

Draft version, Januari, 2003

Modeling Assessment Data

Willem K.B. Hofstee

1. DEFINING QUALITIES

By assessment I mean the attribution of a quality or qualification Q to a person or object P by judges or assessors B . Assessments constitute an important and consequential part of human discourse. Moreover, much of the data in the social and behavioral sciences consists of assessments, rather than measures in any strict sense. The question here is how to conceive of such qualifications, and how to handle them. Traditionally, qualifications are conceived as measurements on a relative metric scale, and handled through classical psychometrics and statistics. I question the appropriateness of that approach and the assumptions that it implies. I develop a more elementary or primitive conception that I argue to be more adequate in the typical case.

1.1. Assessment Versus Measurement

Treating assessments as measurements is to make a shortcut whereby assessors are seen as instruments among other instruments. That conception is quite defensible if the data are factual. For example, one could think of measuring an individual's age or sex in some objective manner, but in ordinary circumstances it is more efficient to ask someone, preferably even the person himself or herself. With qualities, however, things are quite different. If we want to establish a person's integrity, for example, we would realize that any assessment of it is subjective; we would not be inclined to believe the assessor at his or her word, least of all if that assessor were the person himself or herself. Believing John's assessment of his own integrity would be begging the question; asking Mary about her modesty would be paradoxical. Assessments are as much a function of B as of P , and that problem is only aggravated if B and P happen to be one and the same person. To ignore B 's role – as in calling questionnaires “tests” and their scores “measurements” – is to neglect a basic problem in the behavioral and social sciences.

Qualities can be defined in contrast with factual attributes: Qualities do not refer to an objective state of affairs. They are judgmental in principle. In practice, assessments of qualities can sometimes be successfully predicted from objective measures: intelligence from a psychological test (although the example is debatable), and perhaps in the future, personality characteristics from genetic polymorphisms. But even in such cases, the reference is the other way around: Measures derive their validity from their

ability to simulate the assessment; the primary definition of the quality is a matter of human judgment. Only in the long run, successful indicators may substitute judgments of a quality. Subjective assessments of facts are unproblematic because they refer to an objective criterion; they can safely be used precisely because there is no real need to use them. With qualities, there is no such safeguard. To overlook their judgmental nature is like walking on quicksand without taking proper precautions.

1.2. Quality Versus Taste

At the other end of the subjective-objective spectrum, qualities can be distinguished from matters of personal taste. Taste, in the sense that “tastes differ”, is irrevocably subjective. It is indisputable: If you like Wagner’s music and I don’t, we do not have a disagreement that we could solve by trying to convince each other. It is as if we float in different worlds; we do not touch base with a common ground. Likes and dislikes are a subject matter for solipsistic philosophers. A tenet in classical economics concerns the impossibility of interpersonal comparison of utility. All this is not to say that personal taste makes no sense, or that tastes have no place in everyday discourse. However, in accordance with common understanding, I conceive of qualifications in a more official sense. In the typical case where qualities and qualifications are at issue, *B*’s are commissioned by some authority *A* to assess them. Such assessments are supposed to be representative. Qualities are thus understood as *intersubjective*. It does make sense for you to try to convince me that Wagner is a great composer, and I might even agree in the end, disgusting though I keep finding his music.

A complication is that sometimes people are said to have *good* taste, meaning “the ability to notice, appreciate, and judge what is beautiful, appropriate, or harmonious, or what is excellent in art, music, decoration, clothing, etc.” (Webster), or the “faculty of discerning and enjoying beauty or other excellence especially in art and literature, or appropriateness of conduct” (Oxford). Such notions can easily lead to an elitist conception of qualities: It is as if quality is defined by the taste of a happy few. Here, however, the point of reference shall be the judgment – not the taste – of people in general. That does not exclude the possibility that some assessors perform consistently better than others, and in that sense appear to have better “taste”. But it would mean that they represent the common point of view better than others, which is different from having a privileged position. So what the complication amounts to, is that assessors may receive different weights in the averaging of their judgments.

