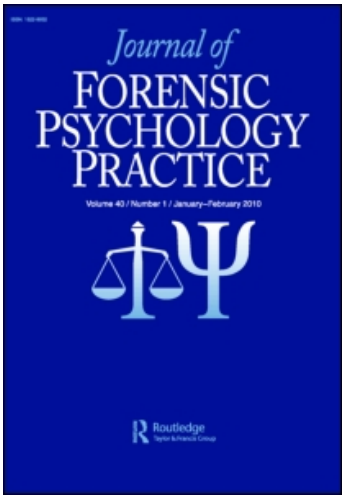

Journal of Forensic Psychology Practice
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t792304004

Formalizing Clinical Decisions in Individual Treatments: Some First Steps

Marinus Spreen[a]; Marieke E. Timmerman[b]; Paul Ter Horst[c]; Erwin Schuringa[d]
[a] Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands Institute for Applied Research, Stenden University, Leeuwarden, The Netherlands [b] Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands [c] Geestelijke GezondheidsZorg Eindhoven en de Kempen, Eindhoven, The Netherlands [d] Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands

Online publication date: 27 July 2010

To cite this Article Spreen, Marinus , Timmerman, Marieke E. , Horst, Paul Ter and Schuringa, Erwin(2010) 'Formalizing Clinical Decisions in Individual Treatments: Some First Steps', Journal of Forensic Psychology Practice, 10: 4, 285 — 299
To link to this Article: DOI: 10.1080/15228932.2010.481233
URL: http://dx.doi.org/10.1080/15228932.2010.481233

Journal of Forensic Psychology Practice, 10:285–299, 2010

ISSN: 1522-8932 print/1522-9092 online
DOI: 10.1080/15228932.2010.481233

Routledge
Taylor & Francis Group

Formalizing Clinical Decisions in Individual Treatments: Some First Steps

MARINUS SPREEN, PhD
Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands Institute for Applied Research, Stenden University, Leeuwarden, The Netherlands

MARIEKE E. TIMMERMAN, PhD
Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

PAUL TER HORST, PhD Candidate
Geestelijke GezondheidsZorg Eindhoven en de Kempen, Eindhoven, The Netherlands

ERWIN SCHURINGA, PhD Candidate
Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands

A fundamental problem in forensic psychology practice is the lack of formal statistical methods to support team decisions about an individual patient's progress during intramural treatment. It is common practice to base decisions about the progress of a treatment on subjective clinical impressions of therapists. In this article, an approach is proposed that can be seen as a contribution to bridge the gap between formal statistical decision making and subjective clinical decision making. To formalize decisions in individual treatments, we have elaborated a statistical decision technique based on degrees of belief. In this article, this so-called N = 1 *analysis is explained and illustrated by a hypothetical case.*

KEYWORDS decision making, treatment evaluation, formal methods, single case

Address correspondence to Marinus Spreen, Research Department, FPC dr S. van Mesdag, P. O. Box 30.002, 9700 RC Groningen, The Netherlands. E-mail: Marinus.Spreen@Stenden.com

A fundamental problem in forensic psychology practice is the lack of formal statistical methods to support decisions about an individual patient's progress during intramural treatment. In a forensic clinical treatment, two basic types of decisions can be discerned: decisions of diagnosis and decisions whether a patient has shown sufficient progress to enter a next phase of the therapy (Streiner & Norman, 2003). In this article, we focus on treatment progress decisions. Conventionally, the decision whether a certain patient has shown sufficient progress is primarily based on intrinsic arguments deduced from clinical impressions of a patient's past intramural behavior. Such clinical impressions can be highly subjective (i.e., depend on the personal view and comparative assessments of the therapist involved). Furthermore, individual therapists may also get used to (extreme) behaviours of patients, which may lead to underreporting or underestimating such behaviours. From the positivist paradigm, which has as a starting point the existence of a univocal objective reality, clinical impressions about one patient cannot be scientific. However, in this article, we argue that decisions about treatment progress of just one individual patient can be formalised applying conventional methodological standards.

Therefore, consider a quotation from Walgrave (2008):

> Methodology is meant to channel the researcher's intuitions and suspicions through a systematic and controllable procedure of thinking and data collection: a well-considered, open problem analysis based on the best available knowledge, and a step-by-step account of all moves in the process of constructing data and drawing conclusions. Results and views based on good scientific research are systematically investigated, contextualized, and controllable.

