Qual Quant (2014) 48:2225–2240
DOI 10.1007/s11135-013-9888-3

A fuzzy set theory based computational model to represent the quality of inter-rater agreement

Enrico Ciavolino · Sergio Salvatore · Antonio Calcagnì

Published online: 6 July 2013

Abstract In this paper we present a method to evaluate the quality of a rater's judgement, which can integrate and enrich the use of inter-rater agreement as a reliability measure. Our proposal is an integrative one and evaluates the quality of a rater's performance through an analysis of the profile of that individual rater's performance. We discuss its rationale on the basis of the interpretation of inter-rater agreement, highlighting some critical issues. For this purpose, we adopt a computational model based on fuzzy set theory, demonstrating its main characteristics with an exemplary case study.

Keywords Inter-rater agreement · Fuzzy set theory based models · Performance measurement · Quality measurement

1 Introduction

In the field of psychology, and in the social sciences generally, rating scales (e.g., PQS) and coding systems (e.g.: the GMI, Auletta et al. 2012; for the social science see: Marradi 1981; Corbetta 1999) are widely used. Such measures are generally based on evaluations performed by independent competent raters. As a consequence of such an involvement, these measures have to be shown not to suffer from the reliability problems potentially associated with any human judgement (Gigerenzer and Todd 1999). To this end, researchers are accustomed

E. Ciavolino (✉) · S. Salvatore
Department of History, Society and Human Studies, University of Salento,
Palazzo Parlangeli, Via V.M. Stampacchia, 45, 73100 Lecce, Italy
e-mail: enrico.ciavolino@unisalento.it

S. Salvatore
e-mail: sergio.salvatore@unisalento.it

A. Calcagnì
Department of Psychology and Cognitive Science, University of Trento,
Corso Bettini, 84, 38068 Rovereto, TN, Italy
e-mail: antonio.calcagni@unitn.it

to adopting inter-rater agreement as the standard method of estimation of reliability. The rationale of such a method is grounded in the classic theory of measurement: any instance of measurement is composed of the true depiction of the measured object plus an error component; the former component being invariant across instances, and the latter being the more limited the more similar are the instances of the output. From this is drawn the conclusion that the level of inter-rater agreement is an estimate of the incidence of the true depiction of the object.

Despite its widespread adoption, such a rationale possesses a logical flaw that cannot but have relevant consequences for the interpretation of the inter-rater agreement as a measurement of reliability. To put it in general terms, if two instances of the measurement (of the same object) had no error component, then they would produce the same output. Yet the converse is not necessarily true: the fact that two outputs are identical does not mean that there is no error component. And the same can be said in the case of disagreement among raters: it does not necessarily mean that the error component is high. In sum, a high inter-rater agreement is neither a necessary nor a sufficient condition for considering a measure to be reliable, and neither is a low inter-rater agreement a necessary or sufficient condition for considering a measure to be unreliable (Gwet 2001).

The logical flaw mentioned above does not mean that the inter-rater agreement has to be abandoned. Rather, it means that it has to be integrated with other sources of information. The aim of this paper is to present a method of the appraisal of the quality of a raters judgement that can integrate the use of inter-rater agreement as a measure of reliability. To this end, first we discuss the rationale underpinning the interpretation of the inter-rater agreement and highlight some of the relevant critical issues. Second, we present our integrative proposal: a method of analysis of the quality of a raters performance based on the analysis of the profile of the individual raters performance. In brief, such a method moves the focus from the agreement among judges to the inner quality of the individual performance of the evaluation. To this end, a computational model based on fuzzy set theory (FST, Zadeh 1965) is adopted. Third, the rationale of the model and its computational characteristics are discussed. Finally, a case study is presented for the purpose of giving an example.

2 Inter-rater agreement as an index of reliability

Figure 1 maps the basic possible relationships between the level of agreement and the actual reliability of a generic measure based on human judgements. For the sake of simplicity, both dimensions are considered as dichotomic YES/NO and only two raters measuring one item are considered. As is shown, the agreement may be the consequence of a high proportion of the true component in both instances (Hit), and the disagreement may be the consequence of a marginality of the true component in both instances (Correct Refusal). These are the two conditions referred to by the rationale for inter-rater reliability. Yet there are two other possibilities: the disagreement may also depend on the marginality of the true component in only one instance (Omission), or the agreement may also depends on the sharing of the error component, the true component being marginal (False Alarm). In what follows, we focus on the former type of error, both because it is the one with greater cost and because is less controllable (see below).

Several criteria and procedures have been adopted to reduce the risk of False Alarm: e.g., the use of trained raters and increasing their number. These procedures make sense but are not decisive. This is evident as one takes into account how the psychological processes

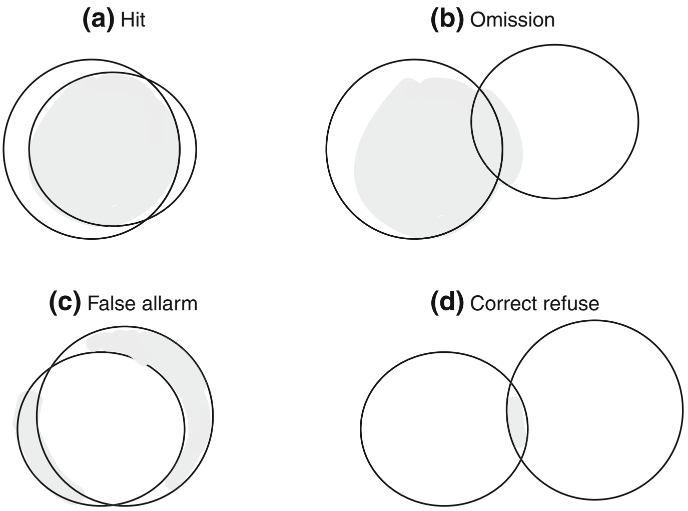


Fig. 1 Possible relations between true and error components in an agreement measure (the true component is represented in *gray*, the false component in *white*)

