Quality & Quantity **29**: 241–250, 1995.

Estimating qualifications in a self-evaluating group

I. BOMZE and W. GUTJAHR
Department of Statistics, OR, and Computer Science, University of Vienna, Austria

Abstract. We shall propose a method for assigning qualification values to individuals in a group, based on a cross-ratings matrix of its members. The method requires specification of a so-called 'calibration function' describing the dependence of judging competence on qualification.

1. Introduction and theory

In the evaluation of scientific or other institutions, two approaches may be followed: The first one is *evaluation from outside*, i.e., by persons not belonging to the institution under consideration. The other one is *self-evaluation*: the performance of individual members is judged by members of the institution. The term 'institution' may be understood in a very broad sense; for example, it may refer to the entire scientific community. In the latter case, some form of self-evaluation becomes indispensible: presumably, support for a system where scientific activities and results are judged only by non scientists will be lacking.

One important aspect of evaluation is attributing degrees of *qualification* to individuals. (This seems indeed to be a major component of every educational system.) In connection with self-evaluation, the aspect of measuring qualifications raises special problems, some of a definitory kind.

Consider, for example, a group of physicists who want to acquire information on the qualification of each member of the group. For this purpose, each physicist rates all of his colleagues. What definition could be used to characterize 'good' physicists, based on the ratings? The most straightforward characterization seems to be the following:

> A good physicist is one who is rated to be good by a majority of physicists. (1)

However, this definition does not adjust for the varying competence of physicists to judge (other) physicists. It might be conjectured that good physicists are more competent in judging their colleagues than the less good ones. So the following characterization appears to be more appropriate:

A good physicist is one who is rated to be good by the majority of *good* physicists. (2)

Because of its evident circularity, this definition might seem hopelessly impractical. The same problem occurs if we want to evaluate *artists*, sharing Elster's intuitively appealing opinion that the best judge on art is a *good* artist (Elster, 1987, p. 179).

Nevertheless, it is shown in Gutjahr (1994) that such definitions can make sense, provided they have a *fixed point*. Fixed-point approaches to circularity problems have been used in different areas of research: In philosophical logic, they are essential ingredients of Kripke's theory of truth (Kripke, 1975); in the foundations of artificial intelligence, they form a central tool for autoepistemic logic systems (Moore, 1988); in biology, they are closely connected to Maturana's and Varela's (1972) conception of *autopoiesis*; in the political sciences, Brams (1976) and Elster (1987) applied them to understand specific paradoxes of social behaviour; and in sociology, it was Luhmann (1984) who introduced, as a basic concept, v. Förster's fixed-point analysis of *observation* (cf. v. Förster 1985).

From a fixed-point interpretation of circular definitions, a definition of the form $x := \phi(x)$ is meaningful (but possibly not unique) if there exists an expression x which, inserted on both sides of $x = \phi(x)$, satisfies the latter equation. Now, for a given group $P_1, \ldots, P_n$ of physicists, let the variable x_j assume the value 1 if P_j is a good physicists, the value 0 otherwise ($j = 1, \ldots, n$). Then (2) may be reformulated as

$$x_j := \begin{cases} 1, & \text{if } P_j \text{ is rated to be good by the majority of } P_i \text{ with } x_i = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $j = 1, \ldots, n$. For given ratings and with $x = (x_1, \ldots, x_n)$, this yields an equation of the form $x = \phi(x)$ for the vector x. It can be demonstrated that there are cases for which a fixed-point solution x of (3) does not exist: interpreting 0 and 1 as truth values, one finds ratings where (3) has the same logical structure as the well-known Epimenides paradox (see Gutjahr, 1994).

In order to improve the chances for a fixed-point solution, it was suggested in Gutjahr (1994) to consider quantitative ratings defined on some interval $[0, M]$ of the real line, whereby one posits that the ratings are unique to a multiplicative scale constant c, $c > 0$, i.e., 'ratio-scale' properties for the ratings are presumed. For technical reasons, it is convenient to normalize the qualification values of the n persons considered, dividing them by the sum of all qualification values. Then the normalized or relative qualifications are numbers $x_1, \ldots, x_n$, with $0 \leq x_i \leq 1$ and $\Sigma_{i=1}^{n} x_i = 1$.