Replace in this quotation researcher by therapist and a proper scientific decision method about treatment progress of an individual patient has to meet a systematic and controllable data collection method on which valid conclusions can be drawn. In this article, we argue that decisions based on subjective clinical impressions do not necessarily imply poor decisions. In general, an experienced therapist will have sufficient skills and insights to accurately establish current levels of functioning of the patient by employing clear and specific constructs. From a methodological perspective, a decision based on clinical impressions is debatable whenever the decision rules by therapists are not explicit, controllable, and repeatable and/or the specific constructs on which the decision is based are not univocal.

To support clinical decisions, forensic therapists can apply measurement instruments. The purpose of such instruments is to measure some relevant construct that is thought to be important for the decision. A construct represents something that is believed to exist although, strictly speaking, whatever it is that one is referring to can never be directly observed (Sheskin, 2004).

Examples of such constructs are anxiety, psychopathy, social skills, impulsiveness, and so on. Constructs can be measured with self-report tests, physiological measures, or peer or expert ratings or through direct observation of specific target behaviors that are considered critical indicators for the presence of that construct.

In this article, we focus on the role of observational data to evaluate the treatment of one single patient in a forensic intramural setting. In general, observational information by therapists is assumed to be more reliable than self-reported information by forensic patients because of social desirability problems. However, in daily practice of a forensic psychiatric hospital, the quality of observational data is often restricted by the amount of available time therapists allow themselves to complete some instrument. Furthermore, the information that is available to the therapist is often imprecise, incomplete, or not totally reliable (Zadeh, 1984). In this article, we propose a decision method that tries to reckon with standard methodological rules and the practical limitations of daily practice in a forensic psychiatric center. The basic design to evaluate an individual treatment contains the following steps:

(1) Decide which hypothesis and interventions will be applied and evaluated on the patient.
(2) Define the construct(s) to be measured and select a measurement instrument.
(3) Before the actual intervention period, the measurement instruments are, independent of one another, completed by a group of therapists (baseline).
(4) Based on the results of Step 3, interventions can be adjusted and conducted.
(5) At the end of the intervention period, the observation list is again administered and completed by a group of therapists independent of one another.
(6) Based on the results, it is decided whether the intervention can be evaluated as sufficiently successful.

This design has been elaborated from daily treatment procedures of the Forensic Psychiatric Centre (FPC) Dr. S. van Mesdag in The Netherlands. We wish to emphasize that the decision method has been elaborated in a pilot study; this article presents a first step toward a formally based evaluation. Currently, the decision method is being examined in a large-scale study. In this article, we illustrate and explain the decision method using a hypothetical example that can be understood as representative for the daily practice of treatment evaluation in the FPC Dr. S. Van Mesdag. The primary goal of the decision method is to define some formal procedures and decision rules for evaluating individual treatment progress. First, some

theoretical and practical considerations will be discussed. Subsequently, the statistical method is illustrated with a hypothetical case study. Finally, some remarks and comments are given.

THEORETICAL AND PRACTICAL CONSIDERATIONS

Suppose the treatment of patient *Hypothetical* must be evaluated. A reason for this can be routine outcome assessment or the evaluation of some specific intervention. Let us further assume that patient *Hypothetical* has some serious impulsiveness problem and, based on this diagnosis, it is decided to give him some specific treatment. This treatment can be behavioral and/or medically oriented. In methodological terminology, the impulsiveness of patient *Hypothetical* is the dependent variable, and the time (respectively before and after the proposed intervention) is the independent variable. According to Step 1 of our design, we want to test the hypothesis whether and to what extent the impulsiveness of patient *Hypothetical* has been reduced after the intervention.

Once the therapist has decided which hypothesis to test, he or she now has to decide how to measure the effect of the intervention (i.e., Step 2 of the design). How to measure the construct, here impulsiveness, is not only a theoretical but a practical question. The therapist could use some standard instrument that has been properly evaluated. Such instruments usually consist of a series of items that are observable manifestations of the construct of interest (Nunally & Bernstein, 1994). For a therapist, a standard instrument is not always the best choice: A proper instrument may not be available to measure the construct of interest, or the instrument may be unsuitable to use for the specific patient (group) concerned. Furthermore, standard instruments tend to consist of many items per measured construct to yield sufficient measurement precision. In daily practice of a forensic psychological treatment, therapists are often reluctant to complete long lists of observational questions.