of decision making work and how they are performed in the socio-institutional context of psychotherapy research. More particularly, three aspects must be considered. (A) As a huge number of authors have shown, human judgements are not performed in terms of an acontextual implementation of invariant, abstract rules (Gigerenzer and Todd 1999). Rather, they are guided by more or less implicit heuristics, grounded on context-sensitive, culturally guided basic assumptions (Salvatore and Freda 2011; Valsiner 2007). (B) Many rating scales used in the psychosocial field are very demanding in terms of cost, time, and effort; this has two implications: on the one hand, most measures have little diffusion across research groups, their use being associated with a specific research group; on the other hand, in most cases the judges are selected from within the same research environment as that of the study being performed. (C) Each research group, as with any other social/work group (Moscovici 1976), is characterized by a peculiar, more or less implicit, network of epistemological, theoretical, and socio-cultural assumptions (Weick 1995), reflecting the position of the members within the scientific community as well as the cultural world (Matusov on science as daily work). As a consequence of the combination of these elements, it is plausible that in the process of measurement a role is played by what we propose to call the localism effect, i.e., the tendency of a shared system of assumptions—and the related interests, desires, and organizational modalities—to orient the judgements performed by the members of the research group. Insofar as the degree of inter-rater agreement depends on the localism effect, i.e., the raters common membership, the less one can interpret it as an index of reliability. In the final analysis, the more the autochthony of the practice of research, i.e., the use of members of the research group as raters, the more is the probability that the agreement reflects a systematic error due to the localism effect rather than the true component of the measure. It is worth noting that the main modality adopted for increasing the reliability and its estimation is not able, by itself, to control the localism effect. As a matter of fact, the localism effect is not affected by the number of raters, it depends on their membership and the level of their identification with the culture of the research group. The involvement of raters in training as well as in consensual meetings can even increase the salience of the localism effect, in the event of such procedures' being implemented in a context of autochthony.[1]

3 A method of appraisal of a rater's performance: *marginal sensitivity*

These considerations lead to the conclusion that the logic of conformism used in inter-rater agreement has to be combined with other modalities of reliability estimation of measures based on human judgement. In accordance with this perspective, our proposal is to focus on the quality of a rater's performance. More specifically, we propose to consider such quality in terms of the rater's marginal sensitivity (MS). The greater is such a capability, the more is the information the rater is able to take into account, the greater is the precision of the evaluation. Accordingly, MS lends itself to be conceptualized as reflecting the rater's autonomy by implicit higher ordered assumptions guiding the judgements. What follows will clarify this statement.

First, consider a set of q items of the kth dimension D_O^k (e.g., latent variable) of the object O to be measured. Moreover, define H_D to be the global representation that the rater elaborates as to the set of dimensions $D_O = \{D_O^1, D_O^2, \ldots, D_O^K\}$, where K is the number of dimensions considered. Thus, H_{D_O} is a higher-ordered concept for the representation of the singular items (Ciavolino 2012); in other words, it is the representation of the class of which the items are members.

Now, one can define two basic, opposite heuristics as the ground of the rating.

- The *top-down way*. According to such a way, the ratings are guided by H_{D_O}. This is the same as saying that a (generally implicit) higher ordered assumption concerning the whole dimension D_O tends to guide the ratings of the singular aspects as, for this reason, we call such a heuristic a *typological judgement*.
- The *bottom-up way*. In this case any item is rated in accordance with its own specific content; such ratings being independent by H_{D_O}. For this reason, we call such a heuristic a *content specific judgement*.

As is largely recognized in the psychological literature, the way of thinking in daily life works mainly in top-down ways. Only in specific and limited circumstances do persons retrieve a global representation by a systematic computation of its components. Usually, persons use a few cues from global, implicit representations for interpreting the singular elements of experience (e.g., Gigerenzer and Todd 1999; Mannarini et al. 2012). On the other hand, such a heuristic reduces the MS, because the elements tend to be homogenized by reason of their common membership in the higher-ordered class. Thus, one is led to conclude that the MS depends on the adoption of the individual judgement as a heuristic. In the next section, a method able to estimate the rater's MS will be provided.

[1] A strategy that tries to avoid such problems is the centralization of the training, namely only a specific research group, usually the one that has developed the instrument, is recognized as legitimated to train the raters, and this is often formalized in terms of a specific program to be attended for formal certificating (e.g., the procedure for obtaining the certification required for using the Adult Attachment Interview). Yet this solution is only partial, and has more costs than benefits. First, the centralization of the training does not solve the problem of the application of the instrument to clinical material reflecting a different cultural context. Thus, rater B can be trained by rater A whose way of coding is taken as normative; and so B can get a satisfactory level of agreement. Yet such agreement concerns specific objects, the ones used in the training and it is not obvious that it can be generalized to Bs cultural or research contexts. Secondly, due to its cost, the centralization of the training is a mechanism that can treat only a very limited subset of raters and instruments. Above all, the centralization of the training, as with any form of monopoly, reduces the free circulation and exchange of knowledge that is the ground of any scientific community.

4 An FST based method to detect the marginal sensitivity

In this section we suggest a formal framework suitable to set up a specific method capable of capturing a rater's MS, as previously stated. For this purpose, in the next paragraph we briefly introduce the main key concepts of FST, then we describe the method based on FST.

4.1 Fuzzy set theory: a brief introduction and some key concepts

Commonly, FST is understood as a theory of vagueness or fuzziness (Klir and Yuan 1995; Ross 2009; Zadeh 1965) and can be used to represent 'fuzzy concepts' or vague states of the world, the ones that are typically employed in psychology (Hesketh et al. 1988, 1989; Zétényi 1988) and the social sciences (Coppi et al. 2006; Rampone and Russo 2012; Verkuilen and Smithson 2006). To begin with, consider the difference between fuzzy and crisp information:

(a) Tom's age is 35
(b) Tom is young

The first proposition is crisp (non-fuzzy) and expresses 'testable information' (Jaynes 1968), namely one can verify its occurrence at the time and map it in terms of a specific numerical measure. The second proposition express 'fuzzy information', namely: what does 'young' indicate? How can we represent this concept in order to associate to it a specific numerical measure? To represent the first source of information, we can apply a traditional mathematical framework, but for the second one we can use a specific framework, different than the first one and capable of capturing the main features of fuzzy information (Zadeh 2005). Before giving the details of our proposal, it might be useful to present some of the key concepts of fuzzy sets, as follows. Let us consider an universal set $U = (x_1, x_2, \ldots, x_i, \ldots, x_n)$ with n elements; a fuzzy set A is a subset of U whose elements non-strictly can belong to it. In a more general sense, it can be defined by a couple of elements:

$$A = (x_i, \mu(x_i; \mathcal{P})) \quad \forall x_i \in U \tag{1}$$

where the x_i are elements of u and $\mu(x_i, \mathcal{P})$ is a parametric function that assigns to each x_i a positive value in the range $[0, \ldots, 1]$:

$$\mu : U \rightarrow [0, \ldots, 1]$$

where 0 indicates null membership, but 1 signifies full membership. Naturally an element can be described by a fuzzy membership and this is specified by the membership function (via $\mathcal{P}$ parameters). Several membership functions are available, such as the triangular membership function, sigmoidal, Gauss curve, etc. To give an example, Fig. 2 represents a triangular membership function defined as follows:

$$\mu_{tr}(x_i) = \begin{cases} 0, & \text{if } x_i < a \text{ and } x_i > b \\ \frac{a-x_i}{m-a}, & \text{if } a \leq x_i < m \\ \frac{x_i-m}{b-m}, & \text{if } m < x_i \leq b \\ 1, & \text{if } x_i = m \end{cases} \qquad \forall x_i \in A \tag{2}$$

The membership function informs us about the possibility of a value's belonging to a specific set; this is one of the possible interpretation of Zadeh's theory (Verkuilen and Smithson 2006; Zadeh 1999), allowing the definition of a fuzzy set as a specific set of elements with its own characteristic possibility distribution.

To conclude this brief introduction, it is now necessary for our purposes to detail the concept of a fuzzy variable, as follows: let us assume a generical universal set U with a

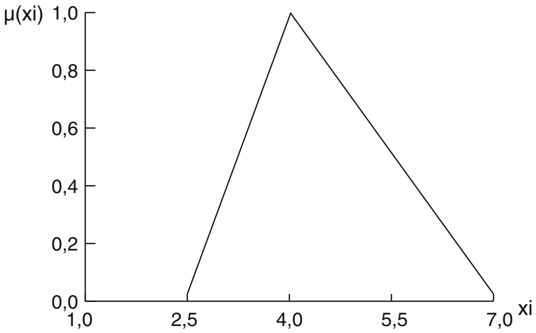


Fig. 2 A triangular fuzzy set (convex set)

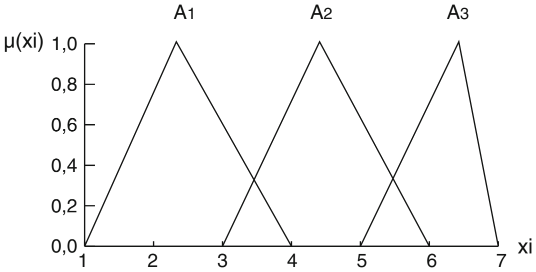


Fig. 3 A graphical representation of triangular fuzzy variable with three overlapping fuzzy sets

group of m fuzzy sets, $A_1, A_2, \ldots, A_i, \ldots, A_m$, as defined above (Eq. 1). A *fuzzy variable* is defined by the following formal structure:

$$f = \langle U; (A_1, A_2, \ldots, A_i, \ldots, A_m) \rangle \tag{3}$$

where U is the domain of the family of fuzzy sets. Because a fuzzy set can be defined by different membership functions, we can obtain different fuzzy variables given the $\mathcal{P}$ parameter defined above.

A fuzzy variable is able to capture so-called ‘empirical knowledge’ (Ross 2009) and, in particular, it can represent a fuzzy object; as we can notice by Fig. 3, the fuzziness of a phenomenon is reproduced mainly by two features: the possibility distribution on U and the overlap space among the fuzzy sets.

4.2 Empirical evidence based on a psychotherapy research study

In the previous sections, we have introduced the key concepts of our proposal, specifically the MS as a possible index of the incidence of bottom-up rating modality. Now, we propose a possible method capable of implementing the information about this process and specifically we measure it by the evaluation of the top-down way. In our assumption, the presence of the first process reduces the presence of the second one. For this purpose, let us deal with the following rationale.

Given a matrix $\boldsymbol{X}$ of n raters who express m judgements (items) of a specific object O. $\boldsymbol{D}$ is a matrix of n raters and K unobservable variables. To define the generic element $\boldsymbol{d}_k$, we use a statistical method (Principal Components Analysis) allowing us to represent each set of judgements in a subspace $\mathbb{R}^K$ of dimension $K < M$. Now, considering a single vector $\mathbf{d}_k$, we can assume a function $\delta : \mathbf{d}_k \rightarrow \mathbb{R}^+$ that assigns to each judgement score a real positive