Now let a_{ij} denote person P_i's rating of person P_j's qualification on a nonnegative rating scale so defined. These values can be found empirically by means of a questionnaire, for an example, see Weiss (1968). Again we may normalize the values a_{ij} such that $\Sigma_{j=1}^{n} a_{ij} = 1$, i.e., each judge P_i distributes a total amount of 1 on the n persons $P_1, \ldots, P_n$, proportional to their respective degrees of qualification (in his or her opinion).

How should the values x_j be determined fiom the values a_{ij}? The simplest way is the following:

$$x_j := \frac{1}{n} \sum_{i=1}^{n} a_{ij} \qquad (j = 1, \ldots, n), \tag{4}$$

which quantifies the (self-evaluative) qualification of person P_i as the *average* score that P_i has obtained by the judges.

Clearly, also other weighted averages instead of the arithmetic mean could be used. In general,

$$x_j := \frac{\Sigma_{i=1}^{n} w_i a_{ij}}{\Sigma_{i=1}^{n} w_i}, \tag{5}$$

with nonnegative weights w_i.

Now let us define a class of qualification indices. Each index is determined by a *calibration function* $f(u)$, $0 \leqslant u \leqslant 1$. The function $f(u)$ specifies the weight assigned to the judgment of a person with qualification u. Thus we set

$$w_i := f(x_i) \qquad (i = 1, \ldots, n). \tag{6}$$

The basic idea is that the ability of judging (other) people's qualification is mainly dependent on one's own qualification.

By taking the special case $f(u) = 1$, one gets (4), which is the quantitative analogue of (1), since no influence of qualification on the judge's competence is assumed. For the special case $f(u) = u$, one obtains (note that always $\Sigma x_i = 1$):

$$x_j := \sum_{i=1}^{n} x_i a_{ij} \qquad (j = 1, \ldots, n). \tag{7}$$

This system of equations can be viewed as a quantitative analogue of (2), since it is based on the assumption that higher qualification results in higher competence to judge. The particular approach described by Equation (7) is

not new; it has been proposed by Lehrer and Wagner (1981) in their theory of rational consensus.

In the general case, inserting (6) in (5) yields:

$$x_j := \frac{\sum_{i=1}^{n} f(x_i) a_{ij}}{\sum_{i=1}^{n} f(x_i)} \qquad (j = 1, \ldots, n). \tag{8}$$

The values x_i occur on both sides of (7) and (8), so these definitions are circular or implicit. The system of equations may or may not have solutions $x = (x_1, \ldots, x_n)$, i.e., fixed points of the vector function defined by the right hand sides. If there exists a solution x, it will be referred to as the vector of *self-evaluative qualifications* of person $P_1, \ldots, P_n$ with respect to calibration function f.

By a formal analogy to the theory of Markov chains (see, e.g., Kannan, 1979, or Kemeny & Snell, 1962, pp. 128–131), it is immediately seen that (7) has a unique solution, except in 'degenerate' cases: Interprete the numbers a_{ij} as the transition probabilities from states i to states j in a Markov chain with state space $\{1, \ldots, n\}$. The necessary conditions $0 \leq a_{ij} \leq 1$ and $\Sigma_j a_{ij} = 1$ are satisfied. Let $x = (x_1, \ldots, x_n)$ be a probability vector. Then (7) is the condition for x to be a *stationary distribution* of the Markov chain. It is well known that if the matrix $A = (a_{ij})$ is ergodic, i.e., irreducible, aperiodic, and positively recurrent, then (7) has a unique solution. For the ergodicity of A, relatively weak conditions are sufficient, for example the condition that the ratings a_{ij} are strictly positive, i.e., $a_{ij} > 0$ $(i = 1, \ldots, n;$ $j = 1, \ldots, n)$. Thus we have (cf. also Lehrer and Wagner, 1981):

If the matrix $A = (a_{ij})$ satisfies $a_{ij} > 0$ then for the identity calibration function $f(u)$, there exists a unique vector of self-evaluative qualifications.

