14. H. J. Tagnon, W. F. Whitmore, Jr., N. R. Shulman, *Cancer* **5**, 9 (1952).
15. E. C. Dodds, L. Golberg, W. Lawson, R. Robinson, *Proc. Roy. Soc. London, Ser. B* **127**, 140 (1939).
16. C. Huggins, R. E. Stevens, Jr., C. V. Hodges, *Arch. Surg.* **43**, 209 (1941).
17. G. T. Beatson, *Lancet* **1896-II**, 104, 162 (1896).
18. J. H. Farrow and F. E. Adair, *Science* **95**, 654 (1942).
19. C. Huggins and D. M. Bergenstal, *Cancer Res.* **12**, 134 (1952).
20. G. W. Woolley, E. Fekete, C. C. Little, *Proc. Nat. Acad. Sci. U.S.* **25**, 277 (1939).
21. R. Luft, H. Olivecrona, B. Sjögren, *Nord Med.* **47**, 351 (1952).
22. A. Haddow, J. M. Watkinson, E. Paterson, *Brit. Med. J.* **2**, 393 (1944).
23. A. Haddow, *J. Pathol. Bacteriol.* **47**, 553 (1938); F. Bielschowsky, *Brit. Med. Bull.* **4**, 382 (1947); L. Foulds, *Brit. J. Cancer* **3**, 345 (1949).
24. O. Mühlbock, in *Endocrine Aspects of Breast Cancer*, A. R. Currie, Ed. (Livingstone, Edinburgh, 1958), p. 291.
25. C. B. Huggins, E. Ford, E. V. Jensen, *Science* **147**, 1153 (1965).
26. L. A. Loeb and H. V. Gelboin, *Proc. Nat. Acad. Sci. U.S.* **52**, 1219 (1964).
27. H. G. Williams-Ashman and C. Huggins, *Med. Exp.* **4**, 223 (1961).
28. J. C. Arcos, A. H. Conney, N. P. Buu-Hoi, *J. Biol. Chem.* **236**, 1291 (1961).
29. C. Huggins and R. Fukunishi, *J. Exp. Med.* **119**, 923 (1964).
30. J. Maisin and M.-L. Coolen, *Compt. Rend. Soc. Biol.* **123**, 159 (1936).
31. H. Shay, *et al.*, *J. Nat. Cancer Inst.* **10**, 255 (1949).
32. C. Huggins and N. C. Yang, *Science* **137**, 257 (1962).
33. C. Huggins, L. C. Grand, F. P. Brillantes, *Nature* **189**, 204 (1961).
34. R. P. Geyer, J. E. Bryant, V. R. Bleisch, E. M. Peirce, F. J. Stare, *Cancer Res.* **13**, 503 (1953); C. Huggins, S. Morii, L. C. Grand, *Ann. Surg.* **154** (Suppl.) 315 (1961).
35. C. Huggins, L. Grand, R. Fukunishi, *Proc. Nat. Acad. Sci. U.S.* **51**, 737 (1964).
36. P. Rous, *J. Exp. Med.* **13**, 397 (1911).
37. T. L. Dao, *Progr. Exp. Tumor Res.* **5**, 157 (1964).
38. E. D. Rees and C. Huggins, *Cancer Res.* **20**, 963 (1960).
39. O. Warburg, *Metabolism of Tumours* (Constable, London, 1930).
40. J. G. Hamilton, P. W. Durbin, M. Parrott, *J. Clin. Endocrinol.* **14**, 1161 (1954); C. J. Shellabarger, E. P. Cronkite, V. P. Bond, S. W. Lippincott, *Radiation Res.* **6**, 501 (1957).
41. E. Ford and C. Huggins, *J. Exp. Med.* **118**, 27 (1963).
42. C. Regaud and J. Blanc, *Compt. Rend. Soc. Biol.* **58**, 163 (1906).
43. C. Huggins and L. C. Grand, *Cancer Res.* **26**, 2255 (1966).
44. C. Huggins, G. Briziarelli, H. Sutton, Jr., *J. Exp. Med.* **109**, 25 (1959).
45. C. Huggins and R. Fukunishi, *Radiation Res.* **20**, 493 (1963).
46. C. Huggins, R. C. Moon, S. Morii, *Proc. Nat. Acad. Sci. U.S.* **48**, 379 (1962).
47. R. L. Landau, E. N. Ehrlich, C. Huggins, *J. Amer. Med. Ass.* **182**, 632 (1962); L. G. Crowley and I. Macdonald, *Cancer* **18**, 436 (1965); B. J. Kennedy, *ibid.*, p. 1551.
48. F. R. Heilman and E. C. Kendall, *Endocrinology* **34**, 416 (1944).
49. J. B. Murphy and E. Sturm, *Science* **98**, 568 (1943).
50. O. H. Pearson, L. P. Eliel, R. W. Rawson, K. Dobriner, C. P. Rhoads, *Cancer* **2**, 943 (1949).
51. T. F. Dougherty and A. White, *Proc. Soc. Exp. Biol. Med.* **55**, 132 (1943).
52. H. W. Balme, *Lancet* **1954-I**, 812 (1954); G. Crile, Jr., *J. Amer. Med. Ass.* **195**, 721 (1966).
53. H. Kirkman, *Nat. Cancer Inst. Monogr.* **1**, 1 (1959).
54. H. J. G. Bloom, C. E. Dukes, B. C. V. Mitchley, *Brit, J. Cancer* **17**, 611 (1963).
55. R. M. Kelley and W. H. Baker, *New Engl. J. Med.* **264**, 216 (1961).
56. O. S. Rodriguez Kees, *J. Urol.* **91**, 665 (1964).
57. H. Kirkman and F. T. Algard, *Cancer Res.* **24**, 1569 (1964).
58. S. W. Nielsen and J. Aftsomis, *J. Amer. Vet. Med. Ass.* **144**, 127 (1964).
59. This investigation was aided by grants from the Jane Coffin Childs Memorial Fund for Medical Research and the American Cancer Society. The Upjohn Company, Kalamazoo, Michigan, through P. Schurr, provided the lipid emulsions of hydrocarbons.

