Promoting intersubjectivity: a recursive-betting model of evaluative judgements

Willem K.B. Hofstee

Evaluative judgement aims at intersubjectivity, rather than objectivity or subjectivity. Multiple judges are widely used to increase its common component. A practical question is how to aggregate the individual judgements, in view of the feed-forward effects that are elicited by aggregation policies. The traditional practice of taking the mean of the individual judges' votes, ratings, or rankings invites improper strategic manoeuvring by judges. The development of a set of more proper aggregation rules requires an analysis of the judgement task, as distinct from measurement, forecasting, preferential choice, and voting. Using psychometric notions, I model judgement as a recursive bet aimed at maximising the common or intersubjective component in the individual judgements. For practical purposes, the analysis leads to a number of recommendations, in particular: Take the median, not the mean, in aggregating individual evaluative judgements, and inform judges about its rationale. (*Netherlands Journal of Psychology* 65, 2-9.)

Keywords: intersubjectivity; common component; recursive bet; scoring rules; majority rule; median

Judgements of quality are ubiquitous, for example, in peer review of proposals and products, quality assessment of public and private services and institutions, personnel assessment, grading of student performance, evaluations of curricula, policies and programmes, and so on. Evaluative judgements of the kind that are at issue here differ from subjective preferences that find their expression in a consumer market, or political voting conceived as expressions of private or group interests: Judges respond to authority (which in turn has a broader responsibility) that commissions the evaluations and takes the decisions. Evaluations are also not objective, in the manner of a measurement, if only because of disagreement between exchangeable judges. However, they are to some extent intersubjective. Typically, authorities enlist multiple judges to promote intersubjectivity by averaging out idiosyncrasies.

My practical focus is on the problem of finding appropriate rules for combining the individual judgements. A traditional practice is to take the unweighted mean – whether formally or in an informal manner – of the judges' ratings, rankings, or votes. In face-to-face committees, con-

The Heymans Institute, University of Groningen, the Netherlands
Correspondence to: Willem K.B. Hofstee, The Heymans Institute, University of Groningen, Grote Kruisstraat 1/2, NL 9712 TS Groningen, e-mail: w.k.b.hofstee@rug.nl
Submitted 5 August 2008; revision accepted 3 December 2008.

sensus is sometimes required. Both procedures invite strategic manoeuvrings by the individual judges, for example, upgrading their preferred candidates, and downgrading ('burying') candidates that they expect to be competitors to the ones they prefer. Co-judges, authorities, candidates, and bystanders may thus come to mistrust the result. Such strategies and misgivings may be counteracted by using more proper (see, e.g., Hofstee, 1984) aggregation procedures.

The derivation of proper aggregation rules requires a theoretical analysis of the judgement task, as distinct from measurement, forecasting, preferential choice, and voting. In attempting this analysis, I take a perspective in which psychometric principles serve to model the relation between involved parties, primarily, the judgement authority and the individual judges. What emerges is a conception of judgement, quality, intersubjectivity, and related notions, as the principal common component that results from a 'recursive bet', that is, a competition among judges to be won by the judge who best predicts their common prediction.

I do not propose literally following rules for combining judgements, for example, in face-to-face committees in which decisions tend to come about through informal processes, which are generally found valuable in themselves. Rather, the point is that models inevitably play a role in these processes, and that they make a difference. For example, if a majority of the committee members finds a particular research proposal excellent, whereas a minority gives it a negative rating, chances are that on balance, the proposal will not get the highest priority. That compromise would be modelled by an unweighted-mean rule. If judgements were to be weighted, formally or informally, according to their correspondence with the others, the minority ratings would be discounted and the proposal would get higher marks. Models are influential even if they are not purposely applied, maybe more so if they are taken for granted. More generally, I recommend models as intellectual tools rather than administrative straightjackets.

Analysing evaluative judgement tasks

Judgement versus measurement

Judgemental data bring about a conception of data treatments (e.g., taking an unweighted mean) as scoring rules (De Finetti, 1965) or policies (Hofstee, 1984), that is, treatments that elicit strategic behaviour by the rational respondent, in this case, the judge. With objective, even if fallible, measuring devices this conception would be out of place: A thermometer, for example, does not let its behaviour be influenced by the diagnostician's expectations or agenda. Ideally, of course, judges would be equally impartial and insensitive to the way their evaluations are processed. In practice, however, judges may be expected to be human. For example, if a conjunctive rule is used, so that a judge's dissenting vote will have veto power, then that judge cannot even escape considering his or her judgement in view of its consequences. Generally, scoring policies have feed-forward effects on judgemental data.