We put forward the idea that a more direct measurement of the construct of interest with a single item would yield a sufficient measurement precision for a sufficiently reliable evaluation of the intervention. This idea contrasts to conventions in item response–theory-based measurement instruments (e.g., Embretson & Reise, 2000), which use a set of items per construct. The idea is that the use of more items increases the measurement precision of the construct of interest. However, this is true only when the items are indicative of the very same construct. Because this is difficult to achieve in practice, some instruments include a series of arbitrary differing items. This increases the reliability coefficients of the instrument. However, it is questionable whether this adds to the discriminative power of the instrument in practice.

Alternatively, a more direct measurement may yield sufficiently reliable and valid scorings. We propose to use for each construct of interest an indicator variable; for the situation of patient *Hypothetical*, this implies simply asking therapists to grade the impulsiveness on, for instance, a five- or more-point scale. To apply such a direct approach (i.e., rating a complex concept by one simple indicator), it is of utmost importance to have clear univocal definitions of both the concept and the answer categories.

The first advantage of this direct investigation is that the relationship between the items and the construct of interest is very clear, which advances a valid measurement. Herewith, the underlying assumption is that therapists have sufficient agreement on the meaning of the construct of interest. This is not unreasonable because therapists have common expert knowledge. Furthermore, the load on the therapist(s) is considerably reduced, because fewer items have to be scored. This increases the willingness to use the evaluation approach and reduces unreliability resulting from motivational problems. The gain in efficiency becomes very important in the frequently occurring situations wherein the progress of some treatment is evaluated on the basis of many constructs.

The outcome of an indicator variable may be understood as a structured professional assessment of the value of the construct. The therapist checks in a standard way a set of issues and weighs them to determine the value of the indicator that best suits the behavior of a patient, observed by the therapist him- or herself. In other words, the outcome of such a structured professional assessment can be viewed as some grading of the construct. Viewing indicators this way is deduced from fuzzy logic reasoning (see Zadeh, 1968,1984, 2006; Mahmoud Taheri, 2003); a detailed discussion of this approach is beyond the scope of this article.

To elaborate a practical decision method using indicators as structured professional assessments, some quality criteria have to be formulated for the validity of the construct. Three basic criteria are of most importance:

(1) The team of therapists who will evaluate the patient must agree with one another on the nature and contents of the construct and the way the indicator is scored.
(2) The indicator must be filled in by at least two persons.
(3) Two therapists with exactly the same observational information about the behavior of the patient must assign independently of each other the same value on the indicator to that behaviour.

As discussed, therapists could use some standard instrument that has been properly evaluated; however, we remark that this is convenient but not necessary. In situations wherein a construct is very patient-specific (e.g., behavior at dinner), a team of therapists can gain consensus on the

formulation of this specific construct. Such a construct can be valid only if Criterion 3 is met. It is obvious that Criterion 3 refers to the degree of agreement: The larger the degree of agreement between a team of therapists, the more evidence that the construct is measured reliable.

Imagine that five therapists have to evaluate the intervention on patient *Hypothetical*. For the sake of argument, the team of five therapists is asked to focus only on the construct "impulsiveness." To measure the construct, they may use a standard test. However, they can also decide to define specific personal behaviors that characterize the impulsiveness of patient *Hypothetical*. It is obvious that those specific behaviors are supposed to be influenced by the proposed intervention. Let us further assume that the team of therapists has reached a consensus and defined an indicator and clear instructions how to weigh different behaviors to graduate "impulsiveness" for the behaviors of patient *Hypothetical*. Let us assume that they value the concept using indicator H on a five-point Likert scale (i.e., a discrete scale of 0–4, where answer category 0 implies the absence of impulsiveness and answer category 4 extreme impulsiveness).

