratings, thus making the analyses less clinically useful to that therapy. If the assignment of assimilation ratings to clients' problems could occur early in their therapy, they would be more clinically useful.

The second concern with this procedure is that cataloging all sessions of a client's therapy takes an enormous amount of time and effort, yet this process has to be complete before the APES ratings can be assigned. Researchers and clinicians who are interested in assessing a client's stage of assimilation may not have the resources needed for this procedure.

The third concern with the method, and perhaps the most important, is the uncertain reliability of the qualitative APES ratings. Assimilation researchers often struggled to come to a consensus on the APES rating of a particular excerpt. Differences can occur because of the various interpretations of the excerpts to be rated and differing familiarity with the cases.

Due to the concerns with the method just described to assign assimilation ratings, a strategy using markers to make APES ratings was pursued. In the following section I describe a marker-based strategy to assign APES ratings.

Assigning APES Ratings Using a Marker-Based Strategy

Markers are easily identifiable signs of a psychological event or a client's psychological state (Honos-Webb, Lani, & Stiles, 1999). Greenberg and Rice (1984) used the concept of markers as part of a strategy to describe and understand the mechanisms of change in psychotherapy. Their notion of markers focused on recurrent episodes in a client's therapy. They stated, "There are episodes or events in therapy that are similar to each other in some important ways… [and] do have some clearly identifiable structural similarities. Markers recur sufficiently often within and across clients to permit a systematic focus on their commonality" (page 19). Thus, markers have a structure or formal characteristic that is independent of that particular event; a marker signifies that the structure is present. To give you a sense by what I mean by a marker, I present an example of the bewilderment marker, a pattern that was observed in several stage 2 excerpts. In this excerpt, the client doesn't understand why she hurts.

> Client: Whether it's giving up, I've given up on him, or just, just let it be I can't change him. What's the point, but then, then inside, it— it still hurts me, um, which doesn't make sense.

To illustrate that markers are formal characteristics that transcends particular content, I present another example of the bewilderment marker. In this excerpt, the client is confused between parental shoulds and her own wants.

> Client: What is the boundary line between what I really should do and what I think my parents would like me to do? And I don't know, it's been a bit confusing too because it's been a bit confusing… (Session 7)

In following section, I briefly explain the usefulness of markers in assigning assimilation ratings and discuss the advantages of markers. I then, describe the process of identifying markers.

The Advantages of Markers In Assigning APES Stages

There are several advantages to using markers to indicate the stage of a client's problem along the APES continuum. Markers could be advantageous in clinical research. Theoretically, a researcher does not need to know the context or theme of a therapy to recognize markers in a client's utterance. Further, markers can be found at any point in a therapy, without the requirement that a therapy be completed. Reliable markers could permit clinicians or researchers to make a quick and accurate assessment of a client's APES stage for a particular problem.

Marker research could be advantageous to practicing therapists. Markers could be used by therapists to efficiently assign APES ratings, they could respond to a client's needs with the most appropriate intervention to facilitate the assimilation of problems. To select an intervention, a therapist could use markers to reliably and accurately judge where a client's problem is located on the APES continuum.

The Identification of Markers

The identification of markers (Honos-Webb, 1999) entails three steps. The first step focuses on recurrent patterns in a client's dialogue. The second step requires researchers to go back and forth between their clinical intuition and their description of patterns and the observations of the pattern in client's dialogue. This process produces improved descriptions of the markers by repeatedly evaluating the descriptions against actual cases. The third step assesses the reliability and validity of the markers.

First, the discovery of markers begins by intensively studying and identifying recurrent patterns or events within a particular stage of assimilation; an extensive search for patterns in excerpts was undertaken at each APES stage. A researcher then uses his or her intuition to select an important pattern or event in the client-therapist dialogue.

Second, after the patterns are observed, the researcher attempts to describe and name the pattern. The cycling back and forth between observation and description, first described by Lewin (1951), refines the description of the pattern.

Third, to complete the marker identification process, a researcher assesses the reliability of a marker and its validity with respect to an APES rating. The markers' reliability is supported when trained raters can consistently identify a marker in excerpts purported to contain a marker. The validity of the marker is supported when the marker is empirically detected in excerpts rated by other means as representing the appropriate APES stage.

Empirical Results of the Honos-Webb Study

Honos-Webb (1999) conducted a study of the development and testing of a marker-based strategy used to assign APES ratings, which I review in detail because my study replicates and extends it. A report of some of these results has recently been published (Honos-Webb et at, 2003). The study, described below, proceeded in four phases: the development of the 1998 Manual (Honos-Webb, Surko, et al., 1998), the selection of excerpts, and raters' endorsements of markers and the assignment of APES ratings to selected excerpts, and the assessment of the reliability and validity of markers and APES ratings.

Phase 1: The Development of the 1998 Manual

The construction of the 1998 Manual began with a search for markers. Honos-Webb, Surko, et al. (1998) collected excerpts from previous studies where the excerpts had been rated on the APES (Honos-Webb, Stiles, Greenberg, & Goldman, 1998; Honos-Webb, Surko, Stiles, & Greenberg, 1998; Varvin & Stiles, 1999). The researchers then used theoretical descriptions and their clinical intuition to identify markers within the excerpts. The process of identifying markers was conducted stage by stage. For a marker to be included in the 1998 Manual, it had to be present in three excerpts rated in one APES stage and not present in excerpts rated at any other APES stage. The five sections of the 1998 Manual were:

1. The assimilation model;
2. Guidelines to assign assimilation ratings using markers;
3. An assimilation rating form;
4. A description of the twenty-five markers with illustrations;
5. A set of heuristics for prioritizing markers.

I will now describe these sections of the manual.

Assimilation model. The 1998 Manual presented a synopsis of the assimilation model, the theory of voices, and a description of the eight APES stages. The purpose of her study was to reliably assign APES ratings using valid markers, thus requiring solid grounding in assimilation theory.

Guidelines to Assigning APES Ratings on the Assimilation Rating Form. The manual included a procedure for making APES ratings. The raters were instructed to first spend some time reading excerpts from clients' therapies. Next, the raters read through the list of markers, endorsing markers, and indicating the level of confidence in their endorsement on the assimilation rating form. The assimilation rating form listed each of the markers and the stage in which it was found. Beside each marker was a "Yes," and "No," to indicate the presence or absence of a marker, and a blank line to indicate if the raters had "low confidence" in his or her endorsement. The raters were to then assign an APES rating to the excerpt and indicate the confidence in their rating.

Description of markers. The manual consisted of 25 markers (see Table 2). Each marker had a description of its formal characteristics, instructions of when to endorse the marker, and clinical illustrations.

Heuristics for Prioritizing markers. The heuristics portion of the 1998 Manual instructed raters on how to select and prioritize markers. The heuristics helped raters distinguish between various stages of assimilation, especially when two stages had similar characteristics. The heuristics were necessary due to differing APES stages sharing similar qualities. For example, clients at both Stage 2 and Stage 4 can show negative affect, or, in Stage 0 and Stage 7, clients' narratives can have an external focus. As examples of an external focus at two different stages of assimilation, the following two illustrations are presented. In this Stage 0 excerpt, Lisa (Honos-Webb, Surko, Stiles & Greenberg, 1998) externalized her problem by expressing her worry about others' feelings, stating, "like the moment somebody else wants something from me, family or friends…I just don't want to disappoint anybody" (Session 1, page 3). Similarly, in this Stage 7 excerpt, Jan (Honos-Webb, Surko, Stiles & Greenberg, 1998) framed her problems in external terms by stating, "…it's okay to ask for help, that I can't manage it or can't do it um, people are not going to think less of me " (Session 16, page 12). In both examples, the criterion of an external focus created uncertainty in the assignment of the APES ratings.

Phase 2: How Excerpts Were Selected

Honos-Webb (1999) selected excerpts to be used in the third phase of her study. To select excerpts, Honos-Webb selected a client's therapy and conducted an assimilation analysis. The therapy was the case of Sarah, a woman seen in Process-Experiential psychotherapy for 18 sessions (Greenberg & Watson, 1998). Sarah was deemed a successful therapy case by objective measures. Honos-Webb conducted an assimilation analysis on the case using the procedure similar to the John Jones analysis described earlier. Her analysis resulted in 45 excerpted passages for use in phase 3.

Phase 3: Raters Endorse Markers and Assign APES Ratings to Excerpts

In this phase of her study, Honos-Webb trained raters to endorse markers and assign APES ratings. To conduct this phase, she used the 45 excerpts from the case of Sarah, the APES rating scale, and the 1998 Manual.

Raters and Training. Honos-Webb's (1999) study used two groups of raters: a high and a low sophistication group. All of her raters read her manual and a paper on the assimilation of voices (Honos-Webb & Stiles, 1998). The high sophistication raters were graduate students who were participants in the Assimilation Research Group, and were already familiar with the assimilation model and qualitative research. The low sophistication raters were undergraduate students, inexperienced in both the assimilation model and qualitative research. They received six one-hour training sessions where they reviewed the markers in the 1998 manual and practiced rating excerpts.

Rating Procedure. Raters read each of the 45 excerpts out of context and out of order. The excerpts were drawn from one of two themes: the caretaker theme or the barriers theme, and each theme included a description of the top-dog and underdog voice for that particular theme. All of the raters were instructed to use the manual, follow the instructions and heuristics therein, and endorse markers and assign APES rating for each excerpt.

Phase 4: Reliability of the Marker-Based APES Ratings and Markers

In this phase of her study, Honos-Webb assessed the reliability of the APES assignments and markers.

APES Rating Reliability Results. Overall, interrater reliability of the APES ratings was acceptable, supporting the Honos-Webb marker-driven strategy. The reliability was indexed by the intraclass correlation coefficient designated ICC (1, k) by Shrout and Fleiss (1979). The ICC of the pooled raters' (i.e., the low and high sophistication groups together) APES ratings was high, ICC (1, 8) = .93. The APES rating agreements were also good for the low, ICC (1,4) = .86, and high, ICC (1,4) = .90, sophistication groups of raters. However, Honos-Webb (1999) reported that the ICC's for individual raters were acceptable only for the high sophistication individuals, (ICC 1,1) = .70, but not for the low sophistication individuals, ICC (1, 1) = .61.

Pairwise Marker Reliability Results. In her study, the reliabilities of the pairwise markers considered separately were quite poor. Honos-Webb (1999) used the Kappa statistic (Barker, Pistrang, & Elliott, 1994; Fleiss, 1973) to assess the proportion of agreement in raters' endorsements of particular markers. Agreement among raters' assignment of the twenty-five markers did not result in *any* marker reaching almost perfect agreement (i.e., kappa > .75), whereas only five of the twenty-five markers reached moderate to substantial agreement (Kappa between .40 and .74).

Questions Stimulated By and Limitations of the Honos-Webb (1999) Study

Several questions were stimulated by the findings of the Honos-Webb (1999) study. In this section I will discuss problematic findings, construct hypotheses as to what may have caused the problem, and review several limitations of her study.

One puzzling finding from the Honos-Webb (1999) study was that raters' marker reliabilities were much lower than the reliabilities of their APES ratings. For example, as stated above, the interrater reliability of the APES was estimated from .61 (for a single low-sophistication rater) to .93 (for 8 raters pooled), whereas the range of the markers' pair-wise reliability was .02 to .58, with a median reliability of .13. These findings are peculiar because raters were supposed to use the markers to assign the APES rating to the excerpts. One of several possible explanations for these results is that the endorsement of markers may have been driven by raters' knowledge of the assimilation model and stages, rather than the other way around. In her study, raters read the manual, which included information on both the assimilation model and markers. Importantly, markers in the rating form were matched and labeled by APES ratings (e.g., marker 1a indicated that the marker was in APES stage 1, marker 2a indicated that the marker was in APES Stage 2, etc.). Raters may have implicitly assessed the excerpt's APES stage, then searched backwards for a marker within that APES stage. Other explanations include raters being more facile with different markers, or perhaps several markers or voices exist in a single excerpt so that finding one marker stopped the search. Any of these explanations could account for raters' ability to reliably assign APES ratings, but not to reliably endorse markers.

A couple of limitations in the Honos-Webb study hindered the usefulness and generalizability of the marker strategy. First, in the Honos-Webb procedure, the raters knew the excerpts' themes when identifying markers and assigning APES ratings. This procedure obviated one of the potential advantages of markers, which is the ability to go into a therapy transcript and rate a problem's APES stage with little or no knowledge of the client's history.

I also wondered whether the markers from her study were applicable to other therapies from different theoretical orientations. The Honos-Webb (1999) study drew primarily on a Process-Experiential psychotherapy (PEP) case material, as described above. I wondered if these markers could be generalized to therapies outside of PEP, such as client-centered, cognitive-behavioral, or psychodynamic case material.

Given the above questions and limitations regarding the Honos-Webb study, I now present my study that address these issues.

Rationale of Study

I examined two primary questions in this study:

1) Can markers of assimilation be reliably identified in excerpted passages of psychotherapy transcripts?
2) Are the identified markers valid indicators of APES stages?