The emphasis in developing aggregation rules should be on taking away any temptation to submit idiosyncratic and biased judgements, rather than just correcting for biases after the fact. Several authors have presented corrective aggregation procedures, most notably, taking the median rank rather than the mean rank – which will appear to agree with one of my recommendations. Basset and Persky (1994) point to the use of median ranks in the judging of figure skating. On the basis of a simulation study, they conclude that median ranks are superior to biased judgements, without loss of efficiency in finding the winner. (As it happens, the rule was abandoned after a scandal involving the 2002 Winter Olympics title, but that does not detract from Basset and Persky's argument). Hurley and Lior (2002) simulate situations in which some judges rank their favourite candidate first, and reverse their rankings of the other candidates to maximally bias the expected outcome; these authors, too, find median ranks superior to partially trimmed means (obtained by discarding pairs of extreme scores, the untrimmed mean and the median being the limiting cases). Moreover, the loss of efficiency through the use of medians appeared to be negligible in their simulations. In an overview on combining forecasts, Armstrong (2001) cites evidence for the median's superiority over the mean's, especially by the criterion of mean absolute error (rather than root mean squared error). First, however, these demonstrations do not take away the sober fact that the median, though more robust against outliers and bias, is less efficient than the mean in normal distributions, which are more likely to arise in the absence of bias. Consequently, one would only opt for trimming when suspecting improper strategic manoeuvrings by the judges. However, competent and conscientious judges might not like the prospect of seeing their judgements discarded for that reason. Authorities might try to impress upon them that policing protects against others' biases rather than their own. However, that Machiavellian argument may not be appreciated in a collegial setting, if it is admissible at all. One would prefer a positive argument to accept the median, which is what I will present. Second, all these demonstrations presuppose that the true quality of a candidate (a skater, a forecaster) is known, if only to the experimenter running the simulation. I argue against that presupposition in the context of evaluative judgement.

Judgement versus forecasting

Evaluative judgements apply to situations in which objective criteria are insufficient. Judges are enlisted to look for something emergent about the candidate's plans or products, which by definition refuses to be caught in established terms; candidates are supposed to be judged on their merit, in other words, they have a part in shaping the very standards by which the quality of their plans, products, and performances will be judged. Objective indicators of quality may support the judging – albeit at the risk of eliciting improper strategic behaviour by candidates – but quality judgement is ultimately a human affair. Put differently, there would be no excuse for using judgements if a particular quality were to be completely covered by a set of objective indicators; for, in such cases objective procedures will be both more valid and efficient than fallible judgements. I thus take it that complete coverage cannot generally be achieved, so that judgements have their place.

Insufficiency of external criteria implies that quality judgements are not to be viewed as predictions or forecasts, unless one wishes to stretch that concept. Reviewing a research proposal, for example, may be constructed as predicting the quality of the proposed research; reviewing a scientific paper may be said to imply a prediction of its contribution to science. However, such criteria are judgemental in their turn; one cannot meaningfully validate an evaluation against a factual future outcome, in the manner of checking a weather forecast or a speculation on the future price of stock. In practice, judgements have criterion rather than predictor status. Therefore, the criterion for individual judgements can only be internal: It relies in the common component of the judgements of the relevant population of judges. The criterion for judgements is in their representativeness, rather than their predictive accuracy. Or, if one prefers a predictive phrasing: The judges' task is to predict the common component in the relevant population of judges.

A familiar objection to adopting internal criteria comes under the label of a Keynesian beauty contest. In a famous passage, Keynes (1973, p. 156) compared the stock market with newspaper competitions in which readers nominated the most beautiful among a number of photographed women. The prize went to a contestant who selected the photographs that were nominated most often. Obviously, the procedure inspires second-guessing of what others would find, rather than expressing one's personal taste. Keynes' problem was that the price of stock seems to be a result of such second-order (or even higher-order) guessing, so that it may become volatile and may have little to do with – what he calls – the intrinsic value of the stock. With respect to the stock market, Keynes' misgivings were undoubtedly to the point, if only because he is known to have made large fortunes on that psychological basis. However, it is difficult to see how they would apply to the beauty contest itself or to more consequential quality judgement. For one thing, one would have to entertain a notion of intrinsic beauty or quality as distinct from what people find, thereby claiming a superhuman position; otherwise, second-guessing can hardly lead to anything but increased validity. For another, an actual beauty contest does not make up the kind of dynamic system in which aberrations get multiplied by positive feedback. The lesson for multiple quality judgements in general is that they should be formulated independently, to avoid contamination and the ensuing loss of degrees of freedom.

