A METHOD FOR SCALING VERBAL DISQUALIFICATION

JANET BEAVIN BAVELAS
University of Victoria

BEVERLY J. SMITH
Oxford University

A disqualification is a message that says something "without really saying it"—one that is evasive, indirect, or effectively ambiguous in some other way. This article describes the rationale for, and development of, an objective method for measuring the degree of disqualification in brief written messages: nonexpert judges independently assess four basic elements of the message (sender, content, receiver, context) for ambiguity, by use of a physical scale of length. The raw scale values are standardized for each judge, and the reliability of these values is quite high. Research uses of the method, as well as its limitations and possible modifications, are discussed.

Disqualification is a direct descendent of the double bind (Bateson, Jackson, Haley, & Weakland, 1956). Like its still active parent, disqualification is both an important and an elusive communicational phenomenon. There does seem to be agreement that disqualification does not equal double bind, principally because disqualification is a property of a message or transaction, whereas double bind refers to a much broader interpersonal situation (Sluzki, Beavin, Tamopolsky, & Verón, 1967). That is, disqualification is discrete and definite rather than global and pervasive.

Another major difference is that disqualification is not necessarily pathological or pathogenic; indeed, our interest in measuring disqualification arose in a project studying its occurrence as normal communication. The present article will describe the first phase of this ongoing research, the ultimate goal of which is to understand the social context of disqualified messages. We assume that "nonstraightforward" messages are part of the fabric of everyday communication—they are not rare, deviant, or random occurrences. On the contrary, we propose that they are predictably elicited by situations that proscribe more direct communication but still require a response. Therefore, ultimately it should be possible (1) to elicit communicational choices in experimentally varied situations, and (2) to tell precisely whether, or to what degree, these messages were disqualified. The skeptical reader would suggest that the second problem is the more critical since, as will be seen below, the several definitions to be found have remained too vague to be useful for the research being contemplated. In any case, it is necessary to demonstrate that there is a phenomenon capable of being identified and measured, before it can be explained.

APPROACHES TO DISQUALIFICATION

In the original clinical literature, disqualification was successively described as incongruent qualification of a message (Haley, 1959a); incongruence between levels of messages (Haley, 1959b); a contradictory message (Jackson, Riskin, & Satir, 1961); an indirect negation of what someone else has said, so that statements are not really met (Jackson & Weakland, 1961); not letting a message stand, clearly and unambiguously (Weakland & Fry, 1962); a technique that enables one to say something without really saying it (Watzlawick, 1964); incongruent messages, that is, messages denying each other or especially the messages of another person (Sluzki et al., 1967); and communication that invalidates one's own communication or that of the other, for example,
self-contradictions, inconsistencies, subject switches, tangentializations, incomplete sentences, misunderstandings, obscure style or mannerisms of speech, the literal interpretations of metaphor and the metaphorical interpretation of literal remarks, etc. (Watzlawick, Beavin, & Jackson, 1967, p. 76)

Similarly, disqualification has been described as a specific technique by which the broader phenomenon of interpersonal disconfirmation may come about (Cisna & Sieburg, 1979).

It is on the various classes of disqualification rather than on precise definitions that agreement seems to emerge. Evasion, tangentialization, literalness, inconsistency or self-disqualification, and so forth, are frequently mentioned. There is probably even better agreement on concrete, specific examples of disqualification:

A. A woman says, "No, no" to a man's advances while moving closer to him. (Haley, 1959a, pp. 322-323)

B. The mother of a student about to begin college writes: "Here is your change, Honey & I know you will do well - You have what it takes - I'm very proud of you - . . . P.S. If it doesn't work out - and [you're] not ready - don't worry - at least you tried." (Weakland & Fry, 1962, p. 620)

C. Daughter: We have always gotten along well.
   Mother: Yes, I've always loved you . . . in the same way. (Sluzki et al., 1967, p. 498)

Stepping back to take a broader, multidisciplinary view reveals that nonstraightforward messages have been described in a surprising variety of fields. The systematic ambiguity we are calling disqualification overlaps, at least in part, with Wiener and Mehrabian's (1968) nonimmediacy; Brown and Levinson's (1978) politeness strategies, such as hedging; Searle's (1975) indirect speech acts and Nofsinger's (1976) indirect responses; some of Grice's (1975) implicatures; and even rhetorical tactics, such as apophasis and enthymeme. What is even more encouraging is an apparent similarity of approach behind these diverse descriptions. All imply strongly not only that there exists a standard of direct and clear response, which is often violated, but also that such "violation" is explicable in terms of the social context of the message. As noted above, the second proposition is the subject of research to be described in a later paper. The immediate problem remains the explication of a model that will go beyond description and exemplification.