Although the procedure to address these questions is described in detail in the Method section, I will briefly sketch out the three-phase plan and the analyses. In Phase 1, four assimilation researchers identified markers in therapy excerpts and constructed a manual of markers of assimilation. In Phase 2, I collected three sets of therapy excerpts--excerpts not used in phase 1--for use in Phase 3. For two of three sets, independent researchers had assigned assimilation ratings to excerpts as part of their research. In Phase 3, fourteen trained undergraduate raters used the manual to identify markers in the Phase 2 excerpts. The reliability with which the markers were identified was assessed. The agreement between raters' marker endorsements and researchers' APES ratings (i.e., convergent validity) was assessed. The relationship between raters' marker endorsements and session number of the excerpt (i.e., construct validity) was evaluated. This relationship between endorsements and session number could provide evidence for the construct validity of the markers in as much as assimilation stages (and the markers associated with these stages) are expected to increase across psychotherapy sessions in successful therapy cases.

The current study improved upon the Honos-Webb (1999) study in several ways. First, in the Honos-Webb (1999) study the reliabilities of the markers were confounded by raters' knowledge of the assimilation model. When raters endorsed markers in her study, they may have drawn on their evaluation of the assimilation stage of the client's problem, and chose a marker within that stage rather than choosing a marker without such assimilation knowledge. In the present study, the raters were not given information on assimilation theory.

Second, raters in the Honos-Webb (1999) study had knowledge of the themes of the client's problem. Knowing how the client progressed with regard to a particular theme may have informed their assessment of APES ratings and their marker endorsement. The current study addressed this limitation by providing excerpts to raters without knowledge of a theme. The marker-based system in the present study was described to raters as a matching task, where markers were to be endorsed if the characteristics of the excerpt met the description of the marker(s).

Third, the convergent validity in the Honos-Webb (1999) study was assessed by, and perhaps compromised by, comparing her raters' APES ratings with her APES ratings. She may have assigned lower or higher APES ratings to passages because she had a greater investment than her raters in assigning APES stages that corresponded to the markers in a passage. That is, she may have had a different interest than her raters in letting the markers in the passage guide her APES ratings. The current study had the advantage of comparing APES ratings assigned by independent researchers who had no such investment in APES ratings with raters' ratings to assess validity.

CHAPTER 2: METHOD

The method used to investigate the two research questions involved three phases: developing a manual of markers, selecting excerpts, and testing whether independent raters can reliably identify markers in new material.

Phase 1: Developing the Marker Manual

The purpose of this phase was to construct a manual of markers.

Investigators

The primary investigator, two other psychology graduate students, and an undergraduate psychology student participated in this phase of the research. I was the primary investigator, a 40-year-old Caucasian male, interested in psychotherapy process research, and using this research as my doctoral dissertation. The graduate students were female second and third year clinical doctoral students who were active assimilation researchers. The fourth researcher was a male, senior undergraduate who participated for course credit while gaining qualitative research experience.

Sources of Excerpts Used in Constructing the Manual

I selected the excerpts used to develop the markers from two types of sources. First, I compiled 357 excerpts from the following completed nine assimilation analyses to find potential markers. These cases represented a diversity of clinical problems and theoretical orientations. The names were the pseudonyms used in the previous studies.

1. Lisa was a depressed woman who felt helplessness in her struggles with her husband's gambling and was treated with 15 sessions of Process-Experiential psychotherapy (Honos-Webb, Surko, Stiles & Greenberg, 1998).

2. Fatima was a refugee who had been arrested and tortured, and experienced the death of her infant. She was treated with 65 sessions of Psychoanalytic therapy (Varvin & Stiles, 1998).

3. John Jones was a man who was anxious about his homosexual impulses, and unclear about his passive and aggressive behaviors. He was treated with 20 sessions of Psychodynamic psychotherapy (Stiles, Meshot et al., 1992).

4. Jan was a depressed woman who felt needy of others, and was treated with 16 sessions of Process-Experiential therapy (Honos-Webb, Surko, Stiles, & Greenberg, 1999).

5. Vicky was a woman struggling with her sexuality, relationship with her parents, and uncertainty with her career choices. She was treated with 18 sessions of Psychoanalytic therapy (Knobloch, Endres, Stiles, & Silberschatz, 2001).

6. Millie was a woman who was struggling to find employment, to develop her social life, and to improve her self-concept (Lani, Stiles, Shaikh, & Silberschatz, 1998). She was treated with 16 sessions of Psychoanalytic psychotherapy.

7. Margaret was a depressed woman who struggled with her lifelong role as a responsible caretaker. She was treated with 17 sessions of Client-Centered therapy (Glick, Stiles, & Greenberg, 2000).

8. Cybil was a woman who dealt with her perfectionism while coping with job stress. She was treated with 16 sessions of Cognitive-Behavioral therapy (Osatuke, Stiles, Shapiro, & Barkham, 2000).

9. Sarah was a 35-year-old woman who was seen for 18 sessions of Process-Experiential psychotherapy for depression (Honos-Webb, 1998).

The second type of source that I used to develop markers was derived from excerpts in the Process-Experiential cases in *Facilitating Emotional Change: The Moment-by-Moment Process* (Greenberg, Rice, & Elliott, 1993). Several of these examples were created from their experiences as long-time psychotherapy researchers (R. Elliott, personal communication, April 6, 2002).

Procedure

First, the primary investigator compiled excerpts from the above-listed completed assimilation analyses and the *Facilitating Emotional Change* unrated excerpts (Greenberg, et al., 1993). A group of assimilation researchers discussed the excerpts from *Facilitating Emotional Change* and came to consensus on the APES ratings.

Second, four participating investigators studied the 1998 manual (Honos-Webb et al., 1998) to familiarize themselves with the markers and the methods of identifying them.

Third, the investigators identified markers by reading the excerpted passages from the completed analyses, the 1998 manual, and *Facilitating Emotional Change* (Greenberg, et al., 1993). To identify markers, the researchers examined patterns of dialogue—potential markers—which represented assimilated or unassimilated voices and metaphors in clients' stories. The investigators then wrote a description of the potential marker and assigned the marker a name. For each excerpt associated with a potential marker, the previously rated APES of the excerpt was confirmed by consensus among the investigators.

Fourth, the investigators reviewed all of the additional excerpts in that APES stage to seek further instances of that potential marker. The investigators cycled between their description of the marker and the observations of the marker in those excerpts. The descriptions of the markers were refined with each iteration.

Fifth, for a marker to be included in the marker manual, the investigators, by consensus, had to have identified at least three instances of that marker in excerpts of the same stage of assimilation, and absent in all excerpts rated at different stages.

The culmination of the five steps resulted in a marker manual of 26 markers, which gave a description and illustrations for each marker. Each marker was intended to identify one stage of one problematic voice. The manual was expected to have markers representing many of the APES stages. However, we did not expect many Stage 0, 6, or 7 markers because these clients would either not present themselves for therapy or would have successfully left therapy.

Phase 2: Selecting Excerpts To Test The Marker-Based Rating Strategy In Phase 3

The purpose of Phase 2 was to select excerpts to be used by trained raters in to endorse markers in Phase 3. I was fortunate to obtain three data sets: the Bill case (Greenberg & Watson, 1998), Detert cases (Detert, 2000; Detert, Llewelyn, Hardy, Barkham, & Stiles, 2002), and Reid case (2001). The latter two cases were selected for practical purposes—they had APES ratings assigned to them. However, these two data sets were atypical therapy cases: Detert's cases had just two sessions per client, and Reid's client was part of a research program that examined the impact of psychotherapy on patients with functional abdominal pain. It is important to note that the excerpts from the Bill, Detert, and Reid cases were not used in the development of the Manual. That is, I did not want the construction and testing of the manual to be confounded.

Bill Case

The transcribed case of Bill was drawn from the York Depression Project (Greenberg & Watson, 1998). The client, Bill (a pseudonym), was a 29-year-old married man who was diagnosed with Major Depression. Prior to therapy, Bill was administered the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the SCL-90 Global Symptom Index (Derogatis, 1983), the Rosenberg Self-Esteem Scale, and the Inventory of Interpersonal Problems (IIP; Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988). He was then seen for 18 sessions of Client-Centered psychotherapy. Bill was administered the same measures post-therapy. His pre- and post-therapy scores were as follows: Beck Depression Inventory (pre-therapy = 31, post-therapy = 5); SCL-90 Global Symptom Index Score (pre-therapy = 1.81, post-therapy = 0.57); Rosenberg Self-Esteem Scale (pre-therapy = 14, post-therapy = 21); Inventory of Interpersonal Problems (pre-therapy = 2.08, post-therapy = 1.87). Bill's decrease in depression, symptoms, and intrapersonal problems, and increase in self-esteem, led researchers to consider him a successful case.

In the case of Bill, I read through the 400-page transcript several times looking for markers in the excerpts. I excerpted passages that contained markers. These excerpts were selected with a minimum of context and each excerpt focused on only one marker. This phase culminated in the collection of 59 excerpts from the Bill case (see Appendix C). The mean number of lines per excerpts was 7.37 (SD=6.88), with a median of 5 lines.

Detert Cases

The second set of excerpts was drawn from eight successful and unsuccessful Cognitive-Behavioral and Psychodynamic-Interpersonal therapies from the Assimilation in 2+1 Brief Therapy study (subsequently referred to as the *Detert Data*; Detert et al., 2002; Barkham, Shapiro, Hardy & Rees, 1999). The patients were all white-collar workers who were seen for depression. The clients had pre-therapy Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) scores of 16-25. The patients were then treated with either three Cognitive-Behavioral or three Psychodynamic-Interpersonal therapy sessions: two one-hour sessions one week apart, and a third session three months later. The successful clients had post-therapy BDI scores of 2 or less, where the unsuccessful clients' scores had post-therapy BDI scores that did not change very much.

Eighty excerpts were selected by Detert from the first two sessions as part of his research through a five stage procedure: (1) he read the patients' transcripts, listened to their tapes, and made notes; (2) he identified themes by using a Personal Questionnaire, Therapy Session Topic review, and notes from the first reading; (3) Detert reread transcript, highlighting excerpts related to the themes; (4) he and his supervisor agreed on a formal set of criteria for selecting excerpts; (5) he selected ten excerpts from the first and second sessions of the eight clients. Detert's 80 excerpts had a mean number of lines per excerpts of 20.20 (SD=8.73), with a median of 19.50 and mode of 14 lines (7.3% of excerpts).

Prior to assigning APES ratings to the 80 excerpts, Detert and his four raters read an assimilation case study (Stiles, et .al, 1991), the voices formulation paper (Honos-Webb & Stiles, 1998), and the 1998 marker manual (Honos-Webb, et .al, 1998). He and his raters then practiced assigning APES ratings in trial excerpts and had 3-two hour meetings to compare, discuss, and arriving at a consensus on their ratings. They then assigned APES ratings to the 80 excerpts. Detert recorded the five individuals' APES ratings and the mean APES ratings for each excerpt.

Reid Case

A third set of excerpts came from the case of Megan (pseudonym), who was one of Reid's patients who participated in a study examining the impact of psychotherapy on patients with functional abdominal pain. The study was named the Functional Abdominal Pain (FAP) study and took place at Tayside University Hospital (Reid, 2001). Megan was a fifty-year-old Scottish woman who was married. She was referred by her gastroenterologist and surgeon, following a third series of GI investigations in 16 years for epigastric and abdominal pain. The therapeutic goals were to improve her relationship with herself, work through her grief and rage, and develop a greater alliance with her body.

Megan was assessed on the Taylor Manifest Anxiety, BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), BAI, IPQ, and two pain questionnaires. She had been seen for 31 weekly sessions of Psychodynamic-Interpersonal therapy. Megan showed clinically significant improvement on these measures, except for the BDI measure (Megan was not depressed pre- or post-therapy), and was considered a successful case.

Reid excerpted passages using the assimilation analysis. From the analysis, two themes were selected, and 106 excerpts representative of the themes were excerpted.

Reid had knowledge of the entire case when making her APES ratings, while her colleague had only the excerpts. To assign APES ratings, Reid and her colleague examined the voices within the themes and used the APES scale. Three sets of APES ratings were obtained: Reid's, her colleagues, and a consensus rating. The consensus rating was obtained by discussing their reasons for their individual ratings. The mean number of lines per excerpts was 14.88 (SD=8.51), with a median of 14.

Phase 3: Raters Endorse Markers In New Material

The purpose of Phase 3 was to test whether trained raters could reliably identify markers in psychotherapy excerpts. In Phase 3, raters were trained to use the manual, practiced identifying markers in excerpts, and then identified markers in the Phase 2 excerpts using the manual.