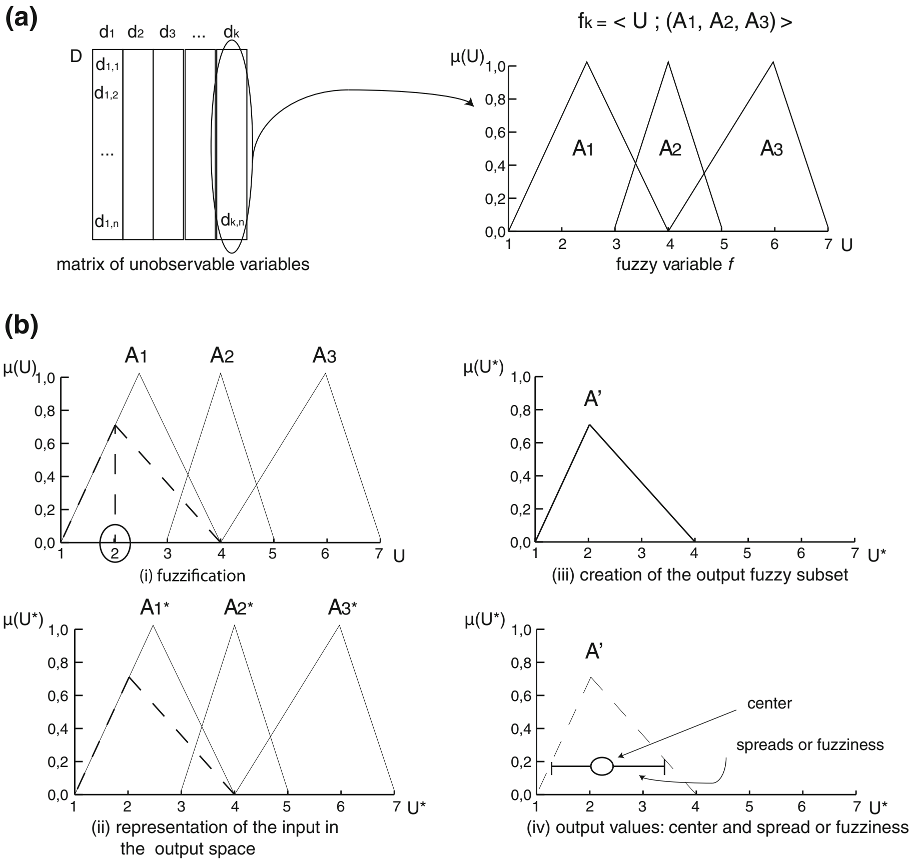

(a)
d1 d2 d3 ... dk
D
d1,1
d1,2
...
d1,n
dk,n
matrix of unobservable variables
fk = < U ; (A1, A2, A3) >
μ(U)
1,0
0,8
0,6
0,4
0,2
0,0
A1
A2
A3
1 2 3 4 5 6 7 U
fuzzy variable f
(b)
A1 A2 A3
μ(U)
(i) fuzzification
A1* A2* A3*
μ(U*)
U*
(ii) representation of the input in the output space
A'
(iii) creation of the output fuzzy subset
center
spreads or fuzziness
(iv) output values: center and spread or fuzziness

Fig. 4 A graphical representation of **a** granulation procedure and **b** fuzzy system methodology

number; we refer to this function as the index of marginal sensitivity (MS-index) informing us about the attractiveness of the variable k in the score of judgement $d_{i,k}$.

4.3 An FST definition of the function δ

To give a computational form for the function δ, we can use the fuzzy system methodology. Considering a set of variables grouped into two classes, the *input* variables and the *output* variables, and a set of *rules* bridging the two, a *system* is a structure composed of variables and rules with a proper behaviour based on these. A *fuzzy system* is a classical system whose variables are fuzzy and rules are chosen considering a class of fuzzy rules (called fuzzy inference rules, expressed by an inference way ‘if A then B’). In our proposal, a fuzzy system must be able to represent, in a fuzzy way, the single score of judgement expressed by a numerical value. Without loss of generality and for our purposes, a so-defined fuzzy system takes as input each fuzzy variable (f_k) representing each unobservable variable ($\mathbf{d}_k$) and has as output a *structural copy* f_k^* of the input (the output variables are the same of input variables). For each input–output couple, we define an implication rule allowing us to represent the input activation in the output space. At once, since each judgement score is represented by a fuzzy subset, we can compute a measure of fuzziness for each score, informing us about the degree of the MS index.

To summarize, we can say that each vector **d** (in the matrix **D**) is represented by a fuzzy variable *f*, the fuzzy variables thus obtained are used as the input of the fuzzy system, and a copy of these represent the output of the system (naturally, if we have several pieces of a priori information, we can use as output other fuzzy variables opportunely chosen) with the set of implication rules based on the Max-Min Mamdani approach (Ross 2009). A graphical summary is presented by Fig. 4.

In Fig. 4a, we can notice a procedure that allows us to represent each unobservable variable through a specific fuzzy variable where the fuzzy levels can be modelled by the empirical data (Medasani et al. 1998) or by an heuristic process based on the information a researcher possesses about the phenomenon considered.

In Fig. 4b we can see a graphical representation of a Mamdani fuzzy system: (1) the fuzzification operation on each $\mathbf{d}_k$, (2) the representation on the output space by the rules defined, (3) the definition of the output fuzzy subset, and (4) the representation of three characteristic of the information: the center value computed through the center of gravity (COG) method of defuzzification (Ross 2009), and the spread value, namely the Euclidean distance computed between the upper/lower bound of the set and the center. The spread information or fuzziness is the value of the MS index we want to capture, thus we have for each $d_{i,k}$ a $\delta(d_{i,k})$ informing us about the marginal sensitivity and the magnitude of the bottom-up way of rating (the greater the MS index, the more presence of the bottom-up rating way, and viceversa).

5 A method of estimation of the marginal sensitivity based on fuzzy set theory (FST). An application

In this section an example of an application of the method and computational strategy is presented.

5.1 Purpose

The aim of the study has been the estimation of the MS of raters applying the Psychotherapy Process Q-Set (PQS, Ablon and Jones 2002; Jones 1985) to the verbatim transcript of a set of sessions from a single psychotherapy. Moreover, the FST based method is compared with an indirect index of the raters' competence.

5.2 Object of the analysis: the PQS

The PQS is a method widely used in the field of psychotherapy process research (Ablon and Jones 1998, 1999, 2002; Ablon et al. 2006; Auletta et al. 2012; Jones et al. 1991), aiming at depicting the global quality and characteristics of the therapist–patient relationship. It is made up of 100 items concerning three areas of investigation: (a) the patient's attitude toward and experience of the therapy; (b) the therapist's actions and attitude; (c) the nature of the interaction and the climate of the encounter between the therapist and the patient. For each session, the rater is asked to sort 100 items on 9 ordinal categories on the continuum from absolutely not so (score 1) to absolutely so (score 9). The Q-set methodology (Block and Reed 1978) implies a forced choice by which each category has to be composed of a specific and already fixed number of items. In the case of the PQS, the forced choices obey a normal distribution (from 5 items to be put in each extreme category to 9 items in the middle one). Several studies have highlighted both the reliability and validity of the PQS method (Ablon and Jones 1999; Jones and Pulos 1993).