How do we know when a given message is a disqualification? The answer has been that experts in interpersonal communication can tell. Obviously, even if highly experienced or expert analysts did demonstrate high interjudge agreement, this would be encouraging but not conclusive. The concept of disqualification must be taken out of the realm of individual intuition and observational skill and into the next scientific stage. By this process of explication, we hope to understand better both the concept and the cues that distinguish it from other kinds of communication. And because of our interest in the pragmatic impact of communications, we need to demonstrate that ordinary persons can notice such phenomena, that these are effective stimuli for the population at large and not merely theoretical entities perceived only by experts. As Wilmot (1980) has noted, the reactions or judgements of "naive others" to communicational behaviors has not been a major area of research.

Expert agreement that a particular message is or is not a disqualification has had another, less obvious limitation. The usual judgment is dichotomous, indicating only the presence or absence of disqualification; yet it seems plausible that there are greater and lesser degrees of disqualification, in which case a yes/no judgement yields less information than was available in the message. Furthermore, dichotomization probably produces apparent unreliability of judgement in borderline cases; that is, different judges may draw the line between what is and is not disqualification in slightly different places. Measurement might better proceed from the assumption that disqualification is a property that messages may show in different degrees, some more extreme than others.
Our methodological aims, then, are specific yet ambitious. Degree of disqualification should be objectively scalable by nonexpert judges with no knowledge of communication theory or literature, whose training could therefore be completely specified and replicated. This should be done within a framework that unifies the diversity of phenomena that might be called disqualification. Finally, the process should support our thesis that these communicational subtleties are apparent to the persons our theories are about, namely, lay communicators.

In order to serve these goals, we have limited ourselves in other areas, focusing for the present on brief, written responses to the experimentally constructed situations we envisage, rather than on extended discourse or conversation within a dyad. While conceding that longish samples of dyadic interaction are more “naturalistic” in the sense of being more typical than short, structured exchanges, we maintain the latter are not rare. Moreover, we do not consider the written-verbal modality a necessarily trivial or artificial one; it depends on how it is handled. First, if one is going to work with verbal material, it is better to use material that is originally purely verbal than to transcribe only the verbal portion of a message that was originally delivered on all channels. In this way, the integrity of the message is preserved. Second, it is likely that when a message is written (rather than spoken), both sender and receiver make the appropriate adjustments. Information that might otherwise be conveyed, say, paralinguistically, is shifted over to the written form—by phrasing, punctuation, fuller verbal expression, and so forth. Finally, there are many nontrivial situations in which relatively short written messages are both natural and potentially rich in disqualification: notes passed in a meeting or class or left by one person for another, social notes (thank you, RSVP) and brief letters (Weakland & Fry, 1962), letters of reference or recommendation, memoranda, telegrams, campaign statements or press releases, written directions or instructions, ads or notices on bulletin boards, and so forth. In the following, we used only such situations, in which a brief written message was the natural and necessary medium.

AN OPERATIONAL DEFINITION

Our first and main problem was to conceptualize disqualification in such a way as to permit its translation into an appropriate yet simple empirical procedure. We found this possibility in what is probably the first article on the subject. When Haley (1959a) approached messages that are “qualified incongruently” (p. 323), he noted that certain fundamental principles of communication were involved. Any communication must contain, implicitly or explicitly, four formal elements:

1. I
2. am saying something
3. to you
4. in this situation. (p. 325)

In other words, there must be a sender, some content, a receiver, and a context. In a perfectly clear or congruent message, all four elements would be clear to the observer (i.e., who is saying what to whom in what context). More interesting for our purposes is the fact that each of these four elements can be rendered ambiguous, or incongruent with the actual state of affairs. Haley went on to illustrate how schizophrenic communication can be seen to evade each or all of these elements: the sender is not me but the spirits that possess me; the content is garbled beyond understanding; the receiver is not really the doctor but a spy; or the context is another situation, time, or place.

This principle can be extended to normal communication as well. A message need not be bizarre to be classified as disqualification. Suppose a student asks a professor about his probable grade in her course. The preface “Our department’s policy is ...” evades what the professor’s own opinion is. The phrasing “On the one hand ... but on the other hand ...” often renders the ultimate content of the message unclear; jargon can serve the same function. The response “Many students have diffi-
culty with this course” leaves ambiguous whether the receiver in particular is being included. Finally, answering a slightly different question to the one asked is a common (and often welcome) means of evading the immediate context, namely, the question one does not wish to answer.