Raters

Fourteen undergraduate students, two male (14.3%) and 12 female (85.7%), were recruited from a Midwestern university to participate in the study. Their average age was 19.93 years (SD=.62), and most (71.4%) were recruited through flyers posted in the Psychology Department. Potential raters were screened by me to assess their interest and knowledge in psychotherapy process research and for their overall motivation to participate in the study. They received 1 unit of academic credit for their participation. All of the 14 raters completed ratings on all three sets of excerpts. The average time to complete the Bill data set was 3.13 hours (SD=.53); Detert data set: 4.68 hours (SD=1.03); Reid data set: 4.07 hours (SD=1.07).

Procedure: Training Raters to Endorse Markers

All 14 raters received approximately six hour (M=6.57 hours, SD=.70) of training by me. The first session provided information to the raters about the purpose of this research, how to identify markers in excerpts, and answered any questions that they may have arisen from the presented information.

The raters read the 82-page manual, and discussed the descriptions and illustrations. To train raters to endorse markers on their own, a 27-page practice excerpt set from the case of Millie (Lani, Stiles, Shaikh, & Silberschatz, 1998) was used. Raters practiced identifying markers in these excerpts using the marker list, illustrations, and the instructions in the manual. The raters used a marker rating form to document the markers that they had identified. Towards the latter part of the training, I also explained the correct marker for a particular excerpt and the reasons for these judgments.

For each excerpt, the raters were instructed to read the excerpts carefully, follow the directions in the front of the manual, and use the marker rating form to record the markers that they have identified (Appendix A).

Raters Identify Markers in Excerpts. The next step involved presenting raters with the excerpts selected in Phase 2, asking them to identify markers using the marker Manual, and documenting the identified markers on the marker rating form. The raters were given three sets of excerpts. The first set contained 59 excerpts from the Bill study. The presentation of these excerpts was not in the sequential sequence of the therapy (i.e., the excerpts were randomized). The raters were then given 82[1] excerpts from Detert's 2+1 study. The last data set was the 106 excerpts from the Reid's FAP study, presented in a random order. The raters' tasks were to identify markers in excerpts using the manual, and record the markers on the marker rater form (Appendix A).

CHAPTER 3: RESULTS

This chapter describes the markers comprising the marker manual, along with data on their reliabilities and their validities. Further descriptions and examples are provided in the marker manual.

The Construction of the Marker Manual

Below are brief descriptions of each marker and its relationship to the assimilation of voices theory.

Marker 1: Body Symptoms (APES = 0). In excerpts exhibiting the Body Symptoms marker, clients spontaneously express a somatic complaint, physical symptom, or presence of a negative physiological process. One aspect of this marker was identified by Honos-Webb, Surko, et al. (1998) as the Somatic Symptoms marker (Table 3). Theoretically, the problematic voice expresses itself through the somatic complaint.

Marker 2: Downplaying Negativity (APES = 1). In excerpts exhibiting the Downplaying Negativity marker, clients state a negative aspect of themselves, then immediately downplay or deny that aspect. Theoretically, the client’s dominant community suppresses the problematic voice.

Marker 3: Avoiding Responsibility (APES = 1). In excerpts exhibiting the Avoiding Responsibility marker, clients focus on events or decisions outside of their control or do not take personal responsibility for events within their control. Theoretically, the dominant voice in stage 1 wants to avoid the problematic voice. This particular marker identifies instances when the dominant voice distances itself from the problematic voice by not taking responsibility for problematic voice’s thoughts, feelings, or actions.

Marker 4: Distancing Language marker (APES = 1). In excerpts exhibiting the Distancing Language marker, clients use second and third person pronouns, instead of first person pronouns, when discussing their own problems. Theoretically, the client does not have the insight to understand that the distancing language being used is partitioning the problematic voice from the dominant voice of the community.

Marker 5: Feeling Surprised At Own Reaction (APES = 1). In excerpts exhibiting the Feeling Surprised at Own Reaction marker, clients state that they were unpleasantly surprised by their own reaction to an event. This marker was identified by Honos-Webb, Surko, et al. (1998) and by Greenberg et al. (1993) as the Problematic Reaction Point marker (Table 3). Theoretically, the dominant voice of the community is surprised by the emergence of the problematic voice.

Marker 6: Fearing Loss of Adaptive Functioning (APES = 1). In excerpts exhibiting the Fearing Loss of Adaptive Functioning marker, clients express the fear of losing their ability to function in their daily activities. This marker was identified by Honos-Webb, Surko, et al. (1998) as the Fear of Losing Control marker (Table 3). Theoretically, a client's dominant voice expresses its fear because the community fears losing the ability to function in daily life.

Marker 7: Feeling Painful Emotions (APES = 2). In excerpts exhibiting the Feeling Painful Emotions marker, clients express psychological pain or report recent events involving painful emotions. This marker was identified by Honos-Webb, Surko, et al. (1998) as the Pain marker (Table 3). Theoretically, this marker indicates a discrepancy between a client's problematic voice and dominant voice.

Marker 8: Feeling Vulnerable (APES = 2). In the Feeling Vulnerable marker, clients defenselessly express a negative emotion about an event or action. This marker was identified by Greenberg, et al. (1993) as the Vulnerability marker (Table 3). Theoretically, a client's affect can be intensively negative in Stage 2; this marker identifies when a client's problematic voice expresses negative affect by communicating despair, regret, or resignation.

Marker 9: Desiring Change (APES = 2). In excerpts exhibiting the Desiring Change marker, clients explicitly express a desire or need for a positive intrapersonal or interpersonal change. Theoretically, Stage 2 describes the emergence of a problematic voice; this marker identifies when the problematic voice emerges to express a desire for the community of voices to change.

Marker 10: Difficulty Articulating What's Wrong (APES = 2). In excerpts exhibiting the Difficulty Articulating What's Wrong marker, clients state that something is wrong, but they can not identify the problem. This marker was identified by Greenberg, et al. (1993), as the Unclear Felt Sense marker (Table 3). Theoretically, a client's problematic voice emerges, yet the voice does not have the language to articulate its internal experience.

Marker 11: Unfinished Business with a Significant Other (APES = 2). In excerpts exhibiting the Unfinished Business with a Significant Other marker, clients state their need to express thoughts and feelings to a significant other (such as a spouse or parent). This marker was identified by Greenberg, et al. (1993) as the Unfinished Business marker (Table 3). Theoretically, the problematic voice emerges into the community of voices' awareness in this stage; this marker describes how the problematic voice emerges to express unfinished business to the community of voices.

Marker 12: Feeling Stuck/Trapped (APES = 2). In excerpts exhibiting the Feeling Stuck-Trapped marker, clients express feeling trapped, held back, or blocked from expressing a thought, emotion, or action. This marker was identified by Honos-Webb, Surko, et al. (1998) as the Stuckness marker (Table 3). Theoretically, clients are aware of their problematic voice, but the dominant community does not have a full conceptualization of the problem, nor the insight, to move beyond the problem.

Marker 13: Feeling Confused (APES = 2). In excerpts exhibiting the Feeling Confused marker, clients express confusion, puzzlement, or bewilderment with their own thoughts, feelings or actions. This marker was identified by Honos-Webb, Surko, et al. (1998) as the

Puzzlement marker (Table 3). Theoretically, this marker identifies when a client's dominant voice is perplexed by the emergence of the problematic voice.

Marker 14: Recurring Problem (APES = 2). In excerpts exhibiting the Recurring Problem marker, clients recognize a recurring interpersonal or intrapersonal problem. Theoretically, a client's community of voices recognizes that a problematic voice recurrently emerges, but the community does not have the insight to understand the root of the problem.

Marker 15: Expressing Then Inhibiting A Need (APES = 3). In excerpts exhibiting the Expressing Then Inhibiting a Need marker, clients express or report a need, then verbally or behaviorally inhibit that need. In theory, the problematic and dominant voices are both salient; the problematic voice expresses a need whereas the dominant voice of the community inhibits the need or suppresses the expression of the need.

Marker 16: Stating Incompatible Goals (APES = 3). In excerpts exhibiting the Stating Incompatible Goals marker, a client's problematic voice and dominant voice express a wish and an incompatible wish in succession, respectively. Theoretically, both voices are salient; the client's problematic voice and dominant voice contradict each other without the awareness of doing so.

Marker 17: Conflicting Wants and Shoulds (APES = 3). In excerpts exhibiting the Conflicting Wants and Shoulds marker, clients express an emotional need, concern, or goal, and then express a conflicting societal or parental should. This marker was identified by Greenberg, et al. (1993) as the Self-evaluative Split marker (Table 3). Theoretically, both the problematic voice and dominant voice are salient in this stage; a client's problematic voice expresses itself and the dominant voice strongly opposing that voice by stating a societal or parental dictate. The dominant voice does not have the insight that the voice is societal or parental.

Marker 18: Taking Other's Values As Your Own (APES = 4). In excerpts exhibiting the Taking Other's Values As Your Own marker, clients state a connection between a current problem or need and a past opinion or action of another. In theory, a client's community has the insight that the dominant or problematic voice is an introjected voice of a significant other (e.g., mother, father, spouse).

Marker 19: Using Old Reactions In A Current Relationship (APES = 4). In excerpts exhibiting the Using Old Reactions in a Current Relationship marker, clients state that actions or reactions in their past relationships are similar to actions or reactions in the their current relationships. Theoretically, a client has an insight that the problematic feelings, thoughts or behaviors were transferred from a past relationship to a current relationship.

Marker 20: Stepping Back To Take A Better Look (APES = 4). In excerpts exhibiting the Stepping Back To Take A Better Look marker, a client is not enmeshed in either the problem or the obstacles to solving the problem; this results in the client viewing these two aspects from a more helpful perspective. Theoretically, clients in this stage gain insight into their problems; this marker identifies when a client has the insight to gain some distance from the problem and subsequently sees the connection between the problematic voice and dominant voice.

Marker 21: Putting Pieces Together In A New Way (APES = 4). In excerpts exhibiting the Putting Pieces Together in a New Way marker, clients formulate aspects of a problem and acknowledge insight into the problem by stating, “Aha,” or “This is new.” Theoretically, a client’s community of voices spontaneously comes to a new understanding of the problematic voice.

Marker 22: Deciding To Act Differently (APES = 5). In excerpts exhibiting the Deciding to Act Differently marker, clients state a decision to behave differently. In theory, a client’s problematic voice and dominant voice jointly decide to make changes based on the needs of both of them.

Marker 23: Almost, But Not Quite, Solving The Problem (APES = 5). In excerpts exhibiting the Almost, But Not Quite, Solving the Problem marker, clients acknowledge trying out new behavior. Theoretically, a client, in an attempt to resolve the problematic experience, applies insight to the problem without complete success.

Marker 24: Successfully Asserting Needs (APES = 5). In excerpts exhibiting the Successfully Asserting Needs marker, clients successfully assert themselves in an interpersonal situation where they previously had not. Theoretically, clients in Stage 5 have a reconstituted community of voices—one in which the problematic voice has successfully assimilated into the community of voices; this marker describes a voice asserting itself from this reconstituted community.

Marker 25: Noticing Change (APES = 6). In excerpts exhibiting the Noticing Change marker, clients (or someone in the clients' life) explicitly notice an intrapersonal or interpersonal change. This marker was identified by Honos-Webb, Surko, et al. (1998) as the Other's Notice Change marker (Table 3). In theory, a client resolves the problematic experience and the resolution of the problem is noticed.

Marker 26: Coming to a Solution (APES = 6). In excerpts exhibiting the Coming to a Solution marker, clients recognize the resolution of their problem. Theoretically, this marker is a generic recognition by a client that he or she has successfully assimilated the problematic voice into community.

Reliability

The kappa statistic (Cohen, 1960; Landis & Koch, 1977) was used to assess reliability. The kappa statistic measures the strength of agreement between two raters on one marker across a set of excerpts. The statistic differs from just examining the proportion of agreement on one marker by two raters by taking into account chance agreement. Table 4 shows how the frequency of observed agreement, the frequency of expected agreement by chance, and the kappa, are calculated.

Aggregating Kappas

A kappa was calculated for each pair of raters for each marker. Because of the difficulty in interpreting thousands of kappas, the kappas were aggregated in four different ways. These four ways of aggregating kappas for this dissertation are named the Rater Mean Kappa, Marker

Mean Kappa, Group Rater Kappa and Group Marker Kappa. The first two terms reflect the reliability of markers by pairs of raters, whereas the latter two terms reflect the reliability of markers based on the work of two groups of raters (Table 5).

Marker Mean Kappa (MMK) and Group Marker Kappa (GMK) refer to the consistency with which particular markers were endorsed. The MMK was calculated by averaging kappas across raters for each marker. The GMK was the kappa for each marker when markers were endorsed by two seven-rater groups.

Rater Mean Kappa (RMK) and the Group Rater Kappa (GRK) are broadly referred to as rater reliability. Rater reliability refers to the consistency with which individual raters endorse markers. Rater Mean Kappa was calculated by averaging kappas across markers for each rater. Group Rater Kappa was the GMK averaged across the markers.