Step 3 of the proposed design is that all five therapists value independent of one another the "impulsiveness" of patient *Hypothetical* based on their individual observations. Obviously, the observation time span should be restricted to some predefined period. To express the degree of agreement between n therapists on indicator H, we employ the following index described by Gower & Legendre (1986) (hereafter GL-index),

$$\frac{1}{\binom{n}{2}} \sum_{(i,j)\in T} 1 - \frac{|H_i - H_j|}{rangeH}, \tag{1}$$

where $T = \{i, j, \cdots, n\}$ is the collection of therapists, n the total number of therapists, and H_i the discrete value of therapist i.

GL- index (1) can be understood as follows: For every possible pair of therapists, their absolute difference in values is related to the maximum possible absolute difference (*range H*). If the difference within pair (i, j) is at a maximum, the index will be 0; if the difference is in the middle of the absolute maximum difference, the index tends to 0.50; and if there is no difference between two therapists, the index is 1. Thus, the index can be viewed as the average agreement between two arbitrary therapists of the team. To find some formal rule to explore which level of agreement can be considered as meaningful on a five-point Likert scale, we examined the GL-index across all possible combinations of values that therapists may give. We computed for groups of three to eight therapists all possible combinations of scores and computed per group the GL- index. In this way, the frequency distribution is obtained of the GL-index per specific number of therapists. For each number of therapists, we considered the point GL-index score at 90%

TABLE 1 The Gower-Legendre Index at Significance Level 0.05 and 0.10 for Groups of Three to Eight Therapists

Significance level	Three therapists $n = 125$	Four raters $n = 625$	Five raters $n = 3{,}125$	Sic raters $n = 1{,}5625$	Seven raters $n = 78{,}125$	Eight raters $n = 390, 625$
0.05	0.83	0.83	0.80	0.78	0.76	0.74
0.10	Not defined	0.75	0.75	0.73	0.71	0.70

and 95% of the frequency distribution. The interpretation of this point GL-index score at 90% and 95% is that there are 10% or 5% other combinations, which results in a higher agreement. In Table 1, the point GL-index score at 90% and 95% are displayed for groups of three to eight therapists.

To illustrate Table 1, consider the group of five therapists. A group of five therapists can generate 3,125 possible combinations of scores. Of those 3,125 combinations, 10% have a GL-index of 0.75 or higher and 5% a GL-index of 0.80 or higher. In other words, if a team of five therapists have a GL-index of 0.75 on some indicator, they could have chosen only 10% other combinations, which results in more agreement.

To return to the case of patient *Hypothetical*, Table 2 displays the discrete values the five therapists have assigned to indicator variable "Impulsiveness" (column Baseline Measurement). The agreement between the five therapists at the Baseline Measurement is about 0.73.

Summarizing the first three steps of the proposed design, we may say that the construct "impulsiveness" special modeled to the evaluation situation by indicator *H* of patient *Hypothetical* seems rather reliably scored: There is an intrinsic agreement among the five therapists concerning the construct resulting in indicator *H*, and the empirical agreement on this indicator was about 0.73.

To formalize, some rule has to be implemented to decide whether a construct is measured reliably. The intrinsic agreement about the "score" on the construct for the patient involved can be determined in a qualitative way. However, if all members of a team agree with the formulations, preferably

TABLE 2 Scores of Five Therapists for Patient *Hypothetical* on Indicator "Impulsiveness"

Therapists	Baseline measurement	Repeated measurement
Therapist *i*	3	1
Therapist *j*	1	0
Therapist *k*	1	0
Therapist *l*	1	2
Therapist *m*	2	2
Sum score	8	5

backed up by evidence-based information, there is less reason to suspect that the construct will be of poor reliability. The actual agreement of the rating on the indicator can be quantified. In the FPC Dr. S. van Mesdag, indicators are considered sufficiently reliable scored if index (1) is higher than 0.70 (see Table 1). Note that this value is arbitrary to a certain extent. Furthermore, it is essential to recognize the importance of the contextual situation to the interpretation of the results. This implies that a low agreement does not necessarily imply a poor validity of the indicator, because specific contextual factors may account for such a low agreement. For instance, a patient's social behavior at the work place in a clinic can be rather different than his behavior in the living group. Hence, whenever low agreement among all or part of the therapists is being observed, one should carefully evaluate the source of this low agreement. Assessing the story behind the scores may yield a more complete picture of the patient concerned. Such observations are very informative for treatment evaluation. However, note that two therapists who both, for example, work on the living group of the patient should have a high agreement.