The assumption upon which our work is based is that all disqualification involves one (or more) of these four elements. That is, all disqualification renders at least somewhat ambiguous the sender’s own opinion, exactly what is being said, the fact that the message is addressed to the recipient, the context in which the message is sent, or some combination of these. Essentially, our measurement method asked judges to tell us how clear each of these four elements was for a particular message. Since this is a fundamental assumption in the present work, it is important to state it explicitly: There are no messages that would intuitively or otherwise be called disqualifications that do not involve ambiguity in one of the four basic elements. Therefore, if the clarity of these four elements is assessed, then disqualifications will be identified. The term disqualification was never used with judges; nor were any related technical terms or categories used (incongruence, evasion, change of levels, etc.). Rather, we asked them to focus on the following four questions, in this order:

1. How clear is this message, in terms of just what is being said? [content]
2. To what extent is this message the writer’s own opinion? [sender]
3. To what extent is the message addressed to the other person in the situation? [receiver]
4. To what extent is this a direct answer to the (implicit or explicit) question? [context]

Applying these questions to the examples given earlier should illustrate: Examples A and B are unclear in content; B is also ambiguous about the sender’s own opinion (of her son’s chances); and C changes context by addressing a slightly different issue. There are two differences between this approach and that of others to similar phenomena (e.g., Brown & Levinson, 1978; Grice, 1975; Sliuzki et al., 1967; and Wiener & Mehrabian, 1968). First, the earlier approaches have been taxonomic in the sense that they generate a series of categories that describe techniques of disqualification, and so forth. We have attempted to state a principle by which disqualification operates, bypassing qualitative classification. Second, categorization tends subtly to direct attention to details of the properties of messages rather than to their effects on recipients. That is, one might also wish to know what a receiver (in the person of lay judges) is able to make of the message, rather than how it might be classified. Obviously, we are not condemning “expert” analyses of communication, but only exploring a less-used alternative; the two are undoubtedly complementary.

SCALING QUANTITY AS PHYSICAL LENGTH

A second goal was to measure degree of disqualification. Suppose now that we have drawn our subjects’ (judges’) attention to the above four basic elements of a message and want them to tell us how clear each element is in a given message. The usual approach would be to adopt a standard rating scale, for example, a seven-point Likert scale, with verbal labels for degrees of clarity corresponding to each number. Although such scaling techniques are convenient for the experimenter (in the sense that they yield data directly in numerical form), they are typically inconvenient for the subjects who must use them.

The problem is that the task rapidly becomes complex and unnatural for the subject, who must pay careful attention to stimuli and dimensions and also use an unfamiliar and restrictive means of expression. In such a situation, errors and even arbitrary responses are likely to arise. It is as if subjects were taught a little of a foreign language and then asked to respond to complex questions in this, rather than their native, language. If the primary
goal is to have the subject tell us where a given stimulus lies on a specified dimension, then the communication of that judgment should be made as natural and convenient as possible—for the subject.

Fortunately, recent advances in basic research on how people make judgments about stimuli (psychophysics) have implications for the way to make the practical process of judgment easier for the subject. To the extent that we know how people naturally tend to make complex judgments, we have a basis for setting up a natural, rather than an intrusive or arbitrary, method for letting them do so. As it turns out, this results in a method that is remarkably simple and undoubtedly not original with us.

Mathematical psychophysicists from Stevens (1966) to Levine (1974) have proposed that the process of making any judgment is a matching, mapping, or analogue process. The subject matches his or her impression of the quantitative value of the stimulus to the appropriate scale point. In other words, he or she maps the dimension of interest onto a useful set, say, the number set, or uses some other analogue for the dimension to be scaled. Levine’s (1974) proposal (see especially pp. 203-208) is that the most natural analogue is simple geometric length. Any quantitative attribute of a stimulus is transformed into a continuous quantity, such as length. Levine argues that even conventional numbers are encoded as length; it is these lengths that are manipulated and compared and then translated back into numbers for the required response. Subjects are seldom cognizant of this complex, learned mental geometry, any more than a skilled bicycle rider is aware of the complexity of that task. But the data suggest that “analogue mental arithmetic” or “geometric projection” is the most likely process by which humans make extremely complex judgments—for example, adding sums in their heads, finding a light that is exactly half as bright as the stimulus light, or adjusting a light until it is as bright as a sound is loud (cross-modality matching).