All of the kappa coefficients were evaluated using the guidelines outlined by Landis and Koch (1977), where the strength of the kappa coefficients were designated as follows: kappa = 0.01 - 0.20 slight; 0.21 - 0.40 fair; 0.41 - 0.60 moderate; 0.61 - 0.80 substantial; 0.81 – 1.00 almost perfect.

Pairwise Reliabilities

Kappa. To assess how reliably a marker was endorsed, a kappa coefficient was calculated for each marker for each pair of raters for each of the three data sets. For each marker in each data set, there were 104 kappas calculated—one kappa for each of the 14 raters compared with every other rater. In all, over 8000 kappas were calculated: 14 raters by 26 markers by 3 data sets (i.e., 104 kappas by 26 markers by 3 data sets = 8112 kappas).

Marker Mean Kappa (MMK). The MMK is a statistic created to assess, on average, how reliably a particular marker was endorsed by single raters. To calculate MMKs, kappas were averaged across all pairs of rater for each marker in each data set. Seventy-eight MMKs were calculated (i.e., 26 markers by 3 data sets).

Rater Mean Kappa (RMK). The RMK is a statistic created to assess, on average, how reliably each rater endorsed markers. To calculate RMKs, the kappas for each rater were averaged across the 26 markers for each data set. Forty-two RMK were calculated (i.e., 14 raters by 3 data sets).

Group Reliabilities

The groupwise reliabilities (GMK and GRK) are analogous to the pairwise kappa reliabilities, except that the source of the data is the group's representative marker rather than a rater's marker endorsement. Group kappas were calculated using two arbitrarily selected seven-rater groups. For each excerpt, any marker endorsed by three of the seven raters in a group was considered to have been selected by that group; this marker was named the group's *representative marker*. For example, if five raters of a group endorsed marker 1 and two raters endorsed marker 2 for a passage, the representative marker would be marker 1. The criterion of three of seven raters agreeing on a marker within an excerpt accomplished two things. First, the criterion reflected at least a moderate level of agreement on a marker within that group (i.e., over 40% agreement). Second, the criterion yielded over 200 passages that contained a single representative marker.

Group Marker Kappa. The group marker kappa was a name given to kappas derived from the groups' representative markers. A group marker kappa was calculated for each marker using the groups' representative marker and shows how reliably a marker was endorsed by two groups of raters. The actual calculation can be seen in Table 4, substituting *group* for *rater* and *representative marker* for *marker*. Seventy-eight group marker kappas were calculated (i.e., 26 markers by 3 data sets).

Group Rater Kappa. The Group Rater Kappa is simply the Group Marker Kappa averaged across markers (i.e., 1 Group Rater Kappa by 3 data sets). The statistic shows the average kappa in a data set.

Reliability Results

One main research question was whether markers are reliable. Before presenting pairwise marker reliabilities (MMK), I describe how reliably each rater identified markers (i.e., RMK).

Rater Kappas

Table 6 presents how reliably, on average, individual raters endorsed markers in each data set. The results in the Bill data set showed the Rater Marker Kappas (RMK) ranged from 0.32 to 0.42 (M = 0.39, SD = 0.03), indicating that the raters, overall, had fair to moderate agreement. The RMKs in the Detert data set showed the rater reliabilities varying from 0.17 to 0.23 (M = 0.19, SD = 0.02), demonstrating slight to fair rater agreement. The RMKs in the Reid data set showed the rater reliabilities varying from 0.12 to 0.19 (M = 0.15, SD = 0.02), indicating slight rater agreement. These results show that, although raters applied markers more reliably in the Bill data set, overall, pairwise raters' level of agreement using markers was slight. The results further show that raters were not substantially different from one another with respect to their reliabilities within a particular data set.

Marker Reliability

Table 7 presents the reliability with which the average rater identified each of the 26 markers and the frequency in which the average rater endorsed markers in a particular set of excerpts. The kappas in the table describe how reliably a particular marker was endorsed in each data set, averaged across every pair of raters. The frequency listed next to each kappa is the frequency with which the average rater endorsed that marker. For example, Table 7 shows that the fourteen raters endorsed the body symptom marker an average of 1.14 times in the 59 Bill excerpts.

The Marker Mean Kappas (MMKs) varied by data set. Eleven markers had slight agreement, one had fair agreement, nine had moderate agreement, three markers had substantial agreement, and two had almost perfect agreement in the Bill data set (Table 7), with MMKs ranging from 0.00 to 0.85 (M = 0.36). The two markers with almost perfect agreement were the Feeling Stuck and Using Old Reactions in a Current Relationship. The markers with substantial agreement were the Distancing Language, Desiring Change, and Noticing Change markers. The markers with moderate agreement were Body Symptoms, Feeling Surprise, Feeling Pain, Recurring Problem, Expressing/ Inhibiting a Need, Conflicting Wants and Shoulds. The MMKs in the Detert data set ranged from 0.00 to 0.51 (Table 7), with an average kappa of 0.18. Three markers reached moderate agreement: Body Symptoms, Feeling Stuck, and Difficulty

Articulating What's Wrong markers. Six markers reached fair agreement: Feeling Confused, Recurring Problems, Old Reactions, Deciding to Act Differently, Noticing Change and Asserting Needs. The remaining seventeen markers had slight agreement. The MMKs in the Reid data set ranged from 0.00 to 0.44 (M = 0.15), with only one marker (Noticing Change) reaching moderate agreement, four markers reaching fair agreement (Recurring Problem, Others' Values, Putting Pieces Together, and Asserting Needs), and the remaining twenty-one markers reaching slight agreement (Table 7).

Correlations were computed between the MMKs and the average frequency that raters endorsed markers in each data set (Table 7). The correlations were conducted to examine whether the frequency of particular markers within a data set related to the markers' reliability (i.e., data sets having many or few examples of a marker might reveal high or low reliability coefficients of these markers). The correlation between the Bill MMKs and the frequency with which individual markers were endorsed was statistically significant, *r* (26) = .61, *p* < .001. The correlation computed between the Detert MMKs and the frequency with which an individual marker was endorsed resulted in a non-significant relationship, *r* (26) = .25, *ns*. The correlation between the Reid MMKs and the frequency with which the markers were endorsed showed a statistically significant relationship, *r* (26) = .66, *p* < .001. The significant correlations in the Bill and Reid data indicate that the more frequently a marker was used, the greater the marker's reliability, whereas the non-significant correlation indicates that there was no significant relationship between the Detert MMKs and frequency the markers were endorsed. These results suggest that some of the differences in marker reliabilities may be accounted for by a data set providing more examples of one marker than of another marker.

The pairwise kappa values in each pair of data sets (Bill and Detert, Bill and Reid, and Detert and Reid) were correlated across the 26 markers (Table 7). There was a significant positive relationship between the Bill and Detert kappas (*r* = .40, *p* < .05), but not between the Bill and Reid kappas (*r* = .25, *ns*), nor between the Detert and Reid kappas (*r* = .33, *ns*). The significant correlation implies that marker reliabilities tended to be consistent across data sets.

Group Reliability Results

Group Marker Kappas

Table 8 presents the strength of agreement between the two groups of seven raters on the 26 markers for each of the three data sets. In all three data sets, the strength of agreement ranged from .00 to 1.00. In the Bill data set eight markers reached perfect agreement, one reached almost perfect agreement, five reached substantial agreement, three reached moderate agreement, seven reached slight agreement, and three markers did not have a representative marker associated with them. In the Detert data set, four markers reached perfect agreement, one marker reached almost perfect agreement, three reached substantial agreement, three reached moderate agreement, four reached fair agreement, ten reached slight agreement, and one marker did not have a representative marker. In the Reid data set, two markers reached perfect agreement, four reached substantial agreement, six reached moderate agreement, five reached fair agreement, seven reached slight agreement, and 2 markers did not have a representative marker.

The groupwise kappa values in each pair of data sets (Bill and Detert, Bill and Reid, and Detert and Reid) were correlated across the 26 markers (Table 7). Using groups' representative markers, there were no significant relationships between Bill and Detert kappas (r = -.04, *ns*), between Bill and Reid kappas (r = .03, *ns*), nor between Detert and Reid kappas (r = .39, *ns*). These results indicate that the groupwise marker reliabilities were not consistent from data set to data set.

Group Rater Kappa

The Group Rater Kappa assessed the average strength of agreement, between the two groups, across the markers for a data set. The Group Rater Kappas were calculated for the two seven-rater groups' representative markers using the formula described in Table 4. Averaged across the 26 markers, groups of raters showed moderate reliability in the Bill data set (0.56), and fair reliability in both the Detert (0.39) and Reid (0.39) data sets.

Confusion Ratio

A confusion ratio, a statistic created for this dissertation, examines the extent to which the raters confused markers with one another. Three-hundred-and-twenty-four unique marker pairs were examined (i.e., marker 1/ marker 2, marker 1/ marker 3, … , marker 25/ marker 26), and a confusion ratio was calculated for each pair. Each of the 324 confusion ratios was calculated by dividing an "observed confusion" value by an "expected confusion" value.

An observed confusion value indicates how many excerpts were rated as containing marker A by one rater and marker B by another rater. Table 9a graphically shows a scaled-down version of observed confusion values with two raters and three markers. This table shows nine agreements: five excerpts where raters 1 and 2 both identified marker 1 and four excerpts where raters 1 and 2 both identified marker 3. There were also six excerpts where rater 1 identified marker 1, while rater 2 identified marker 2. There were also two excerpts where rater 1 identified marker 3 in a excerpt, while rater 2 identified marker 1.

The expected confusion values reflect how likely two markers are expected to be confused with one another by chance alone; the expected confusion is based upon the frequency with which markers were used. The expected confusion values are calculated by obtaining the product of the number of times each of the two markers were used, and dividing by the total number of excerpts assigned markers. In this example, cell 1's expected confusion value was obtained by calculating the product of cell 1's row total (i.e., 7) and cell 1's column total (i.e., 11), and dividing by total excerpts endorsed by markers (i.e., 17). Therefore, the expected confusion for cell 1 equaled (5 x 11) / 17 or 4.53 (see Table 9b). A similar procedure is carried out for the other cells 8 cells in this example.

The confusion ratios shown in Table 9c were constructed dividing cell 1's observed confusion value by cell 1's expected confusion value (i.e., 5/4.53 = 1.10). The same procedure was carried out for the other eight cells.

The matrix descriptions above were meant to give the reader a feel for how these ratios were calculated. In the reported results, the observed, expected, and ratio values were aggregated over raters and over symmetrical cells (e.g., A vs. B was aggregated with B vs. A). The observed values equaled the number of times two specific markers were endorsed in the same excerpt by a pair of raters, summed across all pairs of raters. The expected values equaled the product of the number of times each of two markers were used overall divided by the frequency with which all markers were used. The ratio equaled the observed confusion value divided by the expected confusion value.

Results of the Confusion Ratio Calculations

Table 10 presents fourteen marker pairs that had confusion ratios over 2.00. Each of the 324 marker pairs had a confusion ratio; 2.00 was selected as an arbitrary cutoff. Of note are four pairs of markers with confusion ratios greater than 3.00, and several "clusters" of confused markers. The marker pairs with the four largest ratios are identified below and an example is given of an excerpt where each confusion occurred. The clusters of confused markers describe situations where one marker in the table was confused with several other markers (see note in Table 10); these clusters are discussed below.

Four Largest Confusion Ratio Marker Pairs. Marker 5, Feeling Surprised, and marker 11, Difficulty Articulating What's Wrong, had a confusion ratio of 3.29. One of the confused excerpts was excerpt 23 from the Bill data set (below) where the client stated, "I'm not sure what it is," which some raters incorrectly interpreted as the client not being able to communicate their internal experience (i.e., marker 11—Difficulty Articulating marker). However, to endorse the Difficulty Articulating marker, the client must explicitly state that they have difficulty communicating their internal experience. Other raters accurately applied marker 5 to the client's statement, "it's puzzling…"; this statement met the marker's manual description.

"Th: Something about that grant is – Cl: Yeah, I don't know. Th: you're not sure. Cl: no, I'm not sure what it is, its puzzling, why…why I'm uh, getting that tense about."

Marker 10 (Feeling Confused) and marker 11 (Difficulty Articulating What's Wrong) had a confusion ratio of 4.31. Examining excerpts where confusions occurred, it appears that both markers were plausible, but the raters missed the other marker. It is important to note that when raters endorsed both markers, the two markers would be considered confused. This shows a limitation of the confusion ratio. In effect, the confusion ratio assumes that markers are mutually exclusive.

“Th: and yet you condemn yourself and you end up feeling pretty wrong. Cl: yes…I’m not sure how to describe it further. Th:…Do you have some sort of image of it, do you feel it inside? Cl: I feel like I don’t know how I feel, I feel confused.” (Bill, excerpt 59)

Marker 16 (Incompatible Goals) and marker17 (Conflicting Wants and Shoulds) had a confusion ratio of 3.09. In some excerpts, it appears that raters endorsed a marker similar to the appropriately applied marker. For example, in the excerpt below, it appears that the client was conflicted about whether to stay at her current position or not. However, the raters did not recall that an explicit “Should” is required to endorse marker 17.