Based on the scores and other information of the baseline measurement, interventions may be adjusted before the actual intervention period starts. Step 5 of the design is at the end of the intervention period. Five therapists (preferably, but not necessarily, the same therapists as before) again value, independent of one another, indicator "Impulsiveness" of patient *Hypothetical*. These values are listed in the Repeated Measurement column in Table 2. According to GL-index (1), the group of five therapists have now an agreement of 0.70, which is satisfactory. Thus, both measurements can be conceived as reliable under the defined rule.

The next question in our design is whether we are allowed to decide that the impulsiveness of patient *Hypothetical* has significantly reduced (Step 6). To this end, a statistical decision method will be proposed in the next section.

Forensic $N = 1$ Decision Theory

The proposed statistical decision method has been elaborated, employed, and studied in the FPC Dr. S. van Mesdag, Groningen, The Netherlands. The method departs from the assumption that therapists are professionals and that the value they give to an indicator can be regarded as their subjective degree of belief based on all sorts of evidence (Van Lente, 1993).

Considering a discrete value a therapist has given to some indicator, we make the following assumptions:

(1) The assigned discrete answer category (i.e., raw score) is the mode of the subjective degree of belief (SDB) distribution of the therapist and

reflects his or her best educated guess for the observed behavior of a patient.
(2) The dispersion of the SDB corresponds to the degree of uncertainty about the raw score of the therapist.
(3) The further away a discrete answer category from the raw score, the less likely the observation of the considered behavior.

We could ask therapists to give their values on the SDB by posing questions such as "If you consider the behavior of patient *Hypothetical* the last 6 months, what is your score on indicator Impulsiveness and what is your degree of belief in terms of percentage that this value describes the behavior; can you also give your percentages of belief to the other answer categories?" However, from our experience, it is very impractical to obtain the SDBs of therapists. The posed questions are very time-consuming for therapists and therefore, in most forensic psychological practice settings, rather impractical. Forcing therapists to provide their SDBs appears to be a bad idea, because unmotivated and unsympathetic responding results in invalid responses. With this in mind, we focus in this article on the situation that therapists give only one value on the indicator value. Thus, for patient *Hypothetical*, we have for the basement measurement five values (see Table 2).

To approximate the unknown SDB of a therapist, some plausible SDB must be defined. For the sake of explanation, consider the value of therapist *i* at the baseline measurement. Therapist *i* has valued the "impulsiveness" of patient *Hypothetical* with discrete score 3. His subjective belief is that this score reflects the behavior of patient *Hypothetical* best and is thus most likely in relation to the sort of evidence he has considered. The raw score 3 can be understood as the mode of his unknown SDB. To compute the SDB of therapist *i*, we first have to decide which percentage of degree of belief we accept as plausible for the mode of the SDB; second, we have to decide how the remaining scoring categories are dispersed. The top of the SDB is related to the quality of the indicator. If we are confident that our indicator is univocally defined (and maybe from a well-validated list), we may feel sure to assume that the degree of belief on the validity of the mode may be close to 95% for each therapist. However, when we are less sure, we may allow more uncertainty on the mode by choosing a lower percentage of the mode value. In our hypothetical example, we will choose a percentage of 80% of the mode value for each therapist.

Our next consideration is to disperse the remaining percentage of degrees of belief (in our example, 20%) over the other answer categories. The answer categories closest to the mode are symmetrically weighed by the remaining numbers of answer categories. Because of Assumption 3, the further away from the raw score, the less likely the observation of the considered behavior is. For therapist *i* with raw score 3, we assigned as degrees of belief for values 0 to 4 are, respectively, *.2, .4, .7, .80,* and *.7*.

These *degrees of belief* can be interpreted as follows: If therapist i would have repeatedly valued 100 times the same patient at the same time for the situation that he had 80% belief on the top of his mode, we would expect for this particular situation that he would have chosen *7* times discrete category 4, *80* times discrete category 3, *7* times discrete category 2, *4* times discrete category 1, and *2* times discrete category 0. Note that these percentages are dependent of the applied assumptions.