# Scaling Data on Inter-Nation Action

A standard scale is developed for comparing international conflict in a variety of situations.

Lincoln E. Moses, Richard A. Brody, Ole R. Holsti, Joseph B. Kadane, Jeffrey S. Milstein

In empirical research it is helpful to be able to quantify the things being studied. In the physical sciences the development of suitable units of measurement for various phenomena has been part and parcel of the progress of these disciplines. In the social sciences the development of scales for "measuring" intelligence and various personality traits has been of great benefit in the posing, sharpening, and testing of hypotheses and the studying of relationships between qualities which vary in intensity. Such variables are difficult to "measure." It is easy enough to recognize that there are variations in the degree of such qualities as beauty, leadership, hostility, aggression, and cheerfulness. Study of such variables is greatly facilitated if they can be reasonably quantified.

One approach to quantifying such variables is the method of ranking various specimens with respect to the attribute in question. Thus, several trained judges may agree well, in independent ordering of several individuals interviewed, as to the degree of cooperativeness which they manifest, or several judges may agree well in ordering a group of bathing beauties as to pulchritude. Where reliable judgments of this kind are obtainable, the ranks within the group serve as quantitative indices of the quality being studied. Where there are very many specimens—say 100—it may be convenient to use a coarse ranking into nine categories, requiring the judges to put 5 percent in the top group, 8 percent in the next-to-top group, 12 percent in the third group, and so forth (*1*). Such a coarse ranking, because it resembles the judging task in Q-sort technique, we here call a Q-sort.

The Institute of Political Studies at Stanford has for some years been studying international political crises. The official statements made by political figures during crises under study have been scored by judges, in a forced approximate ranking of the Q-sort type, with respect to variables such as hostility, frustration, friendship, and satisfaction. Further, statements descriptive of the actions of the states involved in the crisis have been assembled, coded so as to mask the identities of the actors in the crisis, and judged in the same way for such variables as violence and conflict. The approximate ranks obtained by means of these procedures have permitted study of such questions as which nation evidenced the most hostility in the crisis, how concordant or disparate were the hostility of the verbal messages and the aggressiveness of the acts of the various participating nations, and what were the time trends of hostile statements and violent actions as the crisis unfolded (*2*).

Dr. Moses is professor of statistics, Stanford University, Stanford, California; Dr. Brody is associate professor of political science at Stanford; Dr. Holsti is associate professor of political science at the University of British Columbia, Vancouver; Dr. Kadane is assistant professor of statistics at Yale University, New Haven, Connecticut; Mr. Milstein is a graduate student in political science at Stanford.

From time to time it has seemed desirable to be able to compare the messages or the acts of a given country in two or more different crises in which it was a participant. The scores obtained by the forced-category ranking (or, for that matter, by a straight rank-ordering) do not permit such comparison. In such ranking, whatever the average intensity may be for a body of specimens, 5 percent of the specimens will be given the highest score, 8 percent the next highest, and so on (or ranks beginning with the integer 1 will be given in each case). Such scores, then, cannot reflect differences in general level of intensity for two separately judged bodies of data.