“Cl: She phoned me at ten minutes to four and said they were going to offer me the job….she said, ‘you sound a bit hesitant.’ I came for the interview because I wanted it. I don’t know whether or not I really want it. I want it, yes. Th: you want it, but you aren’t sure you want it. Cl: …that part of me wants to stay where I am.” (Reid, excerpt 79)

Marker 26 (Coming to a Solution) was confused with marker 22 (Almost Solving the Problem), resulting in a confusion ratio of 3.03. It appears that raters had difficulty identifying markers in excerpts without context; raters were confused as to whether the client had resolved the problem or had only attempted to resolve the problem. For example, in the Reid excerpt 32, it may have been unclear to raters whether the client solved or attempted to solve her problem by not answering the door.

“Cl: to be honest, when the bell rang I let someone else answer it. Even though I knew that this lady would want me, but for once I didn’t care, I just had enough, and let someone else take care of it.”

Three Clusters of Confused Marker Pairs. Three clusters of confused marker pairs emerged from the data (Table 10). The term cluster is meant to suggest that a collection of markers are systematically confused with each other. Markers might be regularly confused with each other for two primary reasons: 1) the two markers are in the same passage and the rater missed one or the other, or 2) two markers have the same overarching (and sometimes implicit) construct (e.g., stage of assimilation), are connected in the mind of the rater, and interchangeably endorsed by the raters.

The three clusters of confused markers were: Cluster 1, comprised of Marker 5 (Surprised at Own Reaction), 10 (Feeling Confused), 11 (Difficulty Articulating), and 21 (Putting Pieces Together); Cluster 2 was comprised of Markers 15 (Expressing, Inhibiting a Need), 16 (Incompatible Goals), and 17 (Wants and Shoulds); Cluster 3 was comprised of Markers 22 (Almost, But Not Quite), 23 (Deciding to Act Differently), 24 (Noticing Change), 25 (Asserting Needs), and 26 (Coming to Solution).

The confusion clusters illuminate different kinds of confusion. The first is when two markers are tapping the same construct, as in the excerpted example of markers 16 and 17 above. A second possibility is that two markers are present in the same excerpt. One might reasonably expect more than one marker from the same or adjacent stage given a problem's level of assimilation. For example, in the excerpt below, raters apparently selected one of several markers because they thought they knew whether the client had reached a solution (Marker 26, Stage 6), a partial solution (Marker 22, Stage 5), or whether several other markers applied to the excerpt (e.g., Asserting Needs—marker 25, Stage 6; Noticing Change—marker 24 Stage 5; or Deciding to Act Differently—marker 23, Stage 5). The manual states that every marker that is present should be endorsed; in this passage, it seems that markers 23, 24, and 26 meet the criteria and should be endorsed.

> Th: … you are thinking about things differently [Noticing Change], it seems. I'm also remembering a funeral recently that you decided not to go to [Deciding To Act Differently], in spite of feeling the very real pressure coming from the outside.
>
> Cl: Hm, yes.
>
> Th: you were actually able to say no [Asserting Needs], it does not feel right for me to go.

> Cl: Hum. I never thought of that. Yeah, cause that could have been quite a ..a hard time. I could have gone, to please my sister and that…but I wouldn't have felt good about it myself. (Mary, excerpt 68, session 14, line 409-415).

Validity

Two types of validity were examined: convergent validity and construct validity. To assess the convergent validity, I analyzed the extent to which APES stages associated with raters' marker endorsements agreed with Reid's and Detert's APES ratings. To assess the construct validity, I examined the extent to which APES stage of a passage (i.e., APES stages associated with raters' marker endorsements—described in detail immediately below) increased with the session number of an excerpt in a successful therapy case, as predicted by theory.

Associating Raters' Marker Endorsements to APES Stages

Every marker is a marker of a particular APES stage. In Table 11, the two left-most columns list the 26 markers and their corresponding APES stage. For the validity analyses that follow, each of the fourteen rater's marker endorsements was associated with its corresponding APES stage. That is, although my raters did not actually assign APES stage ratings, for expository purposes, *raters' APES ratings* refer to APES stages associated with raters' corresponding markers endorsements. If no marker was endorsed, there would be no rater APES rating.

Determining a Representative APES Ratings for Each Excerpt Based on the Fourteen Raters' APES Ratings

To conduct the correlation and crosstab statistics, a single APES rating, based on the work of the fourteen raters, was derived for each excerpt—this single APES rating will be called the *representative APES rating*. A representative APES rating was chosen using the criterion that at least five of the fourteen raters had to have selected markers indicating the same APES rating. For example, in excerpt 1 of the Detert data set, eleven raters endorsed markers indicating APES stage 1, whereas three raters endorsed markers indicating APES stage 2; therefore, stage 1 was designated as the representative APES stage for this excerpt. In cases where five (or more) raters endorsed markers indicating one APES rating and five (or more) raters did not endorse any markers, I used the one APES rating as the representative APES rating. In cases where markers indicating two APES ratings were endorsed by five (or more) raters each, I omitted both APES ratings from the analyses. The criterion of at least five raters agreeing on an APES rating yielded a large proportion of the excerpts (157 of 188) as having a representative APES rating. Theoretically, lowering the criterion to two or three raters agreeing on a marker could have led to divergent APES ratings for a passage; if situations such as this occurred, no representative APES rating would be assigned.

Determining a Mean and Modal APES Ratings from Detert's APES Ratings

To conduct correlation and crosstab statistics, two sets of APES ratings were calculated from Detert and his raters' APES ratings (Detert et al., 2002): a mean and modal APES ratings for each excerpt. Calculating these two different sets of APES ratings (mean and mode) permitted the examination of several statistical relationships (e.g., a correlation between Detert's mean APES ratings and raters' representative APES rating, a correlation between Detert's modal APES ratings and raters' representative APES rating, and a crosstab between Detert's mean APES ratings and raters' representative APES rating). Detert's mean APES ratings were calculated by averaging his and his four raters' APES ratings. Detert's modal ratings were calculated by assessing the most frequent APES ratings (rounded to the nearest whole APES stage) that his and his colleagues assigned to a particular excerpt.

Determining a Mean and Consensus APES Ratings from Reid's APES Ratings

Two sets of APES ratings were obtained from Reid (Reid, et .al, 2001): a mean APES rating and a consensus APES rating. These ratings were used in the crosstab and correlation analyses. For each excerpt in the Reid data set, a mean APES rating was calculated by averaging Reid and her colleague's APES ratings for each excerpt. A consensus APES rating was obtained from Reid when, after discussion, she and her colleague came to a consensus on an APES rating for an excerpt.

Convergent Validity in the Detert Data

Assessing Convergent Validity by Correlating Detert's and Raters' APES Ratings. Raters' representative APES ratings were moderately correlated to the Detert's mean APES ratings, $r = .52$ (n=51), $p < .001$, and to the Detert's modal APES ratings, $r = .42$ (n=29), $p < .05$. These correlations indicate that my raters assigned APES ratings in a similar fashion as Detert's raters APES ratings. For example, when Detert's raters assigned late stage APES ratings to excerpts, my raters assigned late stage APES ratings to those excerpts.

Assessing Convergent Validity by a Crosstab between Detert's and Raters' APES Ratings. To give a more detailed picture of the relationship between my raters' APES ratings and Detert's mean APES ratings, I conducted a cross tabulation to describe their level of agreement. Table 11 presents a crosstab with Detert's mean APES ratings along the columns and the raters' APES ratings along the rows. Table 11 describes the frequency of agreements and disagreements between raters' APES ratings and Detert's APES ratings. For example, Table 11 shows that my raters endorsed marker 8 (corresponding APES stage 2) in eighteen excerpts across the Detert data set. For these 18 excerpts rated at stage 2 by my raters, Detert's raters did not endorse APES stage 0 in any excerpt, while stage 1 was endorsed in one excerpt, stage 2 was endorsed in ten excerpts, stage 3 was endorsed in six excerpts, stage 4 was endorsed in one excerpt, and stages 5-7 were not endorsed in any excerpts. These ten stage 2 endorsements represent agreements.

I calculated a percentage of agreement between raters' APES ratings and Detert's raters' APES ratings by dividing the frequency of excerpts that my raters agreed with Detert's APES ratings by the frequency of excerpts in which the marker was used (shown in the right most column). For example, in Table 11, marker 8, the raters' APES ratings agreed with Detert's APES ratings in 10 of 18 excerpts or 55.6% of the time. An asterisk in the table indicates when the percentage of agreement between my raters' APES ratings and Detert's APES ratings was 40% or greater.

Convergent Validity in the Reid Data

Assessing Convergent Validity by Correlating Reid's and Raters' APES Ratings. Correlations between Reid's ratings and my raters' representative stage ratings showed that my raters' APES ratings were moderately correlated to both Reid's mean APES ratings, $r = .61$ (n=34), $p < .001$, and to Reid's consensus APES ratings, $r = .59$ (n=27), $p < .001$. These correlations indicate that my raters assigned APES ratings in a similar fashion as Reid and her colleague's APES ratings. For example, when Reid and her colleague assigned late stage APES ratings to excerpts, my raters assigned late stage APES ratings to those excerpts.

Assessing Convergent Validity by a Crosstab between Reid's and Raters' APES Ratings. Table 12 shows the crosstab between Reid's consensus APES ratings and my raters' APES ratings (again, based on the work of all fourteen raters). In nine instances—indicated by the asterisks—my raters' APES ratings had 40% or more agreement with the Reid's consensus APES ratings.

Construct Validity of Assimilation Markers

To assess the construct validity of the Assimilation Markers, a correlation was computed between my raters' APES ratings and the session number of the Reid therapy. The results showed a moderately, positive correlation, $r = .51$ (n= 52), $p < .001$. An additional correlation was conducted between my raters' APES ratings and the session number of the Bill therapy. The results were statistically significant, $r = .40$ (n=48), $p < .01$. These results provides support for the construct validity of the markers in as much as assimilation stages are expected to increase across psychotherapy sessions in successful therapy cases, as in the successful cases of Reid and Bill. An analysis was not conducted on the Detert data set because half of the patients were unsuccessful and the range of sessions for all patients was quite small (only two sessions).

CHAPTER 4: DISCUSSION

In this study, I examined whether markers of assimilation may be reliably identified in excerpts of psychotherapy transcripts and whether these markers are valid indicators of APES stages. In examining these questions, this chapter: 1) summarizes the reliability of the markers; 2) conceptually applies Signal Detection Theory (Green & Swets, 1988) to help understand the varying reliabilities across data sets and the impact of the criteria I used in the data analyses, 3) discusses the similarity of rater reliabilities within data sets; 4) discusses and qualifies the validity of the markers; 5) evaluates the success of the markers; 6) proposes methods to improve marker reliability; and 7) draws conclusions.

Reliability of Markers

The markers' reliabilities were assessed by kappa coefficients. All of the kappa coefficients were evaluated using the Landis and Koch (1977) standards (kappa from 0.01 - 0.20 = slight agreement; 0.21 - 0.40 = fair; 0.41 - 0.60 = moderate; 0.61 - 0.80 = substantial; 0.81 – 1.00 = almost perfect agreement to perfect agreement). The kappa coefficients were calculated using raters' marker endorsements. Two types of marker reliabilities were derived from these endorsements: pairwise reliabilities and group reliabilities. This section will describe the pairwise and groupwise reliabilities, an explanation for why the groupwise reliabilities were higher than the pairwise reliabilities, and a comparison of these findings to Honos-Webb's (1999; Honos-Webb et al., 2003) findings.

In the pairwise reliabilities, three markers had moderate agreement (.41 ≤ kappa ≤ .60) in two of three data sets (Table 7). Group reliabilities are the kappa coefficients based on aggregated marker endorsements (i.e., at least three of seven raters agreeing on a marker). In the groupwise analyses, three markers had perfect to almost perfect agreement (.81 ≤ kappa ≤ 1.00) in two of the three data sets; three different markers had substantial agreement (.61 ≤ kappa ≥ .80) in two of the three data sets; and the 20 other markers had only moderate, fair, or slight agreements (.01 ≤ kappa ≥ .60) in two of the three data sets (Table 8).

The number of markers with at least moderate agreement differed between the groupwise and pairwise analyses. The increased number of markers with at least moderate reliabilities in the groupwise analyses, compared to the pairwise analyses, reflects the calculations of the groupwise and pairwise kappas. The pairwise kappas reflect every instance that a rater missed a marker, whereas the groupwise kappas permitted a few raters to miss a marker without lowering the reliability (i.e., perfect reliability was possible even if four of seven raters in each group missed the marker).