If the above is true, then a practical application would be to capitalize on what is, in effect, a natural language for expressing quantitative judgments: physical length. We therefore made our four dimensions concrete and physical. Each is represented as an actual line, on which the stimulus (message) can simply be placed. Exactly where the message is placed indicates how much of the property it has. The line does not contain any numbers or intervals, and only the end points are labeled. The numbers are measured off later by the experimenter and then manipulated to produce desirable psychometric properties. But the subject is not concerned with these refinements and is therefore free to concentrate on the task at hand, namely, the messages and the dimensions.

Over the past few years, an informal research group at the University of Victoria has been using this method of an extended line on which the actual stimuli could be placed for various projects—for example, Lee (1977), MacGregor (1975), Schaefer (1979), and Smith (1979). Various abstract, nonphysical properties of stimuli have been reliably ratio-scaled in this way. Indeed, the more complex the stimulus, dimensions, or both are, the more necessary a simple response format becomes.

THE SCALING METHOD

Overview

We will summarize here the training phase of our method as replicated with five consecutive sets of judges, trained by six different experimenters. This training is intended to be procedural, not substantive, aimed only at explaining what should be focused on in this situation. That is, we “train” what to look at; the judge tells us what he or she sees. In general, we aim to make the dimensions clear and concrete for the judge, to make the differences among dimensions apparent, to keep the judge constantly close to the actual message, and to emphasize that his or her own judgment is what is wanted—that there is no “right” answer. Finally, we train each judge in rather slow, small steps, rather than overwhelm them with information at the beginning.
FIGURE 1
Summary of the Four Dimensions

1. a. How clear is this message, in terms of just what is being said?
   b. COMPLETELY CLEAR
      direct, easy to understand;
      there is only one possible meaning.
   c. "I have the flu."
   d. Between these two extremes would go messages that are somewhat vague,
      ambiguously phrased, or hard to make out exactly; they may have several
      possible meanings or be contradictory.

2. a. To what extent is this message the writer's own opinion?
   b. DEFINITELY
      it is very clear that the message
      is his opinion; it is obviously
      his personal opinion, no one else's.
   c. "I'm sure I'm coming down with
      the flu."
   d. Whatever the message is, and no matter how clear or unclear it is on the
      first dimension, does the writer state his own opinion or not? Or is it
      something in between...you can't tell for sure if it's his or someone else's.

3. a. To what extent is the message addressed to the other person in the situation?
   (That person is eg. Harry.)
   b. DEFINITELY
      clearly addressing the other person
      in the situation; could be addressing
      only that person.
   c. "I'm fine, thank you, Harry."
   d. If from the message you can infer it's addressing the other person in the
      situation, but you could also infer it's addressing anyone or someone else—
      then it's not obviously addressing that one person and so can't be on the
      far left. You decide anywhere else.

4. a. To what extent is this a direct answer to the question?
   (The situation and question are eg. "How are you?")
   b. DEFINITELY
      this is a direct answer to the
      question
   c. "I am lousy."
   d. In other words, is this a good answer to the question given, or is there
      another question it would be a better answer to?

Essentially, the experimenter defines and explains the four dimensions to the judge (see summary in Figure 1) and then asks him or her to place written messages on a line so as to represent "how much" of each dimension the message has. Note that each dimension is ultimately defined by (a) a question, (b) two endpoint definitions, (c) two endpoint examples, and (d) a "rule of thumb."
The order of dimensions, shown in Figure 1, was dictated by the following considerations: the content dimension is judged first, in order to focus the judge's attention firmly on concrete messages, rather than on the intentions of the writer or other inferences. Second, we want to capture "first impressions" of the clarity of content, without the other three aspects, each of which is also necessarily based largely on content as well. Third, we add extra information only as necessary for specific dimensions. Thus, the judge places messages on Dimension 1 (content) in a virtual vacuum, "as if you found it lying in a desk." Little additional information is needed for Dimension 2 (writer); in some cases, the writer's status, name, or role might be relevant and would be given. Dimension 3 (receiver) requires at least the name of the person to whom the message is addressed and, in particular cases, the role relationship between the writer and receiver. Dimension 4 requires a general statement of a situation, with emphasis on a particular preceding question. (For example, Aunt Mary writes to Niece Jane, who usually visits her over the holidays: "Please let me know right away whether you're coming for Christmas.")