In conclusion, assessors are interchangeable in principle, though more or less representative in practice. Qualities are primarily defined by some common component in

their judgments: The reference point is neither in an objective measure nor in the supreme autonomy of some individual or class. Even a common component of actual evaluators can only provide an approximation of the true quality of a *P*. In principle, one could form an idea of what this true quality is. One would take a common component over all possible judges, of all times including the future if the quality is conceived in a timeless manner. This idea, impractical though it may be, can nonetheless serve as a conceptual reference point for defining quality and evaluating assessments.

2. SCALING ASSESSMENTS

The intersubjective definition of qualities implies the possibility of aggregating assessments over assessors. The question then is what kind of scale one should use, particularly, whether it makes sense to quantify qualities, which may well sound like a contradiction in terms. The issue will be approached here in a circuitous manner, calling into question the unbounded interval scales that are postulated in classical applied statistics, but not rejecting a priori a quantitative conception of assessments. So at first, I discuss whether the scale should have a natural or an arbitrary zero point, and whether its intervals are arbitrary or fixed (Suppes & Zinnes, 1963; Zegers & Ten Berge, 1985):

	<u>Arbitrary zero point</u>	<u>natural zero point</u>
Arbitrary intervals	interval scale	ratio scale
Natural intervals	<u>difference scale</u>	<u>absolute scale</u>

The question about intervals appears to be associated with the question whether the scale is bounded or unbounded. Finally, since assessors are human, I investigate the feed-forward effects of a particular scale on their rating behavior; that inquiry appears to lead to an extremely simple final solution.

2.1. Bipolarity

By and large, qualifications come in pairs of opposites. The clearest examples are denials of trait adjectives, such as disagreeable, inattentive, and unreliable. In principle, such denials could be interpreted in a unipolar or a bipolar manner: Unfriendliness could point to the absence of friendly behavior, or to its opposite. There can be little doubt that the latter is the case, and that denials are understood in the manner of a *litotes* (Oxford: “expressing of an affirmative by the negative of its contrary”). If a person is judged to be independent, it may mean that he or she is confident; if another person is judged not to be confident, that may mean that she or he is dependent on others. Such exercises would

make no sense if independence were merely null dependence, and lack of confidence merely null confidence. If B_1 judges P to be friendly, whereas B_2 judges that person to be unfriendly, the compromise that P is somewhat friendly rather than somewhat unfriendly would typically not be acceptable to B_2 . Comparably, if P behaved in a friendly manner half of the time, and in an unfriendly way the other half, few assessors would accept the conclusion that this P is somewhat friendly rather than somewhat unfriendly. Thus the most obvious interpretation of assessments is on a bipolar scale. This is not to say that qualities *are* bipolar in some abstract sense; all it says is that assessments are better understood and represented in that manner.

A seeming exception arises in the appraisal of performances. At first sight, it is difficult to think of negative intelligence, achievement, and the like. Surely, absolute negative scores may occur. On a multiple choice test, P may score below chance level, that is, his or her score is negative upon correction for guessing (for example, P scores below 50 on a test of 100 yes-no items). More generally, P may come up with solutions to problems that are counterproductive rather than just unproductive, like mere guessing is. But in these examples, negative quality is an extreme or pathological case. In the general case, the standard or threshold is quite a bit higher, as in a pass-fail threshold on an exam, or in predicting whether an applicant has a positive expected net value for an organization. Still, the scale may be viewed as bipolar, as P 's performance is pitted against a standard, arbitrary though it may be.