I found a similar number of reliable markers as Honos-Webb (1999) found in her study. In my pairwise analyses, three of the 26 markers (Body Symptoms, Feeling Stuck, and Noticing Change) achieved at least moderate reliability (.40 ≤ kappa) in two of the three data sets, whereas she found five of her twenty-five markers to have moderate to substantial reliability (.40 ≤ kappa < .75; Somatic symptoms, Pain, Metaphor of Downward motion, Understanding Personal Historical Roots, and Others Notice Change). Further, two markers, Body Symptoms (Honos-Webb pairwise comparison, kappa =.47) and Noticing Change (Honos-Webb pairwise comparison, kappa =.58), were reliable at this level of agreement in both studies.

Using Signal Detection Theory to Understand the Reliability of Markers

The ability of raters to make reliable marker endorsements may be considered a signal detection problem. One could consider the marker as the signal and the surrounding text as noise. Sometimes markers may have been more easily detected because there was less noise in the passage. Noise in the data can influence rater reliability. For example, using handpicked excerpts (i.e., the Bill excerpts were selected and edited with markers in mind) resulted in greater rater reliability. In the Bill data set, the average kappa in the pairwise analyses was 0.39, and 0.56 in the group analysis. In the Detert and Reid data sets, the average pairwise kappas were 0.19 and 0.15, respectively, and 0.39 for both group analyses. The greater kappa coefficients in the Bill data likely reflect my selecting the excerpts. I excerpted particular passages because of their strong signal of a marker and I did not include much surrounding context—reducing the noise (i.e., context surrounding the marker is noise in this task). These excerpts may have made the raters' task of identifying markers easier. The finding that material selected specifically for marker research resulted in greater agreement among raters, rather than in unedited material, is consistent with Honos-Webb's (1999) study. Honos-Webb found acceptable marker reliabilities in both individual and group raters' ratings using handpicked, edited excerpts. Therefore, excerpts with less noise seem to enhance the markers' reliability.

Although the data sets were differentially noisy, the markers' kappa coefficients were moderately consistent across data sets. When the pairwise kappa values in each pair of data sets (Bill and Detert, Bill and Reid, and Detert and Reid) were correlated across the 26 markers, one of these three correlations (the Bill and Detert correlation) was significant. This indicates that the strength of the kappas in the Bill data were related to the strength of kappas in the Detert data (if there was one unreliable set of kappas, as it appears to be the case in the Reid data, the two good data sets would be correlated with each other, but neither of the good data sets would be correlated with the unreliable data set). This suggests that some markers may have been easier to detect and were at least modestly consistent across data sets.

Signal Detection Theory (Green & Swets, 1988) can be used to understand how the representative marker criterion impacted the group kappa results. Representative markers were identified when three of the seven-rater group agreed on a marker. Raising the criterion to 4 or 5 of 7 agreements within a group would result in fewer hits and false alarms, and more misses and correct rejections. It is important to note that in any cutoff of hits, there is a tradeoff with false alarms. As McFall and Treat (1999) state, "Selection of an optimal cutoff value necessarily involves specification of a function to be maximized [e.g., hits, misses, false alarms]. Thus there is no true and unique optimal cutoff value… It is also important to reiterate that there is no absolute optimal cutoff value….Ultimately, users have no option but to pay their money and make their choice (p233-234)." I sought to identify representative markers in the raters' endorsements if there was a marker present. In signal detection language, when I selected the criterion, I preferred to have hits and false alarms to misses.

Raters' Reliabilities Did Not Differ Within Data Sets

Raters in this sample did not vary much from each other in their skill at rating markers within a particular data set. As understood within the signal detection framework, in principle, some raters' marker endorsements could be made using very strict thresholds (i.e., those afraid to make a mistake, so called nay sayers), resulting in accurately not endorsing markers that are not present (i.e., correct rejection) and missing markers that are legitimately present (i.e., miss). On the other hand, other raters could apply a loose selection threshold when making decisions (i.e., those afraid to miss a marker, so called yea sayers), with the result of correctly identifying markers (i.e., hits) and endorsing markers that are not present (i.e., false alarms). The raters' average kappa ranged from 0.32 to 0.42 (SD = 0.03) in the Bill data set, 0.16 - 0.23 (SD = 0.02) in the Detert data set, and from 0.12 to 0.19 (SD = 0.02) in the Reid data set. These findings suggest that the raters' thresholds were similar to each other.

Validity of Markers

Convergent Validity of Markers

Convergent validity in this study is the extent to which my raters' APES ratings and the independent researchers' APES ratings were in agreement. To assess convergent validity, I examined the correlations between my raters' APES ratings (inferred from their marker endorsements) and the researchers' APES ratings. Statistically significant correlations were found between my raters' APES ratings and Deterts' APES ratings ($.40 \leq r \leq .59$, $p < .001$), and between Reids' APES ratings and my raters' APES ratings ($.42 \leq r \leq .52$, $p < .001$). Although all of these APES ratings reflect some measurement error, these findings suggest that these sets of ratings tended to converge, and support a marker-based method to indicate APES stages.

These findings that raters' and researchers' rating were correlated are similar to Honos-Webb's (1999) findings. She found moderate to strong correlations between her raters' APES ratings and her own APES ratings ($.59 \leq r \leq .82$, $p < .001$). This study differed from hers in that I compared my raters' ratings to independent ratings from two sources (Detert, et al., 2002, Reid, 2001), whereas she used one case to compare her own ratings to her raters' ratings. Honos-Webb's (1999) ratings may have been biased in that she had an investment in assigning APES ratings that were associated with her markers. The present findings suggest that raters' and independent investigators' ratings can converge without any bias from the primary investigator.

The marker-based APES ratings that converged with the independent researchers' APES ratings tended to come from Stages 2 (Vague Awareness) and 3 (Problem Statement) (see Tables 11 and 12). In the cross tabulation analyses, the validity of each marker was assessed by comparing independent investigators' APES ratings to raters' markers. Eleven markers showed greater than 40% agreement between raters' and independent researchers' endorsements. Despite the limited range of APES ratings assigned by an independent researcher (e.g., in the

Reid consensus data, there were no stage 0, 5, 6, 7 ratings), these finding lend further modest support to markers as a method of assigning APES ratings.

Construct Validity of Markers

Two moderately positive correlations were found between my raters' APES ratings and the session number of that excerpt in the Bill (*r* (48) = .40, *p* <.01) and Reid (*r* (57) = .51, *p* < .001) excerpts. The raters did not know what session the excerpts were drawn from or what markers were associated with particular APES stages. For successful therapy cases, such as the Reid and Bill cases, these two correlations provide support for the construct validity of the markers in as much as higher assimilation ratings are expected to be endorsed as the session number of the therapy from which the passage was drawn increases. Detert's excerpts, with only two sessions per case, had too few sessions to adequately assess this type of validity.

Qualification to the Validity of Markers

The criterion I selected for the *rater's APES ratings* likely influenced the degree of convergence between the raters' APES ratings and the independent researchers' APES ratings. In the validity analyses, the rater's APES ratings were based on five of the fourteen raters' agreeing on a marker in an excerpt. In signal detection language, increasing the criterion from five raters to ten raters could have the effect of decreasing the number of hits and false alarms of these markers, and increasing the number of misses of the markers. The criterion I selected capitalized on hits and false alarms, minimized misses, and by doing so, probably impacted the convergent validity results.

Evaluation of Successful, Potentially Successful, and Failed Markers

Based on the level of agreement in the endorsements of two seven-raters groups (i.e., in the group reliability analyses) and the high level of agreement between raters' and independent researchers' endorsements (i.e., in the crosstab analyses assessing validity), six markers (Desiring Change, Getting Stuck, Feeling Confused, Feeling Vulnerable, Recurring Problem, and Difficulty Articulating What's Wrong) could be considered successful because they were reasonably reliable (i.e., they reached at least moderate agreement, kappa ≥ .41, in two of the three data sets. In addition, validity was supported by at least 40% agreement between raters and researchers in the cross tabulation analyses. Seven markers (Body Symptoms, Distancing Language, Taking Other's Values as Your Own, Using Old Reactions in a Current Relationship, Putting Pieces Together, Noticing Change, and Successfully Asserting Needs) achieved at least moderate agreement levels in the group analyses, but not very high agreement in the cross tabulations (i.e., they were reliable but not valid in terms of the criterion I used). Additionally, two of the above markers, Getting Stuck and Noticing Change, were reliable in the Honos-Webb et al. (2000) study. The remaining thirteen markers (Downplaying Negativity, Avoiding Responsibility, Feeling Surprise, Fearing Loss, Feeling Pain, Unfinished Business, Expressing and Inhibiting, Incompatible Goals, Wants and Shoulds, Stepping Back, Almost but not Quite, Deciding to Act Differently, Coming to a Solution) did not reach moderate reliability (kappa < .41) in two of the three group data sets and could be considered failures. One caution to these evaluations is that the markers' reliabilities varied from data set to data set. Therefore, it is advisable to conduct additional tests of these markers on other data sets to examine their reliability and validity.

Improving Marker Reliability

Given the low reliabilities in the pairwise kappas and the results from the confusion ratios, future research could improve the reliability of the markers by improving the marker's descriptions, refining the training of the raters, examining markers a few at a time, and using clinically sophisticated raters.

Improving Marker Reliability By Better Marker Descriptions

The results suggest that the descriptions of the markers could be improved. The results of the group kappas (kappas accommodating some raters missing a marker) indicated stronger reliabilities than the pairwise kappa (kappas that reflected every missed marker). These results suggest that markers can be identified, but that the description (i.e., the signal) can be stronger for some raters than for others such that some raters do not miss the marker. Perhaps better descriptions of the markers could reduce the number of missed markers.

Another indication of the need for better descriptions is that, despite the low frequency of occurrence of many markers in a data set, some markers were more reliable (i.e., they had a stronger signal) than others. Better descriptions could yield stronger signals and better reliability.

One way to write better descriptions would be to interview the raters after missing or incorrectly endorsing (i.e., false alarm) a marker. The raters could point to the part of the description that led to their endorsement or led them to be dissuaded them from endorsing a marker. A second method to strengthen the manual would be to combine the confused markers within the same stage. That is, markers with similar characteristics within the same stage (e.g., Incompatible Goals marker and Wants/Should marker) could be collapsed into a single marker.

Training of Raters May Improve Reliability

Raters were similar to each other in their ability to endorse markers within a particular set of passages. However, their reliability was lower (GMK = .39 to .56) than I had hoped. Future research might examine whether increasing the time to train raters, reducing the raters' training group size, and allowing more time for individualized questions could improve the whole group's ability to reliably endorse markers. Raters make decisions about endorsing markers with some uncertainty. Training and practice would permit raters to more accurately apply the manual when making these decisions. This accuracy could lead to a greater likelihood of correctly endorsing markers that are present, rejecting markers that are absent, while reducing the likelihood of endorsing markers that are not actually present and missing markers that are present.

Additionally, investigators could give more feedback to raters about the kinds of errors they make when endorsing markers so raters may alter the criteria they use. That is, the investigator can help the raters adjust their yea sayer or nay sayer bias.

Examine the Reliability When a Few Markers are Studied at One Time

Reports from my raters, and my observations of the raters' coding excerpts with 26 markers, suggest that the task of endorsing markers was intellectually taxing (e.g., continually referring back through the 82-page manual). Making the task less cognitively complex might result in fewer missed markers. When raters are looking for a marker, the other markers may become a distraction (i.e., noise) with respect to the marker of interest. Thus, future research could examine the reliability of markers when raters are permitted to become sophisticated in understanding and applying a marker from each stage, one at a time, rather than having to consider 26 markers from seven stages at once.

Use Clinically Sophisticated Raters

Using clinically unsophisticated raters to endorse markers may have led to low reliabilities. Honos-Webb (1999) had found that graduate students were more reliable than undergraduate raters (graduate students had reliabilities from .70-.90, while undergraduates had reliabilities from .61-.84). In hindsight, it makes sense that unsophisticated raters who came from many disciplines would introduce more error variance in the data than a group of mostly clinical, graduate researchers. Therefore, having selected clinically unsophisticated raters likely led to lower reliabilities than had I utilized clinically trained graduate students.

Conclusion

This study investigated whether markers of assimilation stages could be reliably identified, and if those markers were valid indicators of assimilation stages. In addressing these questions, this study contributes to the marker literature by: 1) identifying markers in excerpts drawn from several theoretical orientations of psychotherapy and markers in clinical material unrelated to the manual's construction; 2) presenting some evidence that raters can reliably endorse markers in spite of the noise within data sets and without knowledge of assimilation theory; and 3) providing evidence that raters' marker-based APES ratings correspond with Detert's and Reid's APES ratings.

The results support the notion that there are some good and poor markers whose reliabilities are moderated by noise in the data set rather than the raters. These findings suggest that data used to test the markers (e.g., data with varying noise and varying frequencies of markers within the data) and the independent researchers' ratings (e.g., researchers' ratings with limited ranges of APES ratings in the data), play a role in these results, and will likely play a role in future marker research as well.

REFERENCES

Barkham, M., Shapiro, D. A., Hardy, G.E., & Rees, A., (1999). Psychotherapy in tow-plus-one sessions of a randomized controlled trail of cognitive-behavioral and psychodynamic-interpersonal therapy for subsyndromal depression. *Journal of Consulting and Clinical Psychology, 67(2),* 201-211.