Note that, following Sluzki et al. (1967), we have equated context with the preceding message and have, further, made this message a question, actual or implied. This approach is congruent with the conversational principle of the "connectedness of utterances" (Nofsinger, 1976, p. 173), specifically that a question requires an answer and thereby constrains the options of the respondent. Judges are asked to focus on how well the answer fits the question. Suppose, for example, the question is "How was my performance?" and the answer is "Too bad you didn't have more time to practice." This reply is not totally unrelated to the question; most people can see it as somewhat responsive. However, it actually answers a question different from the one asked and leaves the direct answer ("Your performance was poor") to the judge (and the questioner) to fill in by implication. (Grice, 1975, Nofsinger, 1976, and others have hypothesized some of the "missing links" that may connect the overt question and answer.) Since such indirect replies are quite common, they might be seen as a conventional response. However, the judge is asked to use a different criterion: Has the question itself been answered?

It should also be noted that Dimension 2 refers only to the writer's opinion or responsibility for the message, not for his or her actions. It is possible that people disqualify their own behaviors other than overt communication, but we are not at present measuring this phenomenon. (The difference can be illustrated as "The devil made me do it" versus "The devil made me say it.")

Judges

Our judges were 41 undergraduates at the University of Victoria, in groups of 7 to 11 over three years. None had any background in communication or scaling, nor any special knowledge of the research project. They were heterogeneous in age, year at the university, academic major, background, and so on. When they were recruited, judges were told that this was an ongoing project paying $2.50 an hour for participation, and they understood that they would be coming back for several sessions.

Materials

Four strips of masking tape are laid the full length of a small (60.5 × 101.5 cm) table, as far apart as the width of the table permits. A metric tape measure is used to measure the judges' placement of the messages.

The messages are typed on cards approximately 3 × 7.5 cm; that is, four messages are typed on a vertical 3 × 5-inch index card and then cut apart. At the center top of each message is a light, penciled arrow.

Messages

Three different kinds of messages have been used. For training, four or five sets of three to 14 messages each were written to illustrate the dimensions, to draw attention to details, to present par-
ticular problems, and to assess reliability. These messages are often unusual ones and illustrate the scope of conceivable, rather than naturalistic, messages.

Second, the judges have scaled messages written by the investigators as plausible responses to experimental situations presented to other subjects (not the judges); for example, those in Table 3, below. The situations and messages were designed to explore the hypothesis described at the outset, that disqualification is preferred in certain situations. Each set of three or four messages was intended to include at least one message that was more disqualified than the others. The judges' scaling was the independent and final measure of the degree of disqualification in both these and the subject-written messages.

Finally, other subjects (again, not the judges) have written their own brief replies to experimental situations. Sets of up to 20 different messages, each written by a different subject, have been scaled together, on each dimension. These messages are typically longer and more complex than the experimenter-written replies. However, given sufficient experience with the first two kinds of messages, the judges can handle this kind remarkably easily. (We are currently working with a fourth kind of message, namely, spoken messages, with the goal of scaling subject-spoken responses, including their naturally occurring paralinguistic aspects.)

Procedure

The judges are seen individually and told that this is a scaling project in which they will be scaling written messages on various characteristics. In the first session, the four dimensions or characteristics are explained to the judge in stages. Initially, a judge is given a summary sheet of the dimensions for reference; this is based on Figure 1 but has only the definitions of the dimensions and endpoints (a and b). The experimenter reviews this sheet with the judge and adds the endpoint messages (c) as examples. All subsequent messages are placed by the judge.

With Trial 1 the judge learns about the dimensions, endpoint definitions, and endpoint examples. During the scaling of Trials 2 and 3, further information and the "rules of thumb" (d) are given. After each trial the placements of the messages are recorded on the masking tape using a key word from the message.

At the beginning of the second session, the judge is given the full summary sheet (Figure 1), with the definition of the dimensions, the endpoint definition, endpoint examples, and "rule of thumb," and is asked to rescale the Trial 1 messages. The next set of messages, Trial 4, was designed as a test of reliability; it includes a large set of messages, each varying independently on all four dimensions and covering every "trick" of phrasing we could produce. By the end of the second session, the judges are fairly familiar with the system. For the third and subsequent sessions, we simply give them a sheet exactly corresponding to Figure 1, go over it briefly, and let them proceed with few additional instructions.

Monitoring by the Experimenter

An important part of the procedure is the relation between experimenter and judge. In order to insure that the judge understands the dimensions, each is asked in effect to think aloud about why the messages are being placed at particular points. If the reason indicates that the judge does not yet understand the dimension or is confusing two of the dimensions, then further explanation or examples, or both, are given. The experimenter's skill and attention are crucial here, focusing on why, not where. The judge must learn never to expect or infer from the experimenter what the "right" judgment is. That is the judge's task and area of expertise; the experimenter's task is to explain the basics, then to monitor and assist procedurally. By the third scaling session, this task consists mainly of being an attentive listener and making sure that the judge always has a reason for his/her placement.