5.3 Rationale

The FTS based method was performed through the following *ex-post* computational strategy. The PQS scores on 100 Likert scales were subjected to a principal component analysis (PCA),[2] in order to summarize the information in a few dimensions of variability. The factorial dimensions were then mapped in terms of fuzzy variables. In other words, each latent variable was described through a fuzzy variable and each judgement score was represented as a fuzzy subset. The procedure adopted can be defined as *ex post*, because it is based on the construction of fuzzy variables once the judgements have been performed and represented by the latent variables carried out by the PCA.

It is worth noting that this *ex post* procedure is a modified version of the standard procedure, that is the on-time one, performed by means of an analogical modality of answering, mediated by the computer. The on-line procedure works directly on the fuzzy data, namely the data produced directly in terms of fuzzy information. Although the on-time procedure is more suitable, in terms of its capability of preserving the information stored in each empirical rater's judgement, it is based on methods requiring specialized software (e.g., the rater is asked to answer by selecting a point of a scale shown on the display. In so doing, the discrete Likert scale is transformed to a continuum variable enabling the fuzzification procedure). We chose to adopt an *ex post* procedure, because it does not impose additional requirements compared to the paper and pencil modality of performing judgements usually adopted in psychosocial research.

5.4 Data analysis

A four-step procedure of data analysis was adopted.

Step 1—PQS application Six raters with at least medium competence in psychotherapy process research received 40 h of training in the use of the PQS. At the end of the training, the average inter-rater agreement, calculated on a random sample of 10 sessions taken randomly from four psychotherapy session, different in terms of their length and theoretical orientation, was substantial (Cohen's K coefficient = 0.73; cf. Fleiss et al. 2004). Then, the raters applied the PQS to the whole set (100 items) of a 76 session good outcome psychotherapy: the Max Case.

Step 2—Selection of relevant information The 100 items × 76 sessions data matrix thus obtained was preliminarily elaborated, to transform the scale from an ordinal to a quasi-cardinal scale, in order to improve its mathematical properties. The method adopted is based on the Thurtstone rationale, whose main idea is that the criterion of the choice of each interviewee follows a latent variable distributed as a normal distribution (Ciavolino and Dahlgaard 2007). Then, following Ablon and Jones (Ablon and Jones 1998, 2002) the matrix was subjected to a first PCA. This procedure led to the identification of three main factorial dimensions. PQS items more representative of the three factorial dimensions (loading $>$ .55) were selected. Thus 10 items out 100 were maintained in the following steps.

Step 3—Fuzzification For each rater, a matrix was defined, composed of the selected items applied only on the sessions used to calculate the inter-rater agreement. Thus, 6 sessions

[2] The use of the PCA for basic dimension characterization and the description of a clinical process is a procedure widely used in studies using such methods.

Table 1 δ values for the six judges on the three latent variables

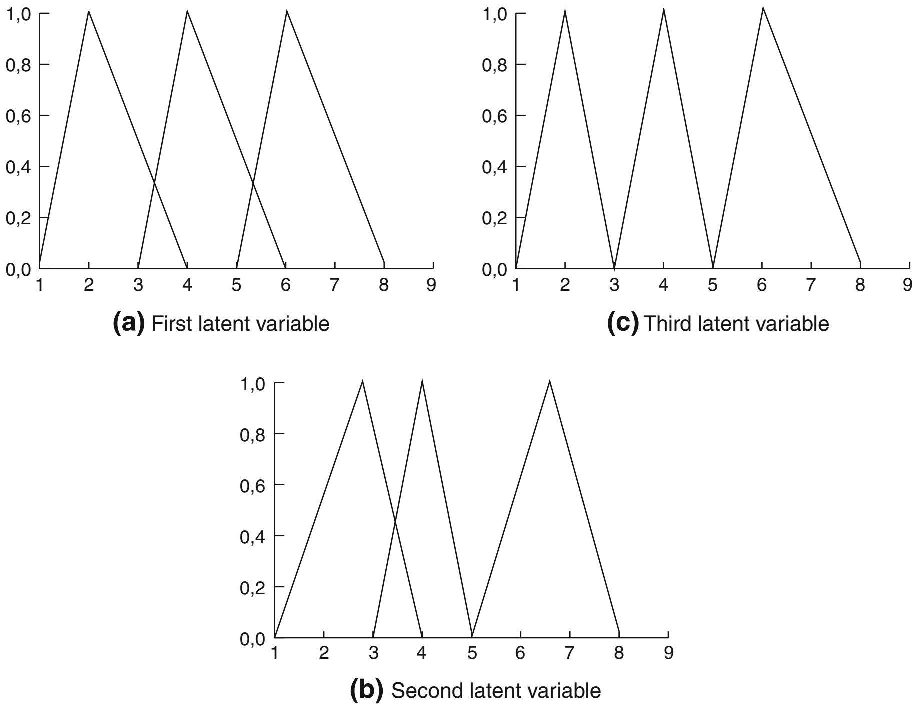
f_1	f_2	f_3
First judge		
3	3	3
3	3	3
3	3	3
5	2	3
3	3	3
3	3	3
3	3	3
3	3	3
3	3	3
3	3	3
Second judge		
2	3	4
2	3	4
2	3	2
2	3	4
3	3	4
5	4	3
5	5	2
3	4	4
3	3	2
3	3	4
Third judge		
3	4	2
3	3	3
4	4	5
3	3	5
4	4	3
3	2	3
3	2	3
3	4	3
4	4	2
3	3	2
Fourth judge		
3	4	2
3	3	3
4	4	5
3	3	5
4	4	3
3	2	3
3	2	3
3	4	3
4	4	2
3	3	2

Table 1 continued

f_1	f_2	f_3
Fifth judge		
2	3	3
2	3	3
2	3	3
2	3	4
2	3	3
2	3	3
2	3	3
2	4	3
2	3	3
2	3	3
Sixth judge		
3	3	3
3	3	3
3	3	2
3	2	3
3	3	3
3	3	5
3	5	5
2	3	3
2	3	3
4	2	3