Another, more exotic, seeming exception to bipolarity is formed by configurations of four traits like the following:

extravagant – generous – thrifty – stingy

Here, it is as if there are three bipolarities since extravagant and generous are of opposite social desirability, as are thrifty and stingy, and since generous and thrifty are of opposite content. The outer terms are popularly interpreted as “too much of a good thing” or “extremes that meet”; “the virtues are in the middle”. However, Peabody (1967) proposed that such configurations are more adequately represented in two dimensions:

generous	thrifty
extravagant	stingy

In Peabody's conception, we have a content contrast – spending *versus* saving – horizontally, and a social desirability contrast vertically, so the opposition is diagonal,

that is, between generous and stingy on the one hand, and thrifty and extravagant on the other: pairs of traits that differ on both content and social desirability. Hofstee and Arends (1994) introduced a further refinement to this “chiasmic” representation, on the basis of a generalized circumplex model (Hofstee, De Raad, & Goldberg, 1992). They argued that sheer social desirability, independent of content, cannot and need not be postulated, and proposed two content contrasts, as follows:

	<u>Spending:</u>	<u>Saving:</u>
Prosocial:	generous	
Neutral:	extravagant	thrifty
Antisocial:		<u>stingy</u>

Hofstee and Arends were successful in constructing new chiasms on the basis of the circumplex model. They concluded that the one-dimensional representation, and the maxims that derive from it, are misleading. So there is no need to worry about nonlinear scales with more than one bipolarity.

2.2. Zero point

Bipolar scales have a zero point at which the quality is reversed. With Likert scales running from 1 to c (for example, 1 to 5), the midpoint $(1 + c)/2$ is the logical candidate (3, in the example). For an appropriate representation of assessments, the value of the midpoint should be subtracted from all scale points, so that a 5-point scale is translated into $[-2, -1, 0, +1, +2]$. An implication of such a natural zero point is that additive translations, as in standard statistics, are inappropriate.

The argument is best illustrated with regard to personal qualities, for example, conscientiousness. In any representative sample, most P will be assessed at the positive side of the scale midpoint. In standard statistics, based on a relative metric (interval scale), that outcome cannot be taken seriously: The only admissible point of reference is the population mean, as estimated by the mean of a representative sample. Here, that mean is above the midpoint, say, at $+0.8$ on the bipolar 5-point scale. P 's relative conscientiousness is thus expressed as a deviation from the mean, typically, in terms of standard deviations or z -scores. Consequently, all scores between the midpoint (0) and the mean ($+0.8$) change sign: The corresponding individuals, who were judged to be on the conscientious side, are now judged to be on the unconscientious side. That is an inadequate translation of these judgments.

A counterargument, familiar in social psychology, is that the relative scaling corrects for a bias. If the individuals themselves make the assessments, they are said to respond socially desirably; if others make them, these assessors are said to be lenient. The fact that judgments by others tend to be positive indeed is explained by the fact that those B who know P well enough to be able to assess him or her, will be friendly to P , and thus constitute a biased sample. What is generally overlooked is that there is no independent proof that assessors are biased. So we are faced with a choice between two explanations. The social-psychological one is relatively complicated as it supposes separate mechanisms for self and others; its main merit seems to be that it justifies the deeply ingrained habit of relative scaling. The other explanation is more parsimonious: It proposes that most people *are* conscientious, or socially desirable in general, and are therefore assessed to be so by both self and others. It is also more plausible, as the human kind would probably not have survived if about half of it would consist of socially undesirable individuals, if about half of its products would have negative quality, and the like.

A further argument in favor of an absolute representation of assessments derives from bivariate distributions like the following, in which, for example, two assessors would have judged P 's conscientiousness on 100 occasions:

		Assessor B_1		Assessor B_2	
		Positive	Negative	Positive	Negative
Assessor	Positive	5	90	90	5
B_2	Negative	0	5	5	0