Practically, the solution x of (7) can be computed by iteration:

$$x_j^{(n+1)} = \sum_{i=1}^{n} x_i^{(n)} a_{ij} \qquad (j = 1, \ldots, n; n = 0, 1, \ldots) \tag{9}$$

with an arbitrary initial vector of qualification estimates $x^{(0)}$, say the arithmetic mean estimate (4). Because of the ergodic theorem (see, e.g., Kemeny & Snell, 1962, pp. 128–131), the following holds:

Under the assumption $a_{ij} > 0$, the sequence of vectors $x^{(n)}$ defined by (9) always converges to the unique fixed point x of (7).

In the context of (self-)evaluation of scientific journals, the iterative procedure (9) was already used by Liebowitz & Palmer (1984). However, they were apparently unaware of the formal analogies to Markov chain theory, because they took great care to choose an appropriate initial qualification estimate $x^{(0)}$, without recognizing that the limiting fixed point solution of (7) does not depend on $x^{(0)}$.

The case of nonlinear calibration functions can be treated similarly, albeit with considerably higher technical effort, see (Bomze & Gutjahr, 1995). While the existence of a fixed point solution of (8) is still guaranteed for all continuous functions f, the solution is not necessarily unique in this situation. Indeed, already in the simple case $f(u) = u^2$ and $n = 2$, there are examples where more than one fixed point of (8) exists. In such cases, the limit of an iteration analogous to (9) *does* depend on the chosen initial value. It seems reasonable to always choose the arithmetic-mean estimate (4) as the initial value of the iteration. This leads to the following convention: The vector of self-evaluative qualifications is defined as that fixed point of (8) that has a domain of attraction containing the arithmetic-mean estimate (4).

In this paper, we restrict ourselves to the class of power functions $f(u) = u^k (k = 0,1, \ldots)$ as calibration functions. The case $k = 0$ yields the arithmetic mean estimates (4), the case $k = 1$ the "Markov chain fixed points" (7). Setting $f(u) = u^k$ with a value $k > 1$ represents the assumption that the competence of judging qualifications, as a function of qualification, grows with a positive second derivative: the influence of qualification on judgment competence is stronger in the high qualified than in the low qualified judges.

2. Empirical results

We emphasize the fact that the self-evaluative qualifications x_i, as defined in Section 1, only reflect subjective estimates. From this point of view, choosing the calibration function f (linking judgment competence to qualification) is *a priori*: If qualification cannot be measured independently of the individual ratings, then it is also impossible to determine the 'true' judgment competence, i.e., the correlation between estimated and real qualifications.

In some cases, however, qualification can be measured by an external criterion, for example an ability test or the estimate of an *external* person whose competence of judgment is commonly accepted. In such a situation, it would be interesting to compare the self-evaluative qualification values, computed via a certain calibration function from the results of a self-evaluation experiment, with the externally determined qualification values. The latter will be called *objective qualifications*.

Table 1. Group sizes

Group	I	II	III
Total no. of students	39	26	40
No. of students in self-evaluating kernel	9	11	7

Three groups of students participated in our experiment. The groups were classes of undergraduate training courses in mathematics (linear algebra and calculus) for students of Computer Science at the University of Vienna (Department of Statistics, OR and Computer Science). The course for Group I took place in winter 1992/93, the courses for Group II and III in summer 1993. Table 1 shows the sizes of the three groups.

During the training courses, the students had to present their solutions to problems from a textbook; they were guided and corrected by an instructor. At the end of the course, a written examination took place, where a score of 0 (worst result) to 40 (best result) could be obtained.

At the beginning of each course, the students were asked to participate in an evaluation experiment. Each volunteer received a questionnaire with the names of the other students of the course. She or he was told to predict for herself (himself) and for each of her (his) colleagues their result at the final examination (a number between 0 and 40). At the end of the course, prior to the examination, the questionnaires were collected.

Only about half the students of each course participated in the experiment. Several of them had missing values in their answers: They did not rate *all* of their participating colleagues. Questionnaires with missing values were eliminated from the evaluation. Thus we finally obtained, for each group, complete (but relatively small) matrices of cross-ratings. The numbers of students in these self-evaluating kernels are also registered in Table 1.

Let us start the analysis of the results with Group I. The complete matrix of the ratings is shown in Table 2. The ith row contains the judgments passed by the ith student.