(rows) per 10 items (columns) were obtained. A further PCA was performed on each matrix in order to extract the latent variables underlying the judgements for each judge. Hence, the first three factorial dimensions for each PCA were considered and each latent variable was subjected to the fuzzificaton procedure. In order to perform the fuzzification procedure, first, as a preliminary, two operation were performed. First, the factorial scores obtained through the second PCA were rescaled from 1 to 9, transforming them to discrete values. Second, an Python script was applied: this was a tailor-made algorithm based on Information Theory, capable of implementing the best fuzzy variables from the empirical data.[3] The six matrices were then subjected to the fuzzification procedure, thus obtaining the fuzzy final data as reported in Fig. 5 (which shows the first judge only). As we can see from the figure, the three fuzzy variables (f_1, f_2, f_3) have three overlapping fuzzy sets; note that the position of the fuzzy sets together with the degree of overlap are set up by the application, in agreement with the empirical data we have used. Thus, each row of the latent variables is represented by a fuzzy subset on f_1, f_2, and f_3, as we can see from Fig. 6. In this way, each fuzzy set

[3] Briefly, we set up a specific algorithm based on the concepts of data information and entropy; in particular, this procedure allowed us to reconstruct an histogram related to a dataset with a specific fuzzy set (e.g., a triangular fuzzy set) capturing the whole information, as measured by the entropy formula, that was present in the empirical data. Strictly speaking, if we have a data vector represented by a specific histogram, we can express it by the best fuzzy set capturing the information present in the vector (using a specific formula, De Luca and Termini's fuzzy weighted entropy), thus the fuzzy set obtained in this way best describes the empirical data vector. For the purposes of this paper, we do not report other details, which can be requested directly from the authors.



(a) First latent variable

(c) Third latent variable

(b) Second latent variable

Fig. 5 A graphical representation of the three latent variables for the first judge only

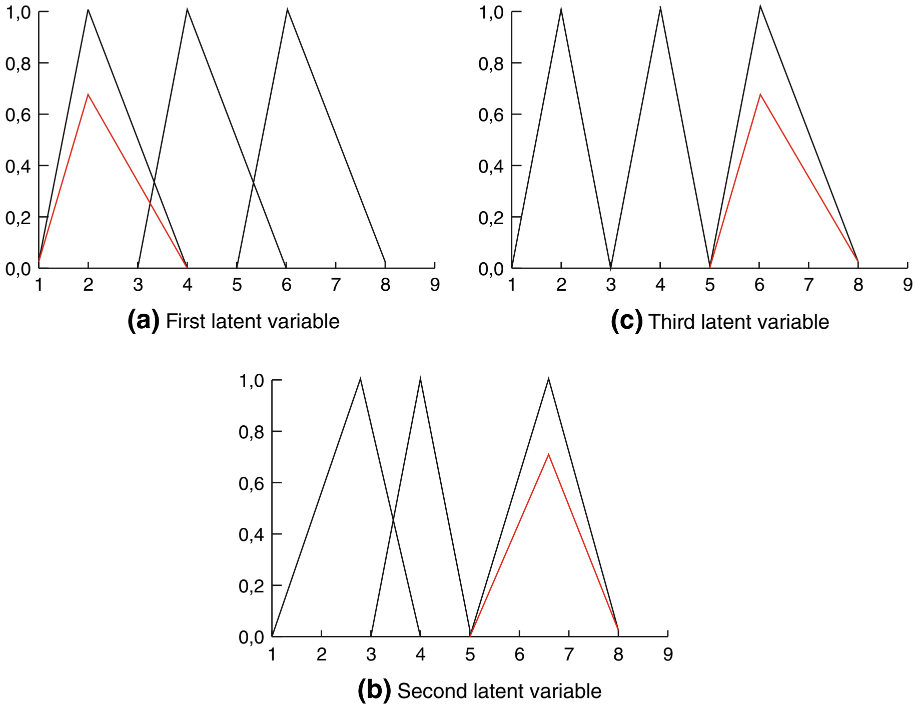
is synthesized from three values: the center, the left spread, and the right spread, as seen in Fig. 4iv.

*Step 4—Calculation of the raters' MS (δ function)*The fuzzification procedure leads to the calculation of the left and right spreads for each variable. This enable computing the marginal sensitivity (MS, or fuzziness value), for each latent variable, as the sum of the two spreads (see Table 3).

5.5 Results and general discussion

Figure 7 shows the MS mean for each judge on the three fuzzified latent variables (f_1, f_2, f_3). Figure 8 shows the judgement trend chart for the variables with the most fuzziness: the line represents the judgement trend among the ten sessions (horizontal axis) while the vertical bars represent the MS of the judgement (the sum of the left and right spreads); to understand these charts we can say how the length of the vertical bars suggests to us the magnitude of the fuzziness (the greater is the fuzziness, the greater is the effect of the bottom-up rating way because the attractiveness of the high-ordered variable upon the judgement is low) while the circles indicate the center of the fuzzy subsets representing the judgement scores.

The fuzziness thus computed for each judge informs us of the magnitude of the bottom-up way and particularly for their MS-capability (see Table 1); specifically, by the analysis of the charts presented in Fig. 8, we can notice two principal patterns: a near-to-stability pattern (judges 1 and 4), and a steadiness pattern (judges 2, 3, 5, and 6). The first of these indicates the position of the judgement scores in the levels of the latent variable considered, although this result does not give us information about the fuzziness of the judgement scores, in fact as we can see from judges 2, 3, 5, and 6, steadiness is not related with the fuzziness that appears



(a) First latent variable

(c) Third latent variable

(b) Second latent variable

Fig. 6 A graphical representation of fuzzified judgement on item 1 on the three fuzzified latent variable (first judge only)

stable (see the vertical bars on the charts) while for the two judges with a stable trend (1 and 4) we notice more fuzziness in several sessions; generally speaking we can notice how only a few sessions have more fuzziness on the latent variable and this result indicates how the capability of capturing marginal information by the judges is related to a few sessions: to give an example, and taking into account judge 2, we can see how sessions 6, 7, and 8 have more fuzziness than the other ones and this can be understood as a greater capability of using the way of bottom-up rating for the judge considered, while judge 4 has a stable fuzziness between the sessions or, rather, he presents a prevalence of the top-down rating way.