The question for each of the two fourfold tables is to what extent the two assessors agree. According to coefficients like Pearson's \mathbf{j} and Cohen's \mathbf{k} that are based on relative scales, the agreement in the left table is slightly positive (both \mathbf{j} and \mathbf{k} are about +.05) and in the right table, slightly negative (both \mathbf{j} and \mathbf{k} being about -.05). The tetrachoric correlation coefficient, which postulates an underlying binormal distribution in addition to relative scaling, is even perfectly positive (+1) in the first table, and perfectly negative in the second. All this is not what one would expect in view of the elementary fact that the assessors have 90% disagreements in the first table, and only 10% in the second. The result comes about because these coefficients calculate relative agreement *given* the assessment rates (the proportions of positive assessments) of the assessors, which are vastly different in the first table and equal in the second. On an absolute scale, the

difference in assessment rates enters into the definition of agreement. Coefficients of absolute agreement are discussed below.

In conclusion, evaluative assessments are best conceived and represented as deviations from an absolute zero point. An important distinction here is between assessment and evaluation on the one hand, and selection or choice on the other. Selection, for example personnel selection, is typically comparative; it is a home base for classical applied statistics. Assessment finds its place in settings like admission and job counseling. An illustrative contamination arises when assessments are made in a selective context, such as personnel selection but also applications for subsidies, grants, or vacancies, with limited budgets or numbers. The primary question to the assessor is about a quality, for example, whether the applicant will be fit for the job. According to reasonable standards, the base rate (the proportion of suitable candidates) in personnel selection is typically in the order of 95% (see, e.g., Van der Maesen de Sombreff, 1992). The assessment rates of selection consultants, however, tend to be much lower (see, e.g., Van Dam, 1996). It is as if these assessment rates are geared toward the selection rates, which are automatically lower than 95% in comparative selection. Thus when the consultant assesses two applicants for a single vacancy, apparently he or she tends to find only one fit for the job. That may be realistic, firstly because only one of the two will be hired anyway, and secondly because the employer may appreciate a clear advice. But the dark side of this confusing of evaluation and selection is that it is unfair to a large percentage of the applicants, namely, the percentage that follows from subtracting the assessment rate from the base rate. Indirectly and in the long run, it may also be detrimental to the public image of the consultancy profession. In the present context, the contamination only illustrates the importance of distinguishing between absolute assessments and relative selection.

2.3. Bounded Scales

So far, the argument was in favor of ratio and absolute scales over difference and interval scales. The choice between ratio and interval scales is next. First observe that in practice, all assessments are made on bounded scales. No B is given the opportunity to judge a P to be infinitely friendly, for the simple reason that this would make assessments by different B 's infinitely incomparable. So here again, there is a certain tension between assessment practice and classical statistics, in which the normal (and thus unbounded) distribution plays a major role. To ensure aggregation over assessors, their assessments should be made on a bounded scale; in representing assessments, it would seem inadequate to stretch that scale to plus and minus infinity.

Percentages and proportions are bounded. They provide a standard for translating Likert scales with different numbers of scale points. With scale points 1 to c , the translation is:

$$(2Q - c - 1)/(c - 1),$$

which does not look particularly transparent but simply means: Set the scale ends at -1 and $+1$, and interpolate linearly. An obvious quantitative interpretation is that a quality Q has been observed on a number of occasions, the assessment representing the balance of positive and negative instances, for example:

positive instances	100	75	50	25	0	
	yes!	yes..	??	no..	no!	
	0	25	50	75	100	negative instances

Note that the scale could accommodate an unlimited number of nuances if a line segment is used rather than a Likert scale.

In sum, the bipolar proportional (biproportional for short) scale has a natural zero point. Moreover, it is standardized at the scale ends. That means that neither additive nor multiplicative transformations are appropriate. It thus forms an absolute scale. For representing assessments, such a scale appears to be more adequate than the unbounded interval scale of classical statistics. A further trimming of the assessment scale will be put forward in the shape of a binary bipolar scale. However, that development relies on the next Section, particularly, on the choice of an index of association in the context of weighting assessors.