In this article we present a method which has been developed to permit comparison of variables such as aggression and hostility for separately judged situations. The methods should be useful in any branch of social science where numerical scores are best assigned by ranking all specimens (or a sample of specimens) within a domain of discourse.

The principle can be illustrated by the following not very serious example. Suppose that, in each of several countries, beauty contests are held. These would serve to identify the prettiest contestant in each country and would leave unanswered the question, Which country had the prettiest contestants? Application of the method proposed in this article would lead to the formation of a "panel" of beauties, the panel to have two properties: (i) the property that independent judges agree well on the ordering of the panel members with respect to beauty, and (ii) the property that the range of pulchritude stretches from very low to very high, ideally spanning the full range to be found in a beauty contest in any country. This panel of beauties (not identified as panel members) would then be entered in each competition in each country. This would permit inter-country comparisons because, for example, the beauty of the winner in any country could be "measured" in terms of the number of panel entrants she was judged to exceed in pulchritude. This of course would permit assignment of average scores for the entrants in each country, and this would permit inter-country comparisons. If the panel could also consist of members who were ageless, the method would even permit comparison on this important variable, beauty, across time.

## Procedures

In the study discussed in this article, two sets of statements concerning international actions were used. The first set concerned actual historical events. These events are the "things to be measured," and statements concerning them are entered on cards called "data cards." (In the study discussed here, the events were drawn from the July-August 1914 crisis that led to World War I.) In these statements of real events, the names of the actual nations or individuals are masked and the statements are phrased in the general form "Nation A does X to (or with) nation B." Thus, a data card for the event "Germany declares war on France" appears on a card as "Nation A declares war on nation B." This form was adopted to emphasize the actions rather than the actors.

The second set of statements consists of statements covering a wide range of possible inter-nation actions. These statements, entered on cards called "marker cards," are the measuring instrument. In format they are just like the statements on the data cards. For example, a marker card might read, "Nation A federates with nation B," or "Nation A executes nation B's prisoners of war."

Use of a standard set of marker cards intermixed with the data cards permits scoring of the data cards. How this is done is illustrated by the following examples. If the average data-card statement for nation A in situation 1 exceeded in hostility three-fourths of the marker-card statements and the average data-card statement for nation B in situation 2 exceeded seven-eighths of the marker-card statements, this would permit the inference that the data-card statements for nation B were more hostile than those for nation A. Each data card is given a score equal to the number of marker cards it exceeds in intensity (in this case, in hostility). This score is independent of the other data cards in the set; it depends only on the data card in question and upon the standard set of marker cards. This is the key to intersituational comparability.

A standard set or deck of marker cards should have several properties. (i) It should span a wide range of intensity, otherwise some bodies of historical material entered on data cards may lie wholly outside the range of the marker cards, making valid comparison with other bodies of historical material impossible. (ii) Marker cards must be applicable to a wide range of situations. For example, probably the word *tomahawk* would not be used in a statement on a marker card, since the word would seem bizarre in the context of World War I data and would reveal that the card in question was a marker card. (iii) Marker cards, like data cards, must be in the standard format (they should refer to countries as "nation A," "nation B," and so on, rather than by name) so that the marker cards will not be identifiable as such. (iv) A set of marker cards must be strongly reliable—that is, different judges, in evaluating them on the basis of the intensity of attitude the statements reveal, should put them in almost exactly the same order.

For the experiment reported here, the data cards were prepared as follows. A few hundred (328) actions from the 1914 crisis were noted on separate cards in the standard format. Two judges then independently "Q-sorted" these cards into nine groups according to increasing intensity of conflict. The percentages of the total number of cards in each of the nine groups were as follows: 5, 8, 12, 16, 18, 16, 12, 8, and 5 (*1*). The first group here represents least conflict and the ninth group represents most conflict. The interjudge correlation was .733. Each of the data cards was assigned the mean value of the Q-sort scores assigned by the two judges. These scores provide a basis of comparison for the scaling of these data by the use of standard marker cards.

The marker cards used to scale the data cards were selected in the following way. Five judges independently ranked a number of different candidate marker cards (191 at the first judging, 151 at the second) in terms of the intensity of conflict represented by each item. No ties were allowed in this ranking. We sought to eliminate those candidates from the prototype scale which had the most variance (*3*) in the ranking—that is, those which were least reliable.