Such a procedure might seem vulnerable to "experimenter bias," but we have several reasons for discounting this possibility. First, the place-
### TABLE 1  
Reliability Within Groups of Judges

<table>
<thead>
<tr>
<th>Group</th>
<th>No. of Judges</th>
<th>Scaling Session</th>
<th>Trial</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>I</td>
<td>2</td>
<td>.93</td>
<td>.92</td>
<td>.79</td>
<td>.88</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
<td>I</td>
<td>3</td>
<td>.92</td>
<td>.99</td>
<td>.92</td>
<td>.97</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>II</td>
<td>4</td>
<td>.94</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>I</td>
<td>2</td>
<td>.94</td>
<td>.91</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>I</td>
<td>3</td>
<td>.94</td>
<td>.99</td>
<td>.91</td>
<td>.98</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>II</td>
<td>4</td>
<td>.90</td>
<td>.94</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>II</td>
<td>4</td>
<td>.93</td>
<td>.97</td>
<td>.98</td>
<td>.94</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>I</td>
<td>2</td>
<td>.94</td>
<td>.85</td>
<td>.99</td>
<td>.97</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>I</td>
<td>3</td>
<td>.92</td>
<td>.91</td>
<td>.82</td>
<td>.97</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>II</td>
<td>4</td>
<td>.91</td>
<td>.95</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>E</td>
<td>11</td>
<td>I</td>
<td>2</td>
<td>.96</td>
<td>.88</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>E</td>
<td>11</td>
<td>I</td>
<td>3</td>
<td>.95</td>
<td>.96</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td>E</td>
<td>11</td>
<td>II</td>
<td>4</td>
<td>.96</td>
<td>.98</td>
<td>.98</td>
<td>.98</td>
</tr>
</tbody>
</table>

---

1. Group C consisted of four new judges plus four returning from Group B. One judge in E had been in D.
2. One judge in Group A and two in Group D were unable to return for the second scaling session.
3. Each trial is a different set of messages; there are 7 messages each in Trials 2 and 3, and 14 in Trial 4.

...ments that the experimenter sees the judge making are not the final message scores, which are obtained later by transformation, as described below. It would be very hard to guess what the transformed score will be, so it is unlikely that the experimenter knows at the time whether judges are agreeing with each other or what the actual scale value will be for a given message. Second, the final scale values for many messages have been different from our initial intuitive "expert" judg-
ments—and we end up agreeing with the judges, rather than the other way around. Finally, we have apparently been quite successful in encouraging our judges' independence. They are quite confident and are more likely, in cases where the experimenter finds the reason unconvincing, to argue the experimenter down rather than to reconsider.

RESULTS: RELIABILITY

Transformation of Raw Scores

The raw data from the procedure just described are lengths, in centimeters, for each judge, dimension, and message. Each dimension yields a matrix of \( n \) judges by \( m \) messages, with a particular entry having the following meaning: when scaling the messages on Dimension 1, the first judge placed this message at 10 cm on a scale that ranged from 1 to 101 cm. The reliability of these raw-score values is encouraging but not high by rigorous standards.

The problem with the raw scores is a familiar one in the use of judgments, namely, the judges appear to use systematic, individual frames of reference; see Winer (1962, pp. 128-130). On a given dimension, one judge clusters his or her judgments around another slightly different point, but in the same relative order. This is not surprising, since the line between the two endpoints is entirely unlabelled. Rather than section the line in order to get judges to use it in exactly the same way, we chose a less intrusive procedure and transformed the scores after the judges were finished.

Our statistical transformation is an extension of Winer's solution to the frame-of-reference problem. Winer suggested that each judge's set of decisions be considered relative to its own mean. We extended this to include, by analogy to test-score standardization, the judge's standard deviation as well. The procedure is sometimes called "ipsatizing" scores, that is, standardizing a judge's scores relative to the self, rather than to a group. The mean and standard deviation in centimeters is calculated for each set of interdependent scores, say, the set of all the Trial 1 messages for Dimension 1 as placed by the first judge. Then each of these raw scores, \( X \), is transformed as follows: \( (X - \bar{X}) + SD = Z \), the ipsatized message score for that judge.