To combine the scores of the various therapists, we propose to compute the joint SDB across therapists. Herewith, we have to assume that all therapists rated the patient independent of one another. The resulting joint SDB expresses the subjective percentages of belief in the various scores across all therapists.

Let $i_c^{(m)}$ denote discrete answer category c given by therapist i with mode m and $P_{(i_c)}^{(m)}$ denote its subjective degree of belief (expressed in a proportion). Thus, for the specified situation of a 80% belief on the mode, therapist i who has valued the indicator with a 3 (his mode) has subjective degrees of belief of $P_{(i_0)}^{(3)} = 0.2$; $P_{(i_1)}^{(3)} = 0.04$; $P_{(i_2)}^{(3)} = 0.07$; $P_{(i_3)}^{(3)} = 0.80$; $P_{(i_4)}^{(3)} = 0.07$.

The joint SDB of the group of n therapists for the sum score of a certain combination is computed by

$$P_{(i_c)}^{(m)} \times P_{(j_c)}^{(m)} \times \ldots \times P_{(n_c)}^{(m)} \tag{2}$$

for $\{i, j, \ldots, n\}$ therapists.

Employing equation (2) implies that all joint subjective percentages of belief between n therapists for each possible composed sum score are calculated. Let us return to patient *Hypothetical*, who is valued by $n = 5$ therapists on a scale of 0 to 4 (five discrete answer categories).

If we take into account the values of the five therapists of the baseline measurement (see Table 1), the joint SDB of their composed sum score *1* is computed by

$$\begin{aligned}
&\left[P_{(i_1)}^{(1)} \times P_{(j_0)}^{(1)} \times P_{(k_0)}^{(1)} \times P_{(l_0)}^{(3)} \times P_{(m_0)}^{(2)}\right] + \left[P_{(i_0)}^{(1)} \times P_{(j_1)}^{(1)} \times P_{(k_0)}^{(1)} \times P_{(l_0)}^{(3)} \times P_{(m_0)}^{(2)}\right] + \\
&\left[P_{(i_0)}^{(1)} \times P_{(j_0)}^{(1)} \times P_{(k_1)}^{(1)} \times P_{(l_0)}^{(3)} \times P_{(m_0)}^{(2)}\right] + \left[P_{(i_0)}^{(1)} \times P_{(j_0)}^{(1)} \times P_{(k_0)}^{(1)} \times P_{(l_1)}^{(3)} \times P_{(m_0)}^{(2)}\right] + \\
&\left[P_{(i_0)}^{(1)} \times P_{(j_0)}^{(1)} \times P_{(k_0)}^{(1)} \times P_{(l_0)}^{(3)} \times P_{(m_1)}^{(2)}\right].
\end{aligned}$$

Thus, the joint SDB of each sum score is dependent on the composition of the discrete categories and the modes of the five therapists. The joint SDB of the five therapists who have been rating patient *Hypothetical* is shown in Table 3.

TABLE 3 Joint SPB of Composed Sum Scores of the Five Therapists Evaluating Patient *Hypothetical*

Composed sum score	Subjective degree of belief at baseline measurement	Subjective degree of belief at repeated measurement
0	0.00004	0.00223
1	0.00032	0.00902
2	0.00178	0.02934
3	0.00695	0.09762
4	0.02097	0.19143
5	0.05166	0.40084
6	0.09024	0.13363
7	0.18354	0.06894
8	0.35142	0.03607
9	0.13964	0.01780
10	0.07944	0.00715
11	0.04440	0.00333
12	0.01671	0.00151
13	0.00764	0.00065
14	0.00333	0.00026
15	0.00118	0.00011
16	0.00047	0.00004
17	0.00018	0.00002
18	0.00006	0.00001
19	0.00002	0.00000
20	0.00000	0.00000

Table 3 can be interpreted as follows. The joint SDB that those five therapists—who have valued the "Impulsiveness" of patient *Hypothetical* with mode scores (1, 1, 1, 3, 2) at the baseline measurement—would obtain together a sum score of 0 is 0.00004, a sum score of 1 is 0.00032, and so on. It seems very unlikely that those five therapists together would obtain a sum score of 1 in the hypothetical situation that they were allowed infinitely to value independently of one another the "Impulsiveness" at the same moment. Note that the subjective degrees of belief around the modes are understood as uncertainties.