3. CORRESPONDENCE BETWEEN ASSESSORS

A spin-off of the biproportional scale is that it greatly simplifies bivariate and multivariate statistics. Coefficients of correlation, association, similarity, agreement, and the like, are conventionally standardized to a maximum of $+1$ and a minimum of -1 . With the proportional scale, which is already confined within those bounds, additional standardizing is no longer necessary.

3.1. The Likeness Coefficient

An obvious choice for a coefficient of association for biproportional scales is $\Sigma Q_1 Q_2 / N$, the cross-product average of two qualities. It is to biproportional scores what the correlation $\Sigma z_1 z_2 / N$ is to standard scores, and the covariance to deviation scores. The average cross product was named L -coefficient by Hofstee and Ten Berge (in press).

The most conspicuous property of the L -coefficient is that it is sensitive to the size $\Sigma Q^2/N$ or L_{QQ} of the scores. Take the following examples:

<u>Q_1</u>	<u>Q_2</u>		<u>Q_1</u>	<u>Q_2</u>
+1	+1		+.01	+.01
+1	+1		+.01	+.01
<u>-1</u>	<u>-1</u>		<u>-.01</u>	<u>-.01</u>

In the first data set, $L_{12} = 1$. In the second, $L_{12} = .0001$, even though the scores are also identical per pair. L is thus not an identity coefficient, like the e -coefficient proposed by Zegers and Ten Berge (1985), which may be viewed as a size-corrected L -coefficient:

$$e_{12} = 2\Sigma Q_1 Q_2 / (\Sigma Q_1^2 + \Sigma Q_2^2) = L_{12} / [1/2(\Sigma Q_1^2/N + \Sigma Q_2^2/N)],$$

the correction representing the mean of the sizes $\Sigma Q_1^2/N = L_{11}$ and $\Sigma Q_2^2/N = L_{22}$. In both examples, $e = 1$. Contrary to e_{12} , the L -coefficient is better viewed as a similarity or likeness coefficient. It embodies a notion of agreement in which size or saliency is a precondition for similarity: It implies, for example, that two P 's can only be similar on extraversion if they are at all extraverted, or agree on an issue only to the extent that they have an opinion. A related specific property of L is that B can maximize it by giving an extreme rating, thus ± 1 . In the context of weighting assessors, it will be shown that this property leads to further consequences with respect to the scaling of assessments.

3.2. Absolute Agreement

To justify coefficients of absolute agreement in general, take again the case in which an individual would be assessed by B_1 and B_2 to behave in a friendly (+1) or unfriendly (-1) manner in a number of situations. This time, their assessments are uncorrelated (statistically independent) over these situations. The thought experiment gives rise to fourfold tables such as:

		B_1		B_1	
		<u>Friendly</u>	<u>Unfriendly</u>	<u>Friendly</u>	<u>Unfriendly</u>
B_2	Friendly	50	50	81	9
	Unfriendly	<u>50</u>	<u>50</u>	<u>9</u>	<u>1</u>

The difference is only in the base rates, both of which are .5 in the left table and .9 in the right table. In the latter case, the conclusion that the assessments of B_1 and B_2 are similar is inescapable: They have 81% overlap; they can be predicted from one another with a

high degree (90%) of certainty. Since the assessments are as statistically independent as they are in the left table, the assessment rates appear to enter into the definition of similarity; coefficients of relative agreement like \mathbf{j} and \mathbf{k} are insensitive to that.

When assessments are made on a [+1, -1] scale, L (and also e) simplify into:

$$L_{bin} = p - q = 2p - 1,$$

in which p is the sum of the proportions in the diagonal cells, denoting agreements, and $q = 1 - p$ denotes disagreements. In the left table, $L_{bin} = 0$, in the right table, $(.81 + .01) - (.09 + .09) = 2(.81 + .01) - 1 = .64$. The binary L -coefficient is identical to a coefficient proposed by Holly and Guilford (1963), among others (see, Popping, 1983, p. 71 f.). It represents absolute agreement on the binary bipolar scale.