Barkham, M., Stiles, W. B., Hardy, G. E., & Field, S. D. (1996). The assimilation model: Theory, research and practice guidelines. In W. Dryden (Ed.), *Research in Counseling and Psychotherapy: Practical Application* (pp1-24). London: Sage.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561-571.

Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement, 20,* 37-47.

Derogatis, L.R. (1983). *SCL-90-R: Administration, scoring, and procedural manual—II.* Baltimore: Clinical Psychometric Research.

Detert, N. (2000). Assimilation in 2 + 1 Brief Therapy. Unpublished doctoral dissertation, Open University, Oxford, England.

Detert, N., Llewelyn, S., Hardy, G.E., Barkham, M. & Stiles, W. B. (2002). Assimilation in Good- and Poor-Outcome Cases of Very Brief Psychotherapy for Mild Depression. Submitted to Journal of Counseling Psychology.

Glick, M. J., Stiles, W. B., & Greenberg, L. S. (2000). *Assimilation Patterns in a Case of Client-Centered Psychotherapy.* Paper presented at the 2000 Society for Psychotherapy Research. Chicago, IL.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (2nd ed). Peninsula Publishing, Los Altos, CA.

Greenberg, L. S. & Foerster, F. S. (1996). Task analysis exemplified: The process of resolving unfinished business. *Journal of Consulting and Clinical Psychology, 64,* 439-446.

Greenberg, L.S., Rice, L.N., & Elliott, R. (1993). *Facilitating Emotional Change: The Moment-by-Moment Process.* Guilford Press; NY.

Greenberg, L.S. & Watson, J. (1998). Experimental therapy for depression: Differential effects of client-centered relationship conditions and process experiential interventions. *Psychotherapy Research, 8(2),* 210-224.

Honos-Webb, L. (1998). *Development of a Manual for Rating Assimilation in Psychotherapy.* An unpublished dissertation. Miami University: Oxford, OH.

Honos-Webb, L., Lani, J. A., & Stiles, W. B. (1999) Discovering makers of assimilation stages: The fear of losing control marker. *Journal of Clinical Psychology, 55 (12),* 1441-1452.

Honos-Webb, L. & Stiles, W. B. (1998). Reformulation of assimilation analysis in terms of voices. *Psychotherapy, 35*, 23-33.

Honow-Webb, L., Stiles, W.B., Greenberg, L.S. (2003). A method of rating assimilation in psychotherapy based on markers of change. *Journal of Counseling Psychology, 50,* 189-198.

Honos-Webb, L., Stiles, W. B., Greenberg, L. S., & Goldman, R. (1998). Assimilation analysis of process-experiential psychotherapy: A comparison of two cases. *Psychotherapy Research, 8 (3),* 264-286.

Honos-Webb, L., Surko, M., & Stiles, W. B., (1998). Manual for Rating Assimilation in Psychotherapy: February 1998 Version. In L. Honos-Webb's *Development of a Manual for Rating Assimilation in Psychotherapy.* An unpublished dissertation. Miami University: Oxford, OH.

Honos-Webb, L., Surko, M., Stiles, W. B., & Greenberg, L. (1998). Voices in psychotherapy: The case of jan. *Journal of Counseling Psychology, 46 (4),* 1-13.

Horowitz, l.M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56,* 885-892.

Hubert, L. (1977) Kappa revisited. *Psychological Bulletin, 84(2),* 289-297.

Knobloch, L. M., Endres, L. M., Stiles, W. B. & Silberschatz, G. (2001). Convergence and divergence of themes in successful psychotherapy: An assimilation analysis of the case of vicky. *Psychotherapy: Theory, Research, Practice, Training, 38(1),* 31-39.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lani, J. A., Brandenburg, C., Glick, M. J., Osatuke, K., & Stiles, W. B. (2000). *Identification and Validation of markers of Assimilation of Problematic Experiences Scale Stage.* Paper presented at the 2000 Society for Psychotherapy Research. Chicago, IL.

Lani, J.A., Stiles, W.B., Shaikh, A. & Silberscatz, G. (1998). Hearing change in clients' narratives: An assimilation perspective. Paper presented at the Society for Psychotherapy Research Meeting, Snowbird, Utah.

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that "Everyone has won and all must have prizes? *Archives of General Psychiatry, 32,* 995-1008.

McFall, R.M., and Treat, T. A. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology, 50,* 215-241.

Osatuke, K., Stiles, W. B., Shapiro, D. A. & Barkham, M. (2000). *Assimilation Patterns in a Cognitive-Behavior Therapy Case.* Paper presented at the 2000 Society for Psychotherapy Research. Chicago, IL.

Perls, F. (1969). *Gestalt Therapy Verbatim.* Moab, Utah: Real People Press.

Prochaska, J. O., & DiClemente, C. C. (1984). *The transtheoretical approach: Crossing the boundaries of therapy.* Homewood, IL: Dow Jones-Irwin.

Quanta Healthcare Solutions, Inc. (2002, October). Calculating the Kappa Coefficient for 2 Observations by 2 Observers. *The Medical Algorithms Project.* Retrieved January 3, 2003, from http://www.medal.org

Reid, M. & McLeod, J. (2001). *Working with defensive resistance: A psychotherapy case study with a functional abdominal pain patient.* Presented at the Society for Psychotherapy Research. Leiden, Netherlands.

Rice, L. N. & Greenberg, L. S. (1984). The new research paradigm. In L. N. Rice and L. S. Greenberg's *Patterns of Change: Intensive Analysis of Psychotherapy Process.* New York: Guilford Press.

Shrout, P.E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in accessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Spitzer, R.L. & Fleiss, J.L. (1974). A re-analyses of the reliability of psychiatric diagnosis. *British Journal of Psychiatry, 125,* 341-347.

Stiles, W. B. (1993). Quality control in qualitative research. *Clinical Psychology Review, 13,* 593-618.

Stiles, W. B. (1997). Signs and voices: Joining a conversation in progress. *British Journal of Medical Psychology, 70,* 169-176.

Stiles, W. B., Barkham, M., Shapiro, D. A., & Firth-Cozens, J. (1992). Treatment order and thematic continuity between contrasting psychotherapies: Exploring implications of the assimilation model. *Psychotherapy Research, 2 (2),* 112-124.

Stiles, W. B., Elliott, R., Llewelyn, S. P., Firth-Cozens, J. A., Margison, F. R., Shapiro, D. A., & Hardy, G. (1990). Assimilation of problematic experiences by clients in psychotherapy. *Psychotherapy, 27,* 411-420.

Stiles, W. B., Honos-Webb, L., & Lani, J. A. (1999). Some functions of narrative in the assimilation of problematic experiences. *Journal of Clinical Psychology, 55 (10),* 1-14.

Stiles, W. B., Meshot, C. M., Anderson, T. M., & Sloan, W. W. (1992). Assimilation of problematic experiences: The case of john jones. *Psychotherapy Research, 2 (2),* 81-101.

Stiles, W. B., Morrison, L. A., Haw, S. K., Harper, H., Shapiro, D. A., & Firth-Cozens, J. (1991). Longitudinal study of assimilation in exploratory psychotherapy. *Psychotherapy, 28,* 195-206.

Stiles, W. B., Shanklank, M. C., Wright, J., & Field, S. D. (1997). Aptitude-treatment interventions based on clients' assimilation of their presenting problem. *Journal of Consulting and Clinical Psychology, 65 (5),* 889-893.

Stiles, W. B., Shapiro, D. A. & Elliott, R. (1986). Are all psychotherapies equivalent? *American Psychologist, 41(2),* 165-180.

Varvin, S. & Stiles, W.B. (1999). Emergence of severe traumatic experiences: An assimilation of psychoanalytic therapy with a political refugee. *Psychotherapy Research, 9(3),* 381-404.

Author Note

I thank Leslie Greenberg, Neils Detert, Mary Reid, and George Silberschatz for providing clinical material for this dissertation.

Footnotes

[1] Because of their length, four of Detert's excerpts were formed from two original excerpts. This resulted in a total of eighty-two excerpts.

[2] It should be noted, however, that the reliability of groups to endorse markers described in this section (i.e., arbitrarily arranged two groups of seven raters) was different than the groups used to test the validity of the system (i.e., where any five of the fourteen raters identifying the same marker constituted the representative marker).

Table 1

Assimilation of Problematic Experiences Scale

0. Warded off/dissociated. Client is unaware of the problem. Problematic voice is silent. Affect may be minimal, reflecting successful avoidance.

1. Unwanted thoughts/active avoidance. Client prefers not to think about the experience. Problematic voices emerge in response to therapist interventions or external circumstances and are suppressed or avoided. Affect involves unfocused negative feelings; their connection with the content may be unclear.

2. Vague awareness/emergence. Client is aware of a problematic experience but cannot formulate the problem clearly. Problematic voice emerges into sustained awareness. Affect includes acute psychological pain or panic associated with the problematic material.

3. Problem statement/clarification. Content includes a clear statement of a problem -- something that can be worked on. Opposing voices are differentiated and can talk about each other. Affect is negative but manageable, not panicky.

4. Understanding/insight. The problematic experience is formulated and understood in some way. Voices reach an understanding with each other (a meaning bridge). Affect may be mixed, with some unpleasant recognition but also some pleasant surprise.

5. Application/working through. The understanding is used to work on a problem. Voices work together to address problems of living. Affective tone is positive, optimistic.

6. Resource/problem solution. Client achieves a successful solution for a specific problem, representing flexible integration of multiple voices. Affect is positive, satisfied.

7. Integration/mastery. Client automatically generalizes solutions; voices are integrated (serving as resources in new situations). Affect is positive or neutral (i.e., this is no longer something to get excited about).

Table 2.

Markers of Stages of Assimilation used by Honos-Webb (1999).

Marker #	Marker Name
0A	Somatic symptoms
1A	Abrupt change of subject
1B	Contradictory narrative
1C	Fear of losing control
1D	External focus
2A	Problematic reaction point
2B	Pain
2C	Puzzlement
2D	Unequal weighting
2E/3E	Absence of/ reflexivity
2F	Metaphor or downward motion
3A	Emergence from embeddness
3B	Stuckness
3C	Clarity
3D	Equal weighting
4A	Flexible use of voice
4B	Resolution of self-evaluative split
4C	Resolution of unfinished business
4D	Understanding personal historical roots
5A	Exploring possible solutions
5B	Generalized application
6A	Pride marker
6B	Specific success
6C	Others notice change
7	No markers

Table 3.

Cross-Reference: Present Markers to Other Researchers' Markers

Present manual	**Honos-Webb et al. (1998)**	**Greenberg et al. (1993)**
Body Symptoms	Somatic Symptoms	
Fearing Loss of Adaptive Functioning	Fear of Losing Control	
Getting Stuck/Feeling Trapped	Stuckness	
Feeling Painful Emotions	Pain	
Feeling Confused	Puzzlement	
Feeling Surprised at Own Reaction	Problematic Reaction Point	Problematic Reaction Point
Difficulty Articulating What's Wrong		Unclear Felt Sense
Unfinished Business with a Significant Other		Unfinished Business
Feeling Vulnerable		Vulnerability
Conflicting Wants and Shoulds		Self-evaluative Split
Noticing Change	Others' Notice Change	

Table 4.

Calculation of a Kappa Statistic for a Marker Between Rater 1 and Rater 2

	Rater 1		
Rater 2	Marker is present in a set of excerpts	Marker is absent in a set of excerpts	Subtotal
Marker is present in a set of excerpts	A	B	A + B
Marker is absent in a set of excerpts	C	D	C + D
Subtotal	A + C	B + D	A + B + C + D

Observed agreement = (A + D)

Expected Agreement = (((A + B) * (A + C)) + ((C + D) * (B + D))) / (A + B + C + D)

Kappa = ((observed agreement) – (expected agreement)) / ((A + B + C + D) – (expected agreement))

Note: A, B, C, and D are the frequencies in which a marker is identified in same excerpt between rater 1 and rater 2.

(Quanta Healthcare Solutions, Inc., 2002)

Table 5.

Name, Acronym, and Description of Types of Kappa Calculations

Name	Acronym	Description
		Pairwise Ratings
Kappa	Kappa	Strength of agreement between two raters on one marker for a data set
Marker Mean Kappa	MMK	Kappa of a marker averaged across all rater pairs
Rater Mean Kappa	RMK	Kappa for a rater averaged across all markers
		Groupwise Ratings
Group Marker Kappa	GMK	Strength of agreement between groups on one marker
Group Rater Kappa	GRK	Strength of agreement between groups averaged across all markers

Table 6

Rater Mean Kappa for Bill, Detert, and Reid Cases

	Kappa		
Rater	Bill	Detert	Reid
1	.39*	.19	.14
2	.43**	.23*	.18
3	.38*	.21*	.13
4	.37*	.17	.13
5	.39*	.16	.15
6	.41**	.19	.14
7	.35*	.20	.18
8	.40*	.20	.14
9	.32*	.18	.14
10	.39*	.17	.14
11	.42**	.19	.12
12	.39*	.17	.15
13	.42**	.20	.14
14	.39*	.20	.19

Note: ****Almost perfect agreement, ***Substantial agreement, **Moderate agreement, *Fair agreement, according to Landis & Koch (1977). Unless otherwise specified, each coefficient has a slight level of agreement.