The Baseline Measurement column in Table 3 can be conceived as being related to the joint SDB of the H_0-hypothesis: the statement of no effect or no difference. This implies that if there is no change in the "Impulsiveness" of patient *Hypothetical*, the distribution of the baseline measurement is also most likely at the repeated measurement. At the repeated measurement, the five therapists had a composed observed sum score of 5, implying that after the intervention period, the impulsiveness of patient *Hypothetical* has decreased from 8 to 5. The percentage of degree of belief to find under the H_0 assumption a value ≤ 5 is *0.08*. In other words, a value 5 or lower is rather uncharacteristic for the baseline measurement. We may define a formal decision rule whether a value can be considered as characteristic for a distribution by setting $\alpha = 0.1$, which is to a certain

extent an arbitrary choice. If we adopt this rule, we may say that the team of therapists has observed a meaningful change.

The statistical decision method refers only to the degree of belief of a joint value. Therapists are not statisticians; they have to apply and interpret these outcomes of analysis in their daily practice. From our empirical work in two psychiatric centers in The Netherlands (Groningen and Eindhoven), we noted that therapists are more convenient with percentages of change. Therefore, the results are also reported in two types of percentages to the therapists. The first is very simple and informative: dividing the difference in composed sum scores between both measurement moments by the maximum composed sum score, the percentage of change is computed. For patient *Hypothetical*, his change is shown in Table 4.

In Table 4, a positive development is observed. We may say that patient *Hypothetical* has decreased 15% on the indicator "Impulsiveness" (e.g., his observed "impulsiveness" is reduced by 15%). From our statistical $N = 1$ analyses, we already knew that we may consider this a meaningful change.

The second way of using percentages of change in an $N = 1$ analysis is somewhat more complex. Recall that the composed sum score of patient *Hypothetical* was 8 and that the intervention aims at reducing the impulsiveness (i.e., a one-sided hypothesis). In Figure 1, both measured joint SDBs of patient *Hypothetical* are graphically displayed.

TABLE 4 Percentage of Change

Composed sum moment 1/max score	8/20	40%
Composed sum moment 2/max score	5/20	25%
Change		15%

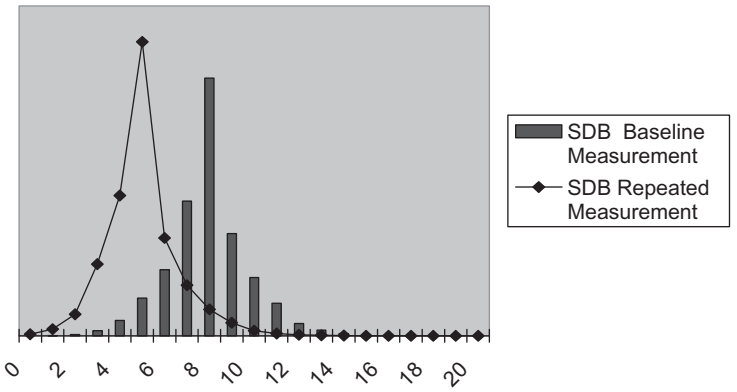

FIGURE 1 Subjective Degree of Belief of Patient *Hypothetical*

In Figure 1, the bars represent the joint subjective degrees of belief of the five therapists at the baseline measurement. The most likely sum score for "impulsiveness" at the baseline for those five therapists is *8*. The intervention aims at reducing the impulsiveness of patient *Hypothetical*, so the area of values ≤ 8 can be defined as the "change area." At the baseline measurement, this change area has a SDB percentage of *0.71*. Suppose that after the intervention period all therapists value the indicator variable with a *0* (i.e., all therapists evaluate that the "Impulsiveness" of patient *Hypothetical* is absent). This situation refers to a maximal change of patient *Hypothetical*. In this maximal change situation, the area of a value ≤ 8 has an SDB of *0.999*. However, based on the SDB of the baseline measurement already, *0.71* was predicted. Subtracting both densities of area ≤ 8 (i.e., *0.99–0.71 = 0.28*) can be understood as the maximal possible change patient *Hypothetical* could make controlled for the SDB of the baseline measurement.