The so-called coefficient of agreement p is widely frowned upon precisely because it does not take the value of 0 when the variables are statistically independent (see, e.g., Popping, 1983, p. 12ff.); the same argument would apply to L_{bin} . However, the argument is mistaken. The marginal distributions are reflected in the value of these coefficients: If two assessors have different assessment rates, L_{bin} implies that they disagree upon relative statistical independence; if their assessment rates are equal, that in itself boosts the coefficient. But all that says is that marginal distributions matter in judging absolute agreement. In a situation of absolute statistical independence, assessors would draw their marginal distributions at random from some universe of such distributions. Under appropriate conditions, for example, a universe consisting of a symmetric distribution of assessment rates, the expected value of the absolute coefficients *would* be zero. In other words, the argument against coefficients of absolute agreement or likeness overlooks the association between the marginals, that is, the assessment rates.

3.3. The Single Case

What coefficients of absolute agreement like L and e have in common is a particular and intriguing property: Unlike coefficients of relative agreement, they are defined in the single case (see also Hofstee & Zegers, 1991). It would technically be meaningless to say, for example, that two assessors are “correlated” with respect to John’s friendliness (unless, of course, one would have a series of observations on that), since the correlation coefficient is undefined in a single pair of observations. Nonetheless, it makes perfect sense to say that two assessors agree with respect to John. The coefficients capture that intuition.

L , but not e , has the additional property that a sample coefficient is the mean of the individual coefficients, as its denominator is a constant. If a sample coefficient is split up over individuals, the individual values are equal to what they were before aggregation.

This property implies that association or likeness is expressed irrespective of other cases, and is, in that sense, absolute. *L* is thus specifically geared to expressing agreement in individual cases.

4. WEIGHTING ASSESSORS

The weighting of assessors is at first unavoidable in the sense that in practice, by far the most *B* must receive null weights: Their number is virtually unlimited, so an assessment authority *A* has to make selections in advance. The selection criterion is supposed representativeness, that is, *A* should try to select assessors whose ratings would maximally correspond with the true quality. This is a judgmental problem by itself: There is no objective criterion; subjective selections are bound to be unrepresentative and lead to a deficient implicit definition of the quality in question; the only alternative is an intersubjective approach. For composing a representative *A*, one would need a meta-authority, and so on ad infinitum. An alternative to selection based on representativeness would be random selection, as in choosing jury members. However, in matters of quality as distinct from guilt, authoritative selection appears to be found more acceptable, for reasons of efficiency.

A second weighting problem presents itself once the assessors are appointed. It would be inconsistent to assume that all of them are equally representative of the common point of view. One would thus think of weighting them in some optimal fashion, that is, according to the correspondence of their assessments with true quality. That criterion is not available as such; however, at this stage a defensible procedure is available that shortcuts an infinite regress. It consists of accepting the common component of the panel as an approximation to the true quality, therefore, as a criterion. As a first step, one would take as a criterion the unweighted mean over all assessors. To the extent that assessors correspond more with that criterion, they should be given higher weights. Next, the procedure should be reiterated taking the weighted mean as the criterion, and so on until convergence occurs. The procedure constitutes a recursive definition of qualities. It strikes a perfect middle between egalitarian (unit weights only) and elitist (zero or unit weights only) definitions of it. In retrospect, both these conceptions appear to special cases of the overarching recursive definition of qualities.

If the initial selection of assessors is unrepresentative, the weighting of panel members can only lead to further bias and derailment. If a panel of assessors for judging creative work in art or science consists mainly of creditable but dull experts, a creative panel member will receive many negative weights, and the inadequacy of the final judgments will only be aggravated. The primary responsibility for a representative

composition of a panel is with the assessment authority A . However, panel members have an important secondary responsibility. They should make sure that they trust each other, in a fairly precise sense: Assuming that they themselves make an effort to take a universal perspective, they should have the expectation that the average other panel member does the same, so that their weight is maximized. Trust also means that an assessor accepts having been wrong some of the time. The requirement of what may be called professionalism on the part of the assessors provides an important even if partial solution to the assessment authority's problem of securing representativeness. Mutual trust among the panel members, and thereby implicit trust in A , is a reasonable warrant for representativeness.