We eventually discarded more than a third of the original candidate marker cards in order to arrive at a set of 120 marker cards with the smallest variances. The marker cards with higher variances of transformed mean scores in this set of 120 items were still concentrated in the low-conflict end of the scale. However, the variances in the re-

| Judge | Order of judging | | | |
|---|---|---|---|---|
| | 2nd | 3rd | 1st | 4th |
| 1 | A1 | B3 | C4 | D2 |
| 4 | B2 | A4 | D3 | C1 |
| 2 | C3 | D1 | A2 | B4 |
| 3 | D4 | C2 | B1 | A3 |

Fig. 1. Experimental design. *A, B, C,* and *D* are decks of 82 cards each with statements concerning historical international actions taken in 1914 (data decks); *1, 2, 3,* and *4* are decks of 30 cards each with standard statements concerning international actions (marker decks).

tained set were not nearly so high as those in the original set, and were of an acceptable level. A set of 120 marker cards was chosen because the number 120 is useful experimentally in that it can be factored in many ways.

It was thought desirable to appraise the marker cards in the following respects.

1) How reliable is a sub-deck of 30 cards?

2) Can a sub-subdeck of only 15 cards do an adequate job?

3) How reproducible are the results obtained by different judges using the marker decks?

4) How "valid" are scores obtained by use of the marker cards in comparison with the Q-sort values earlier assigned to the same data cards by means of the standard procedure of the project?

5) Can the marker cards be simply mixed with the data cards and used with the Q-sort methodology?

To answer these questions, the sets of 328 data cards and 120 marker cards were broken down into four equal decks of 82 data cards (data decks) and four equal decks of 30 marker cards (marker decks). Equality in the distribution of scores in these decks was sought by the assignment of cards to a deck by means of random permutations (*4*). In this random assignment the data cards were arranged in order of Q-sort scores and each set of four successively higher scores (representing increasing intensity of conflict) was randomly assigned to one of the four data decks. Similarly, marker cards were randomly assigned to one of the four marker decks on the basis of transformed mean scores. Thus, each subdeck spanned the whole intensity-of-conflict scale of the original deck.

Four judges (a professor and three graduate students of international relations) who had not participated in the judging involved in the preparation and selection of the experimental decks of marker and data cards, were chosen to participate in the experiment. In the experimental procedure, a judge was presented with a shuffled deck of 112 cards—82 data cards and 30 marker cards. The judge was then instructed to rank all the cards in the combined deck, on the basis of the statements of inter-nation action, in order of increasing conflict.

The experimental design used was that of a Greco-Latin square in which the judges, the order in which the combined decks were judged (first, second, third, or fourth), the set of data cards, and the set of marker cards were factors in the design. Random permutations were again used to govern the assignment of judges, the order of judging, and the selection of data decks and marker decks. The array shown in Fig. 1 is a summary of the experimental design.

In this experimental design, each judge ranked all four data decks and all four marker decks, in combinations of one data deck and one marker deck, but no two judges ranked the same data-deck, marker-deck combination. Similarly, no judge ranked either the marker decks or the data decks in the same order as any other judge. There were thus 16 separate judgments of the four combined decks of 112 cards each. That is, each of the four marker decks was ranked against each of the four data decks by four judges, in four different orders. In this way the experimental design permitted the observation of variability (i) among judges, (ii) in the order of judging (order effects), (iii) in ranking of cards in the marker decks (ideally, this variability should be zero), and (iv) in ranking

Table 1. Average scores for four data decks as measured against all four marker decks, with marker subdecks of odd-numbered cards only, and with marker subdecks of even-numbered cards only.

| Data deck | Average score | | |
|---|---|---|---|
| | All markers | Odd markers | Even markers |
| 1 | 9.27 | 4.37 | 4.89 |
| 2 | 8.62 | 4.10 | 4.52 |
| 3 | 9.54 | 4.52 | 5.01 |
| 4 | 9.50 | 4.58 | 4.92 |
| Range | .92 | .48 | .49 |

Table 2. Average scores for the entire set of 328 data cards as measured against each of the four marker decks (and against their subdecks of odd-numbered and even-numbered cards).

| Marker deck | Average score | | |
|---|---|---|---|
| | All markers | Odd markers | Even markers |
| 1 | 9.98 | 4.80 | 5.17 |
| 2 | 8.81 | 4.17 | 4.64 |
| 3 | 9.22 | 4.37 | 4.85 |
| 4 | 8.92 | 4.23 | 4.69 |
| Range | 1.17 | .63 | .53 |