The ipsatized scores have the following meaning and properties. First, positive values indicate that the message was more disqualified or ambiguous than the mean of messages in the set. Negative values indicate relative clarity. Second, the numbers are completely comparable across judges. A score of +1.5 means that the judge placed the message one and a half of his or her standard deviations above the mean of this set. (It is this property that distinguishes ipsatized scores from simple rank ordering.)

A third and equally important point about the ipsatized scores is that they are relative, not absolute, numbers. They indicate how unclear or disqualified a given message was judged to be, on a given dimension, compared to the other messages in the set with which it was scaled and ipsatized. Therefore, inspection of the raw scores is desirable in interpreting and comparing \( Z \) scores across sets.

Interjudge Agreement

We assessed the reliability among judges' message values by the intraclass correlation coefficient, \( R \); see Winer (1962, pp. 124-128). The results are summarized in Table 1.

The reliability of the ipsatized values started high and got even better as the judges learned the system. We are confident that judges can maintain reliabilities of around .95, unless the message set is homogeneous on the dimension (curtailment of range).

The final score for a given message is the average across a given group of judges for that message on that dimension. Since we have had several successive groups, the reliability of these values can be assessed by simple correlation; see Table 2. Thus, through successive replications of the procedure, virtually identical message scores have been obtained.
TABLE 2
Reliability Across Groups of Judges

<table>
<thead>
<tr>
<th>Group</th>
<th>Experimenter</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B5</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>D</td>
<td>NC &amp; SC</td>
<td>93</td>
<td>95</td>
<td>99</td>
<td>95</td>
<td>93</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>94</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>94</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
</tr>
<tr>
<td>E</td>
<td>NC &amp; JM</td>
<td>93</td>
<td>95</td>
<td>99</td>
<td>95</td>
<td>92</td>
<td>97</td>
<td>99</td>
<td>94</td>
<td>91</td>
<td>96</td>
<td>99</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>94</td>
<td>91</td>
<td>96</td>
<td>99</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93</td>
<td>95</td>
</tr>
</tbody>
</table>

Note: Based on Trial 4 (14 messages).

Sample Message Scores

The last step is to average across judges for each message and dimension. Some examples of message scores are given in Table 3. For this set, the other person was "a fellow student in a class with the writer (of the message to be scaled)." The situation and question were as follows: the other student gave a class presentation and afterwards passed a note asking, "How did I do?" Note the subtle aspects of messages reflected in their placement. On Dimension 1, the third message is relatively unclear; it is a laconic and somewhat contradictory statement that gives criticism and then takes it away. On Dimension 2, the writer's own opinion is conspicuously missing in the same message; it is not clear who thinks "not well"—is the writer only reporting the class's reaction, or is he expressing his own? The second message also lacks a clear sender-commitment, but to a lesser degree. On Dimension 3, the third message is mildly unclear about receiver; the missing "you" is always noted by judges (vs. "You did not do well..." or "...you shouldn't feel bad about it," both of which would have more personal contact and commitment). On Dimension 4 it is the fourth message that emerges with a very high positive score, as a classic evasion of the issue. The first three messages were judged to be relatively more responsive to the question. Note that the meaning of the message (praise or criticism) makes little difference, because the first two messages are scaled very similarly except for a real difference in identification of the sender (D2). Clearly, the third and fourth messages are more disqualified, although in different ways.

DISCUSSION

Our technical goals were to translate Haley's (1959a) definition of disqualification into a specifiable and replicable psychometric procedure that would enable nonexpert judges to indicate with high reliability the degree of disqualification in brief, written messages.

The results described here are quite consistent with these goals; indeed, the ease of replication and very high reliabilities are more than we initially dared hope for.

Our broader goals include measurement that is consistent with our overall theory that disqualification involves properties of messages that are of pragmatic importance, in two senses: first, that naive communicators (receivers) can both perceive and make fine discriminations about these properties; and, second, that these properties will characterize messages chosen by senders in certain social contexts. Underlying all this is the assumption that disqualification is a normal and frequent communicational occurrence that can be elicited and understood as an interpersonal, situational (vs. intrapsychic or pathological) event.
TABLE 3
Sample Message Values: Trial 5