4.1. Empirical assessor weights

Given a selection of B 's, we thus have an operational criterion for assigning weights to them, namely, their correspondence to the observed mean of the assessments. For a first illustration, take the most elementary case in which two assessors rate a single P :

	Q
B_1	.5
B_2	1

Their unweighted mean is .75. The L -coefficient between B_1 and the mean is $.5 \times .75 = .375$; for B_2 , it is $1 \times .75 = .75$. In the next step, the B 's should be weighted proportionally to their L .

To keep the weighted sum on the same scale, the weights w should have a sum of 1; more precisely, the sum of the absolute values $\sum|w|$ of the weights should be set at 1. So the weights should be divided by this sum, being 1.125, which gives:

	Q	L	w	wQ
B_1	.5	.375	.333	.167
B_2	1	<u>.75</u>	<u>.667</u>	<u>.667</u>
		1.125	1	.833

The weighted mean has gone up from .75 to .833. In principle, one should reiterate the procedure and see how the assessors should be weighted against the weighted mean, and so on; in elementary cases like this, however, convergence is immediate.

In weighting assessors, the absolute scale provides the means to assign optimal weights in the one- P design. As the weights w for each B are proportional to their $L_{QM} = QM$, in which M is the average score of the assessors in question, and as M is a constant in comparing these assessors, it appears that their weights are proportional to their ratings. Extreme ratings thus get higher weights; null ratings function like abstentions that are disregarded altogether; negative ratings get negative weights if the average is positive, and vice versa. For another example:

	Q	L	w	wQ
B_1	-.5	-.167	-.333	.167
B_2	0	0	0	0
B_3	1	.333	.667	.667
<i>Mean</i>	.33			.833

4.2. Feed-forward effects

From the point of view of finding an optimal weighting of the assessments in such single cases, or even in small-scale designs in general, little is gained since the weights have wide error margins. The more interesting question here is how weighting would influence the behavior of the assessors if they respond to the mechanics of the weighting rule; in other words, what are the feed-forward effects of weighting. According to the assessment ethic, assessors should maximize their correspondence with the final score. This implies maximizing their own weight, which is what self-respecting assessors would want to do for more mundane reasons. They can do so by submitting extreme ratings, provided that they expect the sign of their rating to correspond with the sign of M , even though they cannot be sure of that.

In evaluating the predictable feed-forward effects of weighting, it should be realized that unweighted averaging – that is, unit weighting – also has its feed-forward effects. The impact of an individual is still maximal upon giving an extreme rating. Here however, an extreme dissenting assessment is not sanctioned; rather, the assessor is invited to impose his or her point of view on the other panel members. Unit weighting is thus clearly incompatible with the assessment ethic, which would require that assessors try to take a universal point of view. This is by no means an academic issue: In the almost universal practice of unit weighting, there is abundant anecdotal evidence of authoritarian assessors who would have their private taste prevail, and get away with disrupting the assessment process and violating the assessment ethic. Weights according to agreement

provide an important safeguard against such transgressions, and appear do so without eliciting middle-of-the-road assessments.

4.3. A binary bipolar scale

As a consequence of the weighting rule, assessors would only assign ratings of ± 1 , with an occasional 0 if they are in perfect doubt whether they hit the consensus, so that the weighted average assessments (for which 0's do not count) would consist of extreme values only, without any nuances. As that scale use is wholly compatible with the assessment ethic, one might as well present the assessors directly with a binary bipolar $[+1, -1]$ scale, while preserving the opportunity to abstain (but not in the sense of taking an in-between position, since abstentions receive zero weights and therefore do not detract from or add to the grand mean, and are treated as missing data). Consequently, the assessment data would take the shape of a sign vector per P , with occasional empty cells.