At the repeated measurement, the sum score of the mode values of the five therapists reduced to *5*. Based on this value, the joint SDB of the five therapists is computed (the line in Figure 1). The subjective probability of a value ≤ 8 at the repeated measurement is *0.97* implying that a score in the change area has become more likely. A part of the degree of belief percentage of the change area at the repeated measurement was already "predicted" by the baseline measurement: Thus, the unique change is $0.97 - 0.71 = 0.26$. We already knew that the maximal possible change based on the baseline measurement was *0.28*. We may say that patient *Hypothetical* has reached 93% ($0.26/0.28 = 0.93$) of the maximum possible change on indicator "Impulsiveness" according to the five therapists.

Summarizing the Case of "Patient Hypothetical"

First, there was substantive consensus on the measurement of the indicator "Impulsiveness." Second, the agreement of the five therapists was at both measurement moments (0.73 and 0.70) satisfactorily. Third, the composed sum score of the five therapists after the intervention was evaluated as not characteristic for the situation that nothing had been changed. As a conclusion, one may decide that patient *Hypothetical* has meaningfully reduced his impulsiveness and that this reduction was 15% in total score and 93% of the maximum possible change. However, the question whether the intervention really caused the reduction remains troublesome to answer. In practice, one cannot isolate one intervention. Consequently, it is almost impossible to unravel which interventions count for which part of the effect. To cope with such problems, one could measure a "period as usual" before actually starting with the intervention, but that is beyond this article.

CONCLUDING REMARKS

In this article, an approach for treatment evaluation of one patient is introduced. The purpose of this approach can be described as follows: trying to meet methodological standards of psychological research and being as practicable as possible (minimal investments for therapists). Due to time restrictions in forensic psychiatric practice, decisions whether a patient has shown some progress are often made on subjective grounds based on ad hoc and uncontrollable procedures. Sometimes those decisions are partly supported by epidemiological norm-data from valid measurements instruments. The proposed approach has been elaborated from the perspective that therapists are professionals. By considering indicator variables as structured professional judgments and assigning degrees of belief to the outcomes of the indicator variables, it is possible to meet clinically efficient criteria. Using indicators can be more efficient because it uses knowledge and experience of the therapists and may motivate therapists to evaluate their therapies in a formal way.

An unexpected, but warmly welcome, spin-off we signaled when implementing this specific $N = 1$ approach in the FPC dr. S. van Mesdag is the "scientification" of the treatments. Because therapists had to disclose their expectations on paper about the effect of the proposed interventions, they were forced to think about the treatments and the state of knowledge of these interventions. Discussing reasons why other therapists observed other values on the constructs seemed to "scientificate" the staff of the psychiatric hospital in a natural way.

The proposed method must be viewed as a first step in trying to bridge the gap between subjective and formal forensic clinical decisions. We proposed a design wherein other assumptions will lead to other plausible SDBs. The validity of the assumptions can be different for different contexts. In practice, one has to strive to direct measurements of the SDBs by training therapists to observe in degrees of belief. However, regardless of the discussion about the plausibility of the assumptions to compute SDBs, in forensic practice each clinical decision in individual treatments must meet formal criteria: a systematic and controllable data collection method, precise operational definition of the construct, and decision rules.

REFERENCES

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Biometrica*, *27*, 5–48.

Mahmoud Taheri, S. (2003). Trends in fuzzy statistics. *Austrian Journal of Statistics*, *32*(3), 239–257.

Nunnaly, J. C., & Bernstein, I. H. (1994).*Psychometric theory*. (3rd ed.). New York: McGraw-Hill.

Sheskin, D. J. (2004) *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales. A practical guide to their development and use* (3rd ed.). Norfolk: Oxford University Press.

Van Lente, J. (1993) *The use of proper scoring rules for eliciting subjective probability distributions*. Leiden: DSWO Press.

Walgrave, L. (2008). Criminology. As I see it ideally. *Newsletter of the European Society of Criminology*, *7*(3), 15–17.

Zadeh, L. A. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, *23*(2), 421–472.

Zadeh, L. A. (1984). Fuzzy probabilities. *Information Processing & Management 20*(3), 363–372.

Zadeh, L. A. (2006). Generalized theory of uncertainty (GTU)—principal concepts and ideas. *Computational Statistics & Data Analysis 51*, 15–46.