Afterwards, we evaluated, for the six judges, the correlation between the mean on their coefficients of agreement (IRA), highlighting the agreement mean over the ten psychotherapy sessions (IRA_{mean}) and the mean of the MS indices (MS_{mean}) over the three latent variables (Tables 2 and 3).

As we can notice from Table 3, the MS-index is associated with the IRA coefficient for the first and second fuzzified latent variable. These results, although showing a weak association, highlight that the MS-index is related to the IRA coefficient and, generally speaking, encourage us to make further investigations in this direction and to consider the application of the proposed method in combination with the traditional ones.

In order to give a general scenario about our results, some computational details have to be explained. As we have previously seen, we use the concept of fuzziness as an index of a judge's capability to use a bottom-up heuristic. To carry this out, we have adopted a suitable procedure based on the fuzzy system methodology. For this purpose, it is necessary to note that our results about fuzziness are related to the empirical data and to the sample size; this latter is a central feature of our procedure because the definition of the fuzzy variable is based

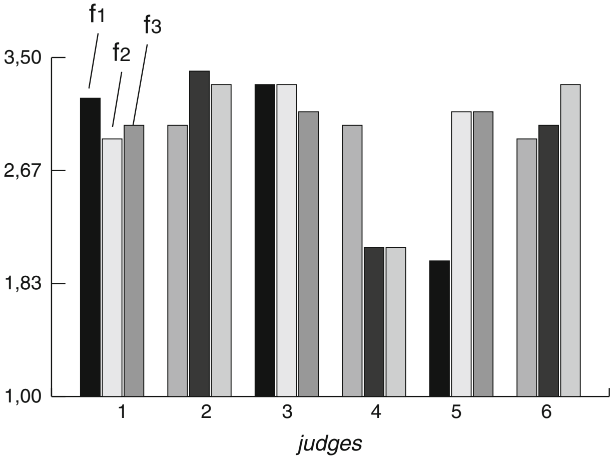


Fig. 7 Marginal sensitivity mean on the three fuzzified latent variables and for the six judges

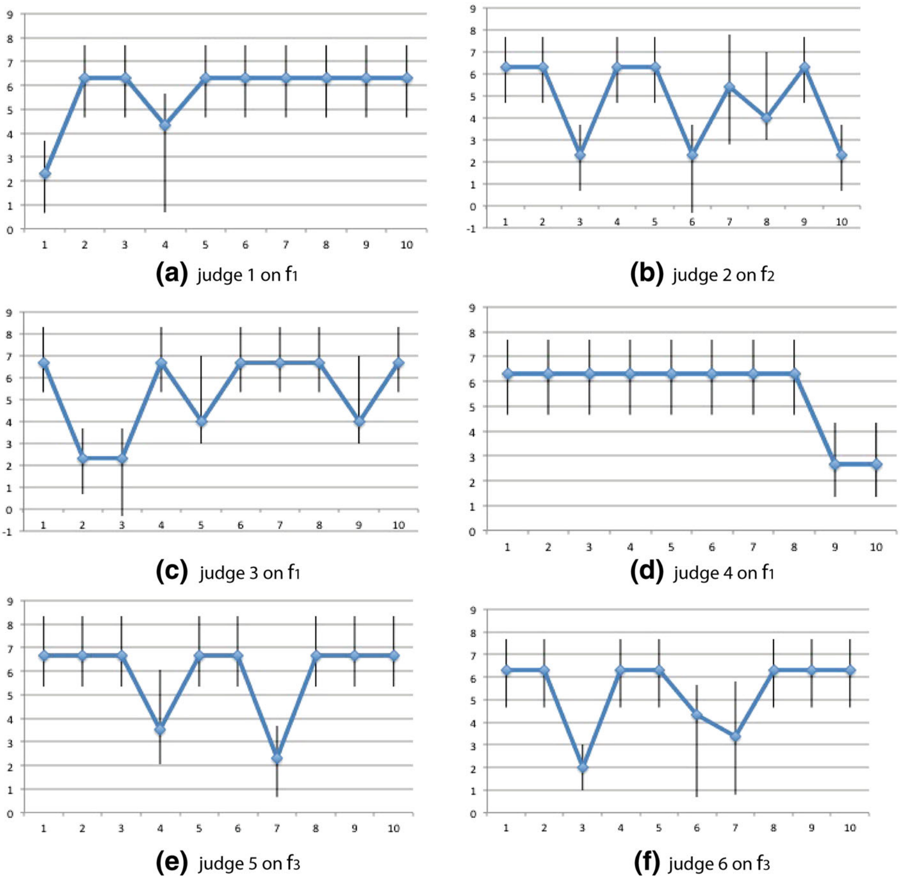


Fig. 8 MS trend chart for the six judges

Table 2 Mean of the CIR and MS for the six judges over the ten psychotherapy sessions

$Judge$	IRA_{mean}	MS^{I}_{mean}	MS^{II}_{mean}	MS^{III}_{mean}
1	0.65	3.2	2.9	3.0
2	0.67	3.0	3.4	3.3
3	0.62	3.3	3.3	3.1
4	0.63	3.0	2.1	2.1
5	0.55	2.0	3.1	3.1
6	0.48	2.9	3.0	3.3

Table 3 Pearson correlation between IRA_{mean} and MS_{mean}

	IRA_{mean}
MS^{I}_{mean}	0.45
MS^{II}_{mean}	−0.023
MS^{III}_{mean}	−0.28

on a data-oriented method. We would like to point out that small samples can be a problem for the working of the algorithm for the construction of the fuzzy sets, and thus the results about the fuzzy sets and their overlap space definition have to be managed with care, and consequently, the same holds for the results about the fuzziness computed (the values of the δ function).[4] This consideration can be extended to the correlation results between the MS and the IRA. This aspect and others allow us to regard this study as a preliminary case exploring our theoretical and methodological proposal and for this reason we have chosen to manage these results with some caution, using an exploratory attitude.