The polarization effect is an outgrowth of adopting L . With continuous scales, that coefficient takes on low values compared to just about all other indices, because it does not correct for size. For example, if the scores on a bipolar scale were in the order of ± 0.5 , the identity coefficient would be 4 times as large as the likeness coefficient. However, the feed-forward effect of using L leads to the adoption of a binary bipolar scale in gathering the data, so that the difference is annihilated: with $[-1, +1]$ ratings, $L = e$. One might say that in the process, L overcompensates for its smallness, as the e -coefficient does not bring about the particular feed-forward effect, and would lead to smaller-sized ratings.

The polarizing effect does not do justice to subtle differences in quality. However, to pose that requirement would be to confuse assessment and selection, or absolute and relative judgment. The binary scale focuses on the central issue in assessment, which is a pass-fail appraisal. Nuances are important only if a further comparative selection has to be made from P 's of positive quality. But for that purpose, absolute scales are irrelevant. So the polarization underlines the fundamental distinction between absolute and relative judgment with a vengeance. For practical purposes, one should install separate procedures for assessing a quality and subsequently selecting the best P 's.

5. CONCLUSIONS

The present analysis leads to a simple and internally coherent set of conclusions. First, weighting of assessors in accordance with the likeness of their ratings to the mean, using a binary bipolar scale, comes down to a simple majority rule: If there are more positive than negative ratings, the final assessment is $+1$ in all cases, and vice versa; any abstentions are discounted. This rule preempts an inelegance that would otherwise have

to be dealt with: Reversing the sign of an assessor's rating is a rude way of treating *B*, that *A* would be wise to avoid. A simple majority rule, however, is hardly objectionable.

Second, the binary absolute scale appears to coincide with the most undemanding scale of all, an elementary nominal scale with two categories, the modal response to which counts as the final assessment (assuming that the probability of an abstention is so small that its preponderance would constitute a pathological case). This 'downscaling' should be welcomed by those who would doubt whether assessors are actually capable of giving judgments on a strong scale. Note that the nominal interpretation was reached without assuming that they are not. Such a negative assumption would have simplified the analysis considerably, but it would have been as strong as its counterpart. In the best of all possible worlds, positive and negative assumptions (even including ordinal versus nominal assumptions, which were not explicitly discussed) come down to one and the same thing in the end.

The third uneasiness that is resolved in the end is the dilemma between the identity and likeness interpretations of the concept of association: With the binary bipolar scale, L and e have themselves become identical. Again, this is not to say that the analysis could have done without the special properties of the likeness coefficient. But the final result allows us to have our cake (likeness) and eat it (identity), without having to accept the conceptual drawbacks of either: the consequence that a non-extreme rating is not completely like itself ($L_{QQ} < 1$), or the puzzling idea that identity can be a matter of degree.

Finally, and most important of all, the analysis was performed without assuming at any time that the quality of any P is comparable to the quality of another P . Comparing assessors was argued to be both inevitable and justified by the very concept of quality, but all arguments were kept intra- P . That possibility arose as a more or less serendipitous spin-off of using coefficients of absolute agreement, particularly, L . However, it is of great significance in situations in which qualities of different P 's are judged against a standard, as opposed to comparing them to each other. First, it is a fact of life that different P 's may pass or fail for quite different reasons. One person may be fit for a job because she is a good problem solver, another because he is good at keeping people together. It is not at all clear if meeting or failing a common standard implies mutual comparability. Second, assessment is a dynamic or historic process. It does not deal with passive things but with human reality. P 's are to some extent assessed on their own merits, rather than strictly bureaucratically. That means no less than an interaction between P and the standard itself; it could even be argued that the essence of a creative

product or performance is that it uproots standards of quality. With wooden standards, human history would come to an end.