Table 3. Average scores (for the entire set of data cards) for four orders of judging, with scores based on all marker decks, on marker subdecks of odd-numbered cards, and on marker subdecks of even-numbered cards.

| Order of judging | Average score | | |
|---|---|---|---|
| | All markers | Odd markers | Even markers |
| 1 | 9.55 | 4.58 | 4.97 |
| 2 | 9.11 | 4.33 | 4.79 |
| 3 | 9.56 | 4.54 | 5.02 |
| 4 | 8.70 | 4.13 | 4.57 |
| Range | .86 | .35 | .45 |

Table 4. Average scores (for the entire set of data cards) assigned by four judges on the basis of all marker decks, of marker subdecks of odd-numbered cards, and of marker subdecks of even-numbered cards.

| Judge | Average score | | |
|---|---|---|---|
| | All markers | Odd markers | Even markers |
| 1 | 9.14 | 4.38 | 4.76 |
| 2 | 9.52 | 4.51 | 5.02 |
| 3 | 8.61 | 4.08 | 4.53 |
| 4 | 9.65 | 4.61 | 5.04 |
| Range | 1.04 | .53 | .51 |

Table 5. Correlations between odd-card scores and even-card scores for 328 judged cards for each judge, each order, each data deck, and each marker deck.

| Designation | Correlation | Designation | Correlation |
|---|---|---|---|
| *Judge* | | *Data deck* | |
| 1 | .98 | 1 | .98 |
| 2 | .98 | 2 | .98 |
| 3 | .97 | 3 | .98 |
| 4 | .98 | 4 | .97 |
| *Order* | | *Marker deck* | |
| 1 | .98 | 1 | .98 |
| 2 | .98 | 2 | .98 |
| 3 | .98 | 3 | .97 |
| 4 | .97 | 4 | .98 |

of cards in the data decks (this variability should also be nearly zero).

The data from the Greco-Latin square consisted of the ordered combined deck (containing 82 data cards and 30 marker cards) for each of the 16 trials. These 16 sets of data made it possible to proceed with other investigations, as follows.

1) To study the variation in the average scores obtained when a single data deck was combined with, and ranked against, each of the four marker decks in turn.

2) To study scores obtained in scoring a data deck against (i) the 15-card subdeck of marker cards which were ordered 1st, 3rd, 5th, . . . 29th in intensity (the "odd-numbered cards") and (i) the 15-card subdeck of marker cards which were ordered 2nd, 4th, 6th, . . . 30th in intensity (the "even-numbered cards").

3) To compare the scores assigned by the various judges to cards in identical data decks.

4) To compare the scores assigned to data cards by the marker-card method with the original Q-sort scores of the data cards.

5) To condense the ranked data into nine ordered categories applicable in Q-sorting, and to give each data card a score representing the number of marker cards in lower Q-sort categories plus half the number of marker cards in its own Q-sort category.

The experiment also permits a direct appraisal of inter-judge agreement in terms of the way in which the judges order the marker cards in their rankings of the 112-card decks. This same information affords a test of the judge's competence, for the rankings should reproduce well the a priori ordering of the marker cards.

## Results

The 328 data cards were divided, as described above, into four decks intended to be exactly equal in terms of Q-scored conflict, and then each of these sets of data cards was ranked four times, once by each of the four judges (once as the first task of a judge, once as the second task, and so on, and each time in combination with a different marker deck). Each time a data deck was judged, every data card was given a score, representing the number of marker cards it was judged to exceed in degree of conflict. Thus, with a 30-card marker deck, the score for each data card could range from 0 to 30, and the average score for a set of 82 data cards could also conceivably range from 0 to 30. The overall average score for the four sets of data cards was 8.92. When the cards in the data deck were mixed with, and judged against, a 15-card subdeck of marker cards, the score for any data card could range only from 0 to 15 and the average score was 4.46. In Table 1 we see that the four data decks, intended to be exactly equal in terms of the degree of conflict represented by the actions described, as indicated by their Q scores, were judged, through mixing and comparison with the marker cards, to be essentially equal, the four data decks having average scores of 9.27, 8.62, 9.54, and 9.50, respectively. The range between the highest and lowest score is .92; this means that the averages for the four data decks differed by less than one "notch" of our 30-notch scale. Closely comparable results were obtained when the 15-card subdecks of marker cards were used.

In Table 2 are shown the average scores for the entire body of 328 data cards when these cards were mixed with and scored against cards in each of the four different marker decks. Here, again, the results are closely comparable, the average score in each case being approximately 9.0. The range is 1.17.