Rater Mean Kappas were calculated for each marker by 1) calculating a kappa for each pair of raters across one set of excerpts, and 2) averaging across all 26 markers.

Table 7

Marker Mean Kappa (and Frequency) for Bill, Detert, and Reid Cases

	Bill Case	Detert Cases	Reid Case
	(59 excerpts)	(82 excerpts)	(106 excerpts)
Marker	Kappa (Freq)	Kappa (Freq)	Kappa (Freq)
1. Body Symptoms	.52 (1.14)**	.46 (2.64)**	.16 (4.00)
2. Downplaying Negativity	.12 (1.07)	.05 (0.93)	.09 (3.00)
3. Avoiding Responsibility	.20 (1.86)	.04 (2.29)	.03 (1.79)
4. Distancing Language	.76 (3.00)***	.07 (2.29)	.12 (3.64)
5. Feeling Surprise	.54 (1.93)**	.07 (0.64)	.10 (1.57)
6. Fearing Loss	.00 (.00)	.07 (1.14)	.00 (0.79)
7. Desiring Change	.64 (4.86)***	.19 (7.43)	.19 (3.79)
8. Feeling Stuck	.83 (5.14)****	.51 (5.36)**	.15 (2.86)
9. Feeling Pain	.56 (6.14)**	.13 (4.21)	.19 (8.86)
10. Feeling Confused	.19 (1.79)	.28 (2.86)*	.13 (4.93)
11. Difficulty Articulating	.35 (1.57)*	.50 (1.57)**	.15 (0.93)
12. Feeling Vulnerable	.34 (5.07)**	.14 (8.36)	.19 (6.93)
13. Unfinished Business	.15 (0.93)	.06 (4.50)	.09 (5.64)
14. Recurring Problem	.46 (2.43)**	.26 (9.86)*	.22 (6.00)*
15. Expressing/Inhibiting	.51 (1.43)**	.05 (3.21)	.07 (3.71)
16. Incompatible Goals	.04 (0.79)	.00 (2.00)	.19 (4.07)
17. Wants and Shoulds	.53 (2.71)**	.07 (3.29)	.18 (6.86)
18. Other's Values	.34 (1.71)*	.19 (2.58)	.22 (2.86)*
19. Old Reactions	.85 (2.93)****	.37 (3.64)*	.07 (2.64)
20. Stepping Back	.15 (2.43)	.07 (3.86)	.08 (5.43)
21. Putting Pieces Together	.27 (2.86)*	.15 (5.14)	.25 (3.79)*
22. Almost, But Not Quite	.12 (2.43)	.05 (4.79)	.03 (3.29)
23. Deciding to Act Diff.	.07 (0.79)	.32 (6.57)*	.11 (2.64)
24. Noticing Change	.62 (6.64)***	.32 (6.21)*	.44 (8.36)**
25. Asserting Needs	.04 (0.21)	.25 (5.71)*	.31 (7.36)*
26. Coming to Solution	.07 (1.43)	.11 (1.36)	.03 (1.71)

Note: ****Almost perfect agreement, ***Substantial agreement, **Moderate agreement, *Fair agreement, according to Landis & Koch (1977). Unless otherwise specified, each coefficient has a slight level of agreement.

The frequencies listed next to the kappas are the average frequency that raters endorsed markers in a particular data set. Marker Mean Kappas were calculated for each rater pair across one set of excerpts, and then the kappas for each marker were averaged across raters.

Table 8

Group Marker Kappa (and Frequency) for Bill, Detert, and Reid Cases

	Bill Case	Detert Cases	Reid Case
	(59 excerpts)	(82 excerpts)	(106 excerpts)
Marker	Kappa (Freq)	Kappa (Freq)	Kappa (Freq)
1. Body Symptoms	1.00 (1.00)****	.85 (3.50)****	.56 (3.50)**
2. Downplaying Negativity	-.02 (1.00)	.00 (0.50)	.00 (1.50)
3. Avoiding Responsibility	.48 (2.00)**	.00 (0.50)	.00 (0.50)
4. Distancing Language	1.00 (3.00)****	.00 (1.00)	.80 (2.50)***
5. Feeling Surprise	.79 (2.50)***	.00 (0.50)	.00 (0.50)
6. Fearing Loss	-- (--)	.00 (0.50)	-- (--)
7. Desiring Change	1.00 (5.00)****	.25 (6.50)*	.49 (2.00)**
8. Feeling Stuck	1.00 (5.00)****	.65 (4.50)***	1.00 (1.00)****
9. Feeling Pain	.90 (5.50)****	-.03 (3.00)	.22 (7.50)*
10. Feeling Confused	.49 (2.00)**	1.00 (1.00)****	.32 (3.00)*
11. Difficulty Articulating	-.02 (1.50)	1.00 (1.00)****	1.00 (1.00)****
12. Feeling Vulnerable	.73 (4.00)***	.28 (6.00)*	.42 (4.50)**
13. Unfinished Business	.00 (0.50)	.39 (2.50)*	.23 (4.00)*
14. Recurring Problem	1.00 (2.00)****	.55 (10.00)**	.23 (4.00)*
15. Expressing/Inhibiting	1.00 (1.00)****	.00 (0.50)	.00 (1.00)
16. Incompatible Goals	.00 (0.50)	-- (--)	.66 (3.00)***
17. Wants and Shoulds	1.00 (2.00)****	.00 (1.00)	.34 (5.50)*
18. Other's Values	.79 (2.50)***	.66 (3.00)***	.49 (2.00)**
19. Old Reactions	1.00 (3.00)**	1.00 (1.00)****	.00 (1.00)
20. Stepping Back	.00 (1.50)	.00 (1.50)	.49 (2.00)**
21. Putting Pieces Together	.79 (2.50)***	.31 (3.00)*	.56 (3.50)**
22. Almost, But Not Quite	-.04 (2.00)	-.01 (1.00)	.00 (0.50)
23. Deciding to Act Diff,	-- (--)	.55 (3.50)**	.00 (0.50)
24. Noticing Change	.68 (7.00)***	.75 (6.50)***	.76 (9.00)***
25. Asserting Needs	-- (--)	.52 (5.50)**	.73 (8.00)***
26. Coming to Solution	.00 (0.50)	1.00 (3.50)****	-- (--)

Note: ****Almost perfect agreement, ***Substantial agreement, **Moderate agreement, *Fair agreement, according to Landis & Koch (1977). Unless otherwise specified or negative, each coefficient has a slight level of agreement. Blank kappas indicate that no representative marker was selected by a group. Raters 1-7 comprised Group 1 and raters 8-14 comprised Group 2.

The frequencies listed next to the kappas are the average frequency that group raters endorsed markers in a particular data set. Group Marker Kappas were calculated the same way as for single raters, except the calculations used group sets of markers, and then the kappas were averaged across all raters for each marker.

Table 9

Observed Confusion Values, Expected Confusion Values, and Confusion Ratios by Marker.

a. Observed Confusion Values

		Rater 1			
		Marker 1	Marker 2	Marker 3	Row Totals
Rater 2	Marker 1	5 (Cell 1)	0	2	7
	Marker 2	6	0	0	6
	Marker 3	0	0	4	4
	Column Total	11	0	6	17

b. Expected Confusion Values

		Rater 1		
		Marker 1	Marker 2	Marker 3
Rater 2	Marker 1	4.53 (Cell 1)	0	2.47
	Marker 2	3.88	0	2.12
	Marker 3	2.59	0	1.41

c. Confusion Ratios

		Rater 1		
		Marker 1	Marker 2	Marker 3
Rater 2	Marker 1	1.10	-	.81
	Marker 2	1.55	-	0
	Marker 3	0	-	2.84

Note: The values in the observed confusion values above are frequencies of excerpts.

Table 10

Confusion Ratios Greater Than 2.0

	Marker A	Marker B	Ratio
Cluster 1			
	5	10	2.74
	5	11	3.29
	5	21	2.21
	10	11	4.31
Cluster 2			
	15	16	2.34
	15	17	2.24
	16	17	3.09
Cluster 3			
	22	23	2.22
	22	24	2.17
	22	25	2.21
	22	26	3.03
	23	26	2.26
	24	26	2.63
	25	26	2.36

Note: Three clusters were observed (Markers 5, 10, 11, and 21; Markers 15, 16, and 17; and Markers 22, 23, 24, 25, and 26). The confusion ratio is the extent to which raters confused markers with each other.

Table 11

Frequencies of Excerpts Where Detert's Mean APES Ratings and Raters' APES Ratings Agreed and Disagreed

Raters'		Detert's Mean APES Stages								
Marker	Stage	0	1	2	3	4	5	6	7	%
1	0	0	0	4	2	0	0	0	0	0.0
2	1	0	0	2	5	1	1	0	0	0.0
3	1	0	2	9	5	3	1	0	0	10.0
4	1	0	0	4	5	5	0	0	0	0.0
5	1	0	0	2	2	1	0	0	0	0.0
6	1	0	0	5	1	0	0	0	0	0.0
7	2	0	3	9	23	3	1	1	0	22.5
8	2	0	1	10*	6	1	0	0	0	55.6
9	2	0	0	9	12	3	0	0	0	37.5
10	2	0	1	7	11	1	0	0	0	35.0
11	2	0	0	3*	3	0	0	0	0	50.0
12	2	0	1	12	21	5	0	0	0	30.8
13	2	0	1	9	15	4	0	0	0	31.0
14	2	0	1	15	21	5	1	1	0	34.1
15	3	0	1	8	14*	2	0	1	0	53.8
16	3	0	2	5	13*	1	0	0	0	61.0
17	3	0	2	3	17*	2	0	1	0	68.0
18	4	0	0	5	4	0	0	0	0	0.0
19	4	0	0	1	10	4	0	1	0	25.0
20	4	0	2	7	14	3	2	1	0	10.3
21	4	0	0	5	15	3	2	2	0	11.1
22	5	0	0	4	12	6	5	3	0	16.7
23	5	0	1	6	10	7	3	2	0	10.3
24	5	0	0	3	9	4	5	3	0	20.8
25	6	0	0	5	10	6	3	2	0	7.7
26	6	0	0	0	5	4	2	1	0	8.3

*Raters' APES ratings agree with Detert's Mean ratings ≥ 40%.

Note: Shaded boxes indicate frequencies of agreement between Detert's mean ratings and Raters' APES stages. The percentage was calculated by dividing the shaded frequency in a row by the sum of frequencies in that row.

Table 12

Frequencies of Excerpts Where Reid's Consensus APES Ratings and Raters' APES Ratings Agreed and Disagreed.

Raters'		Reid's Consensus APES Stages								
Marker	Stage	0	1	2	3	4	5	6	7	%
1	0	0	0	2	1	0	0	0	0	0.0
2	1	0	4*	2	3	0	0	0	0	44.4
3	1	0	3	3	2	0	0	0	0	37.5
4	1	0	1	6	2	0	0	0	0	11.1
5	1	0	2	2	2	0	0	0	0	33.3
6	1	0	0	0	4	0	0	0	0	0.0
7	2	0	0	5*	7	0	0	0	0	41.7
8	2	0	1	7*	5	0	0	0	0	53.8
9	2	0	2	6*	7	0	0	0	0	40.0
10	2	0	4	6	6	0	0	0	0	37.5
11	2	0	0	1	2	0	0	0	0	33.3
12	2	0	3	6*	5	0	0	0	0	42.9
13	2	0	5	7*	3	0	0	0	0	46.7
14	2	0	3	9*	1	0	0	0	0	69.2
15	3	0	3	9	5	0	0	0	0	29.4
16	3	0	1	3	6*	0	0	0	0	60.0
17	3	0	1	7	9*	0	0	0	0	52.9
18	4	0	2	1	2	0	0	0	0	0.0
19	4	0	1	6	1	0	0	0	0	0.0
20	4	0	6	8	4	0	0	0	0	0.0
21	4	0	3	6	4	0	0	0	0	0.0
22	5	0	1	2	8	0	0	0	0	0.0
23	5	0	0	0	11	0	0	0	0	0.0
24	5	0	0	1	10	0	0	0	0	0.0
25	6	0	0	2	11	0	0	0	0	0.0
26	6	0	1	0	5	0	0	0	0	0.0

*Raters' APES ratings agree with Reid's Consensus ratings ≥ 40%.

Note: Shaded boxes indicate frequencies of agreement between Reid's consensus ratings and Raters' APES stages. The percentage was calculated by dividing the shaded frequency in a row by the sum of frequencies in that row.