The true results z_j of the written examination are shown in Table 3.

From the values b_{ij} in Table 2, the self-evaluation matrix $A = (a_{ij})$, as described in Section 1, is computed by normalization of the rows:

$$a_{ij} = \frac{b_{ij}}{\Sigma_k b_{ik}}.$$

The self-evaluative qualifications were computed according to (8) for the class of power functions $f(u) = u^k$ $(k = 0, 1, \ldots)$ as calibration functions.

Table 2. Matrix of ratings b_{ij} (student i's prediction of student j's result) in Group I

	1	2	3	4	5	6	7	8	9
1	29	31	12	29	17	15	37	10	21
2	35	20	20	20	20	25	19	15	20
3	36	33	30	17	16	35	28	27	32
4	30	22	20	20	23	20	19	20	20
5	37	20	27	25	25	15	27	21	19
6	35	25	15	30	25	30	40	20	20
7	37	27	13	30	17	17	30	10	25
8	33	29	32	29	26	31	37	21	35
9	16	29	21	29	17	34	30	36	15

Table 3. Examination results of Group I

Student	1	2	3	4	5	6	7	8	9
Score	26	33	30	28	37	27	39	04	14

Now let

$$s^{(k)} = (s_1^{(k)}, \ldots, s_n^{(k)})$$

denote the fixed point of (8) with $f(u) = u^k$, that has a domain of attraction containing the arithmetic mean estimate $x = s^{(0)}$. Our main question was: Do the fixed point estimates $s^{(k)}$ for some k correlate higher with the true result z than the 'naive' arithmetic mean estimate $s^{(0)}$?

As a measure of correlation, we used the Pearson product-moment correlation coefficient $r(s, z)$.

The Group I column of Table 4 contains the correlation coefficients $r(s^{(k)}, z)$ for $k = 0, \ldots, 10$. Therefore, for the self evaluation matrix of Group 1, the above question may be answered with an affirmative. The best correlation was obtained for $k = 8$.

Nevertheless, the improvement from $r(s^{(0)}, z)$ to $r(s^{(1)}, z)$ is not impressive. A deeper analysis of the data reveals why: Let $a_{i\cdot} = (a_{i1}, \ldots, a_{in})$ and let $\rho_i = r(a_{i\cdot}, z)$ denote the correlation coefficient between the ratings and the examination results. Obviously, ρ_i is a measure for the judgment competence of judge i. Then, contrary to our expectation, the values ρ_i did *not* correlate positively with the qualifications z_i. The correlation was $r(\rho, z) = -0.019$ (cf. Table 4). Roughly speaking, in Group I, good students were on the average not superior to weak students in judging qualifications. However, the best students had an above average competence for judgment.

In Group II, the improvement achieved by computing $s^{(k)}$ for higher values

Table 4. Values $r(\rho, z)$ and $r(s^{(k)}, z)$ for the three groups

Group	I	II	III
$r(\rho, z)$	−0.019	0.090	−0.226
k	$r(s^{(k)}, z)$	$r(s^{(k)}, z)$	$r(s^{(k)}, z)$
0	0.400	0.617	−0.066
1	0.416	0.634	−0.155
2	0.431	0.652	−0.256
3	0.443	0.671	−0.351
4	0.451	0.691	−0.428
5	0.456	0.712	−0.484
6	0.460	0.733	−0.525
7	0.426	0.752	−0.555
8	0.463	0.770	−0.577
9	0.433	0.787	−0.594
10	0.399	0.802	−0.607

of k was more distinct (see Table 4), caused by the slightly positive correlation coefficient $r(\rho, z) = 0.090$ between judgment competence and qualification. A maximum was obtained for $k = 16$, namely $r(s^{(16)}, z) = 0.847$. Compared with the accuracy $r(s^{(0)}, z) = 0.617$ of the arithmetic mean estimate, this is a remarkable increase.