In Table 3 are shown the average scores for the entire body of 328 data cards when the scoring against a particular marker deck was the first, second, third, or fourth task of the judge. If order of judging makes no difference in average scores, these numbers should, ideally, be exactly equal (8.92). They are not exactly equal, but they are very close, the range being .86. There is no evident trend toward higher or lower scores as the order changes from 1 through 4, so we have good grounds for supposing that this range of .86 reflects only the inherent variation that cannot be avoided in this kind of judging task. The fact that .86 is similar to .92 and to 1.17, the ranges, respectively, of average scores for the individual data decks and the complete data deck judged against different marker decks, without regard to order of judging, encourages us to believe that the individual data decks were equivalent, as intended, and that the individual marker decks were also equivalent. Further, the variation from judge to judge (Table 4) is small, with a range of only 1.04, reflecting essential uniformity of judgment. The scores given in Tables 1 though 4 are scores obtained with the 30-card marker decks, but our conclu-

Table 6. Correlations between original Q-sort scores and scores assigned on the basis of all marker decks and marker subdecks of odd-numbered cards and of even-numbered cards, for each judge, each order, each data deck, and each marker deck.

| Designation | Correlation | | | Designation | Correlation | | |
|---|---|---|---|---|---|---|---|
| | All markers | Even markers | Odd markers | | All markers | Even markers | Odd markers |
| | *Judge* | | | | *Data deck* | | |
| 1 | .75 | .73 | .74 | 1 | .75 | .73 | .74 |
| 2 | .71 | .69 | .69 | 2 | .76 | .74 | .74 |
| 3 | .76 | .74 | .74 | 3 | .72 | .72 | .70 |
| 4 | .77 | .75 | .74 | 4 | .72 | .72 | .74 |
| | *Order* | | | | *Marker deck* | | |
| 1 | .73 | .71 | .72 | 1 | .73 | .71 | .71 |
| 2 | .77 | .75 | .74 | 2 | .73 | .71 | .71 |
| 3 | .76 | .75 | .75 | 3 | .76 | .73 | .74 |
| 4 | .73 | .70 | .70 | 4 | .77 | .77 | .76 |

Table 7. Inter-judge correlations for four data decks.

| Judge | Deck 1 Judge | | | Deck 2 Judge | | | Deck 3 Judge | | | Deck 4 Judge | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 |
| 1 | .64 | .81 | .69 | .77 | .69 | .78 | .71 | .64 | .69 | .62 | .73 | .75 |
| 2 | | .56 | .61 | | .57 | .70 | | .66 | .67 | | .68 | .67 |
| 3 | | | .79 | | | .84 | | | .71 | | | .85 |

Table 8. Correlation of judges' ranking of marker cards with the "correct" order of the cards.

| Marker-card deck | Judge | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | .98 | .88 | .91 | .93 |
| 2 | .93 | .90 | .94 | .90 |
| 3 | .93 | .83 | .96 | .96 |
| 4 | .90 | .88 | .97 | .91 |
| All | .94 | .87 | .95 | .93 |

sions are no different when only the two 15-card subdecks are used. These appear to be fully as useful as the 30-card decks.

This equal usefulness of the 15-card subdecks of marker cards is further evidenced by the very high correlations between scores obtained with odd-numbered cards and those obtained with even-numbered cards on the 328 data-card judgments made by each judge, in each order, for each data deck and with each marker deck. Table 5 shows these correlations; they are all either .97 or .98 to the number of significant figures shown.