<table>
<thead>
<tr>
<th>Message</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. &quot;You did very well. I really liked it.&quot;</td>
<td>-.35 (.45)</td>
<td>-.78 (.45)</td>
<td>-.38 (2.71)</td>
<td>-.63 (4.38)</td>
<td>-2.14 (40.51)</td>
</tr>
<tr>
<td>2. &quot;You were terrible, bad job.&quot;</td>
<td>-.50 (.95)</td>
<td>.10 (18.88)</td>
<td>-.32 (28.38)</td>
<td>-.58 (6.13)</td>
<td>-1.30 (62.64)</td>
</tr>
<tr>
<td>3. &quot;Not well, but don't feel bad about it.&quot;</td>
<td>1.02 (38.13)</td>
<td>1.24 (38.25)</td>
<td>.79 (45.13)</td>
<td>-.21 (15.)</td>
<td>2.84 (136.51)</td>
</tr>
<tr>
<td>4. &quot;You were braver than I would be!&quot;</td>
<td>.02 (13.88)</td>
<td>-.56 (8.75)</td>
<td>-.09 (33.63)</td>
<td>1.42 (39.75)</td>
<td>.79 (96.01)</td>
</tr>
</tbody>
</table>

Both ipsatized and raw scores (in parentheses) are given. Recall that negative ipsatized scores indicate relative clarity, and positive scores mean relative ambiguity on the dimension. The scores should be read by column, i.e., across messages, which is how they were scaled and ipsatized.

We did not intend to develop a method encompassing other possible definitions of disqualification, much less other related phenomena. It would be imperialism to attempt to subsume all neighboring concepts. Nor was the goal to replace expert analysis of messages with lay impressions; these are quite different research strategies. Finally, no effort was made to accommodate extended naturalistic or face-to-face communication exchanges. Not surprisingly, then, the method in present form cannot do any of these things.

The reader may have already considered possible modifications to suit these other goals. A sequence of dyadic transactions could be scaled now by experienced judges, if the total number of messages were constrained to, say, 20 or so, which could be scaled as a set. The first dimension, clarity of content, presents no problems, as the messages are self-contained for this dimension. The alternation of source and receiver would have to be kept in mind for Dimensions 2 and 3, making the job somewhat more complex. Dimension 4 presents the more formidable problem, in that each message would become the context for the next. This would require that the judges be able to indicate that message B is more responsive to message A than C is to B—a difficult but not impossible task. Another alternative is to change to expert judges doing traditional content analysis with this system as a guideline. The advantage would be the considerable savings in judges' time and the time spent ipsatizing. Such a modification would move the method from the "naive other" to "external expert" perspective (Wilmot, 1980). A possible disadvantage would lie in our impression that the present number of judges, as well as their independence, result in a healthy and fresh diversity of opinion among them, which might not be true of a small group of experts working closely for a long time.

With respect to the method in present form, the obvious question, "... but is it valid?" has been answered affirmatively, at least in part, by the results described here. The construct validity of a measure is classically established by hypothesis-testing procedures (Cronbach & Meehl, 1955). In
effect, this means establishing whether the measure acts, empirically, as a measure of this construct should act; that is, whether it yields results consistent with its theoretical framework. Thus, the attainment of technical goals, including reliability, does contribute to construct validity (Campbell & Fiske, 1959). We are continuing to explore the broader theory by manipulating the characteristcs of situations and asking subjects to choose or write responses to them. One series of five experiments, to be described in a subsequent paper, shows that when more direct communications would be, for example, tactless or dishonest, then the amount of disqualification increases. In general, disqualification seems to arise in avoidance-avoidance conflicts, where it is the communicational equivalent of "leaving the field."

Thus, the method is proving to be construct valid in the context of our approach. Whether it would be the relevant and valid measure for other conceptual approaches is an open issue. With regard to this problem and the procedure as a whole, we want to emphasize that nothing is cast in stone, immutable. We have presented our approach with special attention to its assumptions and development precisely so that others might question or change these. We are as interested in presenting a way of approaching the problem, one that others might modify to suit their own needs or assumptions, as we are in describing a particular solution to the measurement problem we faced.

NOTES

This work has been generously supported by research grants from the Social Sciences and Humanities Research Council of Canada. We would also like to acknowledge our experimenters, Deborah Bolt, Susan Cathro, Nicole Chovil, Jennifer Mullett, and Diane Russell.

1. A supplementary report containing procedural details sufficient for replication is available from the first author.

2. Note that R is not the average correlation between judges, which is typically much lower, and which is irrelevant when one plans to average across judges (Ebel, 1951).

REFERENCES


HALEY, J. An interactional description of schizophrenia. Psychiatry, 1959, 22, 321-332. (a)

HALEY, J. The family of the schizophrenic: a model system. Journal of Nervous and Mental Diseases, 1959, 129, 357-374. (b)


SLUZKI, C.E., BEAVIN, J., TARNOPOLSKY, A., &