The results of Group III are also interesting. The arithmetic mean estimates $s^{(0)}$ happened to correlate *negatively* ($r = -0.066$) with the examination results. (This seems to be caused by the very small size $n = 7$ of this group, which obviously favored the influence of randomness.) The effect of computing $s^{(k)}$ for higher values of k is not an improvement, but a further deterioration of the estimates. Naturally, an indispensible condition for correctly estimating an objective qualification by any of the proposed self evaluative qualification indices, is a positive correlation between ratings and objective qualification. For Group III, however, we obtained $r(\rho, z) = -0.226$.

An explanation for the surprising fact that, in our experiment, good students did not judge better than weak students, might be the public feedback the instructors of the groups gave during the training courses: Even less qualified students were able, by means of this feedback, to realize whether a presentation of another student was adequate or not. The difference in accuracy between the arithmetic mean estimate $s^{(0)}$ and the 'Markov chain fixed point' estimate $s^{(1)}$, or the estimate based on a calibration function f differing from the identity, would possibly prove much more drastic in an experiment where such feedback was absent.

In general, we expect a (small) positive correlation $r(\rho, z)$ between the competence of judging others and the externally determined objective qualification, but that would be a subject for a large empirical investigation.

Our own experimental data are only intended to illustrate the proposed methodology.

3. Conclusion

A method has been presented for assigning qualification values (with respect to a specified attribute or ability) to individuals from a group, based on a matrix of cross ratings of the members of the group. The method requires the specification of a 'calibration function' f, describing the dependence of competence for judging on qualification. The function f may be obtained in two different ways:

(a) *Empirically*, given that 'objective' qualifications can be measured. In this case, a primary investigation should be carried out in order to determine an optimal shape of f for the qualification attribute under consideration. Once this has been done, the presented method may be used for estimating objective qualification values in similar situations where only cross-ratings are available. On the condition that the calibration function has been chosen appropriately and that there is a positive correlation between average ratings and objective qualification values, the presented method is likely to yield better estimates than the arithmetic means of the ratings.

(b) *Theoretically*, whenever 'objective' qualification values cannot be measured because of the intrinsic 'subjectivity' of the qualification attribute under consideration. In this case, a dependence between competence for judgment and qualification, as specified by some chosen calibration function f, has to be justified by theoretical arguments. For a plausible calibration function, the presented method yields a plausible operational definition of qualification with respect to the subjective qualification attribute inquired by the ratings.

Acknowledgments

The authors are deeply indebted to G. H. Fischer for valuable suggestions and stimulating comments. We also want to thank E. Reschenhofer for drawing our attention to reference (Liebowitz & Palmer, 1984).

References

Bomze, I. M. and Gutjahr, W. (1995). The dynamics of self-evaluation, to appear in: Applied Mathematics and Computation (1995).

Brams, S. (1976). *Paradoxes in Politics*, New York: Free Press.

Gutjahr, W. (1994). Paradoxien der Prognose und der Evaluation: Eine fixpunkttheoretische Analyse, to appear in: *Annals of the Kurt Gödel Society* (1994).

Elster, J. (1987). *Subversion der Rationalität*, Frankfurt/New York: Campus.

Förster, H. v. (1985). *Sicht und Einsicht: Versuche einer operativen Erkenntnistheori*, Braunschweig: Vieweg.

Kannan, D. (1979). *An Introduction to Stochastic Processes*. Amsterdam: North-Holland.

Kemeny, J. G. and Snell, J. L. (1962). *Mathematical Models in the Social Sciences*, Waltham, Mass.: Blaisdell.

Kripke, S. (1975). Outline of a theory of truth, *The Journal of Philosophy* 72: 690–716.

Lehrer, K. and Wagner, W. (1981). *Rational Consensus in Science and Society*, Dordrecht: Reidel.

Liebowitz, S. J. and Palmer, J. P. (1984). Assessing the relative impacts of economics journals, *Journal of Economic Literature* 22: 77–88.

Luhmann, N. (1984). *Soziale Systeme: Grundriß einer allgemeinen Theorie*, Frankfurt: Suhrkamp.

Maturana, H. R. and Varela, F. (1972). *Autopoiesis*, Santiago.

Moore, R. C. (1988). Autoepistemic Logic, in: Smets, Mandani, Dubois and Prade (eds.), *Non-Standard Logic for Automated Reasoning*, London: Academic Press.

Weiss, R. S. (1968). *Statistics in Social Research*, New York: Wiley.