Our conclusions at this point are (i) that the 30-card marker decks and the 15-card marker decks are closely comparable in usefulness; (ii) that they may be used interchangeably by any of our four judges; (iii) that the four data decks are equivalent; and (iv) that the ranking of cards of the data decks is not influenced by the order in which the decks are given the judges. There remains the question: Are the scores obtained by means of marker decks comparable with those obtained by means of the conventional Q-sort method? The rough quantitative answer is that correlation between marker-deck scores and original Q-sort scores is about .75. The correlation between Q-sort scores and marker-deck scores is shown in Table 6, separately for each judge, each order, each data deck, and each marker deck. Since the standard error for any number in Table 6 is approximately .06, we may ignore the slight discrepancies among the values displayed, since none of these differences is statistically significant. A correlation of .75 is not especially heartening. That is, though we have grounds for concluding that both methods of scoring are revealing the same quantity (in this case, degree of conflict), the scores are by no means directly interchangeable. Whenever two quantities are correlated and each (or either) contains a "measurement error," the true correlation between the things being measured will be underestimated. This phenomenon, known as attenuation, can be corrected for by using information about the size of the measurement errors of the two instruments. In the case of our scoring methods, both the original Q-sort scores and the scores obtained through use of the marker cards contain some "noise." In particular, the Q-scores that we used were the average values assigned by two judges whose judgments when they independently scored the entire 328 data cards in the customary Q-sort manner were not very close, the correlation between their two judgments being only .733. Further, the correlation for judgments made by the four judges by means of the marker-card method was not very different from this figure. Table 7 shows separately for each data deck the six intercorrelations among the four judges. The values range from .56 to .85. The data of Table 7 show that the intercorrelations for Judge 2 are rather consistently the poorest, and it may be that a better estimate of the reliability of the method would be obtained if the scores based on his ranking were omitted from the study. The advisability of this is further suggested by the observation (Table 8) that his judgments on the ordering of the marker cards was conspicuously less well correlated with their "correct" or a priori order than were the corresponding judgments of the other three judges. In the light of these considerations, we decided to omit scores attributable to Judge 2 in estimating the inter-judge correlation on marker-card scoring of the data decks. We then averaged the 12 inter-judge correlations of Table 7 which did not involve judgments made by Judge 2 and got an average figure of .75.

At this point we are able to correct for attenuation our correlation, $r_{XQ}$, between original Q-sort score and marker-card score. To do this we define (*5*) the two reliabilities $r_{XX}$ and $r_{QQ}$ and then estimate the attenuation-corrected correlation $r^*$ to be

$$r^* = \frac{r_{XQ}}{(r_{XX}r_{QQ})^{\frac{1}{2}}} \doteq \frac{.74}{[(.75)(.85)]^{\frac{1}{2}}} \doteq .93$$

Such a high correlation indicates that the two judging methods, Q-sort and marker-card scoring, are nearly measures of the "same thing"—presumably conflict.

It cannot be denied that the Q-sort approach may be more convenient to use than the method requiring strict ranking of marker and data cards. Thus it might be desirable to use the marker cards in a Q-sort procedure. This would involve adding a set of marker cards to a set of data cards and then going through the ordinary Q-sort procedure. Then the data cards would be scored, not in the customary Q-sort manner, but rather through assigning to each data card a score equal to the number of marker cards in lower Q-sort categories, augmented by half the number of marker cards in the Q-sort category containing the data card. Although we did not actually have our cards Q-sorted, we were able to simulate this kind of Q-sort by taking the Q-sort results that our rankings implied and then assigning Q-sort values in the way described. The correlation between these modified Q-sort scores and the original (averaged) Q-sort scores is, as in the case of the marker scores, about .75. Table 9 shows the correlations in detail for each judge, each order, each data deck, and each marker

Table 9. Correlation between original Q-sort scores and marker-card scores derived from Q-sort of data cards mixed with marker cards.

| Designation | Correlation | | | Designation | Correlation | | |
|---|---|---|---|---|---|---|---|
| | All markers | Even markers | Odd markers | | All markers | Even markers | Odd markers |
| | *Judge* | | | | *Data deck* | | |
| 1 | .76 | .76 | .76 | 1 | .75 | .75 | .75 |
| 2 | .72 | .72 | .72 | 2 | .75 | .75 | .75 |
| 3 | .75 | .75 | .75 | 3 | .72 | .72 | .72 |
| 4 | .76 | .76 | .76 | 4 | .77 | .77 | .77 |
| | *Order* | | | | *Marker deck* | | |
| 1 | .74 | .74 | .74 | 1 | .72 | .72 | .72 |
| 2 | .77 | .77 | .77 | 2 | .73 | .73 | .73 |
| 3 | .75 | .75 | .75 | 3 | .76 | .76 | .76 |
| 4 | .73 | .73 | .73 | 4 | .77 | .77 | .77 |

deck. The modified Q scores and the raw score for number of marker cards exceeded are very similar in information content, having a correlation of .98. Thus, either method could be used; the choice would depend on convenience.

### Conclusion and Discussion

We conclude (i) that the marker-card method is as reliable a measure as the ordinary Q-sort method; (ii) that Q and the marker-card score are measuring very nearly the same thing—in this case, presumably conflict; and (iii) that the use of marker cards permits comparison of judgments about a given quantity in the context of different situations and may even permit comparison of results obtained by different research teams if the different teams use the same marker decks.