6 Conclusion and further remarks

In this paper we have proposed a new methodology capable of evaluating the quality of inter-rater agreement. Thus, after having discussed some theoretical weaknesses (see Sects. 1 and 2) inherent in the traditional approach to inter-rater agreement, we proposed a specific methodology based on the concepts of fuzzy set theory, in particular the concept of *fuzziness*, understood as an index of a judge's marginal sensitivity. In Sects. 4 and 5 we discussed our proposal and some results obtained through an case study based on psychotherapy research. We can note that our method allows us to capture a piece of 'pure' information about the agreement, exempt from the conformism effect or the 'alone effect': the MS index is based on an individual measure of agreement and thus is not related to other concepts such as the variance of agreement. In future research, we will try to explore a possible definition of the inter-rater agreement entirely based on fuzzy set theory, and thus capturing a new piece of information (the fuzziness of the judgement), informing us about the implicit adhesion of the judge to a prototypical representation of the object evaluated.

[4] The fuzziness computed by the system is related to the definition of the overlap space between the fuzzy sets, and thus the greater is the overlap space, the greater is the possibility of having judgement scores with an high degree of fuzziness.

References

Ablon, J., Jones, E.: How expert clinicians' prototypes of an ideal treatment correlate with outcome in psychodynamic and cognitive-behavioral therapy. Psychother. Res. **8**(1), 71–83 (1998)

Ablon, J., Jones, E.: Psychotherapy process in the national institute of mental health treatment of depression collaborative research program. J. Consult. Clin. Psychol. **67**(1), 64 (1999)

Ablon, J., Jones, E.: Validity of controlled clinical trials of psychotherapy: findings from the nimh treatment of depression collaborative research program. Am. J. Psychiatr. **159**(5), 775–783 (2002)

Ablon, J., Levy, R., Katzenstein, T.: Beyond brand names of psychotherapy: identifying empirically supported change processes. Psychother. Theory Res. Pract. Train. **43**(2), 216 (2006)

Auletta, A., Salvatore, S., Metrangolo, R., Monteforte, G., Pace, V., Puglisi, M.: The grid of the models of interpretations (gmi): a trans-theoretical method to study therapist interpretive activity. J. Psychother. Integr. **22**(2), 61 (2012)

Block, R., Reed, M.: Remembered duration: evidence for a contextual-change hypothesis. J. Exp. Psychol. Hum. Learn. Mem. **4**(6), 656 (1978)

Ciavolino, E.: General distress as second order latent variable estimated trough PLS-PM approach. Electron. J. Appl. Stat. Anal. **5**(3), 458–464 (2012)

Ciavolino, E., Dahlgaard, J.: Ecsi-customer satisfaction modelling and analysis: a case study. Total Qual. Manag. **18**(5), 545–554 (2007)

Coppi, R., Gil, M., Kiers, H.: The fuzzy approach to statistical analysis. Comput. Stat. Data Anal. **51**(1), 1–14 (2006)

Corbetta, P.: Metodologia e tecniche della ricerca sociale. Il Mulino, Bologna (1999)

Fleiss, J., Levin, B., Paik, M.: The measurement of interrater agreement. In: Statistical Methods for Rates and Proportions, 3rd edn. pp. 598–626 (2004)

Gigerenzer, G., Todd, P.: Fast and frugal heuristics: the adaptive toolbox (1999)

Gwet, K.: Handbook of inter-rater reliability. Adv Anal LLC (2001)

Hesketh, B., Pryor, R., Gleitzman, M.: Fuzzy logic: toward measuring gottfredson's concept of occupational social space. J. Couns. Psychol. **36**(1), 103 (1989)

Hesketh, T., Pryor, R., Hesketh, B.: An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences. Int. J. Man Mach. Stud. **29**(1), 21–35 (1988)

Jaynes, E.: Prior probabilities. Syst. Sci. Cybern. IEEE Trans. **4**(3), 227–241 (1968)

Jones, E.: Psychoterapy Process Q-set Coding Manual. University of California, Berkeley (1985)

Jones, E., Hall, S.A., Parke, L.A.: The process of change: the Berkeley Psychotherapy Research Group. In: Beutler L.E., Crago M. (eds.) Psychotherapy Research: An International Review of Programmatic Studies, pp. 99–106. American Psychological Association, Washington (1991)

Jones, E., Pulos, S.: Comparing the process in psychodynamic and cognitive-behavioral therapies. J. Consult. Clin. Psychol. **61**(2), 306 (1993)

Klir, G., Yuan, B.: Fuzzy Sets and Fuzzy Logic. Prentice Hall, New Jersey (1995)

Mannarini, T., Ciavolino, E., Nitti, M., Salvatore, S.: The role of affects in culture-based interventions: implications for practice. Psychology **3**(8), 569–577 (2012)

Marradi, A.: Misurazione e scale: qualche riflessione e una proposta. Quaderni di sociologia **29**, 595–639 (1981)

Medasani, S., Kim, J., Krishnapuram, R.: An overview of membership function generation techniques for pattern recognition. Int. J. Approx. Reason. **19**(3), 391–417 (1998)

Moscovici, S.: La psychanalyse: Son image et son public. Presses Universitaires de France-PUF, Paris (1976)

Rampone, S., Russo, C.: A fuzzified BRAIN algorithm for learning DNF from incomplete data. Electron. J. Appl. Stat. Anal. **5**(2), 256–270 (2012)

Ross, T.: Fuzzy Logic with Engineering Applications. Wiley, New York (2009)

Salvatore, S., Freda, M.: Affect, unconscious and sensemaking. A psychodynamic, semiotic and dialogic model. New Ideas Psychol. **29**(2), 119–135 (2011)

Valsiner, J.: Culture in minds and societies: foundations of cultural psychology. Psychol. Stud. (September 2009) **54**, 238–239 (2007)

Verkuilen, J., Smithson, M.: Fuzzy Set Theory: Applications in the Social Sciences, vol. 147. Sage Publications, Incorporated (2006)

Weick, K.: Sensemaking in Organizations, vol. 3. Sage Publications, Incorporated (1995)

Zadeh, L.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)

Zadeh, L.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **100**, 9–34 (1999)

Zadeh, L.: Toward a generalized theory of uncertainty (GTU)–an outline. Inf. Sci. **172**, 1–40 (2005)

Zétényi, T.: Fuzzy Sets in Psychology, vol. 56. North Holland, Amsterdam (1988)