A fundamental issue involved in scoring the entries on the cards in this study by either Q-sort or marker cards is the "judgability" of the items as indicators of conflict. We noted above that items with the highest variances in scoring, from judge to judge, were concentrated at the low-conflict end of the scale. A possible explanation is that judges are able to make finer distinctions concerning an attribute when it is present than when it is absent. Differences in scoring could also arise from a lack of unidimensionality in the attribute being scaled. If judges find that conflict has several distinct aspects, the task of placing items in a single order becomes more difficult.

Difficulties in making judgments also arose from the fact that the actions were being judged out of context, or in contexts that varied from judge to judge, since no standard context was supplied. But these difficulties lie outside our problem, which was to find an alternative to Q-sorting which would permit intersituational comparisons. Our success in finding an alternative is apparent in our results, but problems of judgability remain with both techniques.

Beyond providing a standard for intersituational comparisons, the marker-card technique has other advantages. It is possible that the use of marker decks will be of help in training judges to score such a variable as conflict. In our study it permitted identification of a judge inadequately trained to do so. Possibly the marker cards will be useful in assigning scale values to batches of data too small for Q-sort, or even to individual items. It further appears that the marker cards should be useful in discriminating among high-conflict items which heretofore would all have tended to appear in the top Q-sort category.

The marker-card technique has given us a reliable alternative to Q-sort for scaling conflict. The method should be capable of extension to dimensions other than conflict.


#### References and Notes

1. The specimens are to be distributed into the nine "levels of intensity" of the quality being judged in these proportions:

| *Level* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *Proportion* | | | | | | | | |
| 5 | 8 | 12 | 16 | 18 | 16 | 12 | 8 | 5 |

Note that the mean and variance are established by the nature of the distribution and would be the same for any such distributions irrespective of the data being scaled.

2. R. North, O. R. Holsti, M. G. Zaninovich, D. A. Zinnes, *Content Analysis* (Northwestern Univ. Press, Evanston, Ill., 1963). For an example of the application of this technique to material from the crisis preceding World War I, see R. North, R. Brody, O. Holsti, *Peace Research Society (International) Papers No. 1* (1964), p. 1.
3. Variance was measured in terms of the arc sine transforms of the ranks, to reduce end effects. In particular, the item ranked $r$ in the set of $N$ marker cards was transformed to:

$$\sin^{-1}\left(\frac{r}{N+1}\right)^{\frac{1}{2}} + \sin^{-1}\left(\frac{r+1}{N+1}\right)^{\frac{1}{2}}$$

4. L. E. Moses and R. V. Oakford, *Tables of Random Permutations* (Stanford Univ. Press, Stanford, Calif., 1963).
5. We take $r_{xx}$ as the average of the inter-judge marker-score correlations. Because Q is the average of two judges' Q values, we must correct for this fact; the Spearman-Brown "prophecy formula" can be shown to be applicable. It yields:

$$r_{QQ} = \frac{2(.733)}{1.733} = .846$$

6. Partial support for the research described was provided by contract NONR 225 (82), Project NR 177 254, Group Psychology Branch, Office of Naval Research. Reproduction of this article in whole or in part is permitted for any purpose of the United States Government. We acknowledge with thanks helpful conversations between one of us (L.E.M.) and Frederick Mosteller of Harvard University. Data analysis for this study was made possible by a grant from the Stanford University Computation Center. We thank Mr. Kuan Lee for his invaluable assistance with the computer programming for this study. Copies of the four marker decks are available upon request from the Stanford Studies in International Conflict and Integration, 550 Salvatierra Street, Stanford, California.


# Water Balance in Desert Arthropods

Despite their small size, arthropods may be highly adapted for life in xeric conditions.


E. B. Edney


Deserts are not homogeneous environments. The surface is indeed often very hot and dry and sometimes very cold, but there are plenty of protected niches in which the climate is much less extreme. Adaptations, therefore, are not always concerned with tolerance of, or regulation against, drought and heat, but often take the form of structural and behavioral characters associated with particular modes of life (for example, the flat shape and scoop-like legs of many dune insects), or of phenological mechanisms that permit the animals to take maximum advantage of short climatically favorable seasons. Such adaptations permit the avoidance of desert conditions and are associated only indirectly, if at all, with water stress. Nevertheless, water shortage and high temperatures are encountered by many desert animals, and this article attempts to consider the structural and functional mechanisms of arthropods which are important in relation to these aspects of desert life.

Nothing like a complete picture of such mechanisms is available at present since the physiology of desert insects, let alone other arthropods, has


The author is professor of zoology at the University of California, Riverside 92502.