Another way to phrase misgivings regarding beauty contests is to say that unusual, novel, and creative plans, products or performances would have no chance of being acknowledged, and that the outcome of the judgement process would be conservative, conformist, risk-aversive, and the like. The underlying reasoning must be that enough individual judges might personally value such unusual candidates (otherwise, there is no issue), but that they either would not trust their fellow-judges to do so, or would second-guess that the others would follow the same mistrustful strategy. The judge could be wrong about this, of course, and could thus lose the bet, which would be all the more of a pity. In any event, the mistrustful and meta-mistrustful strategies are not *a priori* rational, apart from being disloyal to colleagues. Undoubtedly, a defiant attitude has its place if an objective truth is at stake, as in whether a defendant is guilty or not. In quality judgements, however, any appeal to an ultimate truth would rather be presumptuous.

Still another objection against internal criteria would be that judges have no other basis for predicting the joint outcome than their own preferences and dislikes, in other words, that all thinking is wishful. That argument is obviously incorrect in the critical case in which judges wittingly manoeuvre to advance their favourite sons or daughters at the expense of more likely candidates: To do so, they must have an idea of the common component. In other cases, one would have to assume that all judges are completely unaware of their own prejudices and idiosyncrasies, which is patently unrealistic. In a more plausible representation, an individual quality judgement consists of a common and an individual component, orthogonal to each other by definition as in classical true-score theory (Lord & Novick, 1968) or in Spearman's (1904) two-factor model, with the added assumption that judges can distinguish between the two at least to some extent, so that they can increase the relative contribution of the common component if motivated to do so.

Judgements versus preferences

Parallel to the distinction between common and individual components, judges are presumably subject to mixed inclinations, ranging from the kind of self-will whereby one wishes to force his or her private preferences upon others, to professionalism in the basic sense of wishing to maximise intersubjectivity. Consumers may be at liberty to exert their own market preferences and political voters may be entitled to pursue and express their own interests[1]; however, unlike consumers and voters, judges respond to an authority. Presumably, that agent is exclusively interested in the common component of their judgements, and must regard individual components as noise almost by definition. From the authority's point of view, judges are called upon to maximise their representativeness. In this respect, certain requirements of a fair voting system are of lesser relevance in judgement tasks. Particularly, professional judges may accept that they have been wrong, not only in the sense of being fallible or subject to all sorts of random error, but also directionally: They may admit that they have been unable, in a particular case, to abstract from their idiosyncrasies and biases, and therefore agree to having their judgement reversed.

Such a negative weighting runs counter to Arrow's (1951) monotonicity or 'positive association of social and individual values' requirement. That requirement is altogether reasonable in the context of reconciling individual preferences, but out of place in combining judgements of quality. This is not to say that formal results from social preference and voting theory are irrelevant to the theory of quality judgement. In particular, Arrow's impossibility theorem, which states that one cannot find aggregate preferences over three or more candidates in such a way that the aggregate always satisfies a number of reasonable criteria, does not appear to hinge on the monotonicity requirement *per se* (Arrow, 1963). However, questions like whether admitting negative weights lifts the impossibility theorem are outside the scope of this paper. Here, the montonicity issue only serves to accentuate the fundamental difference between preferential choice and evaluative judgement.

1 *Voting theory traditionally adheres to the sort of economic paradigm whereby maximum collective utility is supposed to result from the pursuit of private or group interests, under certain conditions (e.g., one person one vote). However, one might conceive of an alternative and maybe more appropriate reconstruction of political voting, whereby citizens judge the potential contributions of programmes and politicians to the quality of a society, from an intersubjective perspective.*

Maximising representativeness through weighted aggregation

The conclusion from the analysis is that professional judges – as distinct from measurement devices, consumers, political voters, and so on – should welcome or at least accept procedures whereby their judgements are weighted and scored for representativeness. Representativeness may be maximised by weighting individual judgements according to their correspondence with a common component. The appropriate weighting model is Principal Component Analysis (PCA), in this case, raw-scores PCA. Its logic, in the present context, is best understood by following the power algorithm (see, e.g., Horst's 1964 text) to construct the first principal component. The algorithm runs as follows: In a first approximation, one would weight the individual judgements according to their correspondence with the (unweighted) mean of these judgements. By this weighting, however, the mean shifts into a weighted mean. Therefore, the individual judgements should be weighted according to their correspondence with the weighted mean; in other words, the procedure is iterated. Iteration continues until convergence is reached, that is, until the weights and the weighted mean after the last iteration no longer differ (according to a threshold criterion) from the weights and mean after the before-last iteration. The end result consists of principal component weights and weighted means.

Through PCA, representativeness in terms of the size of the common component is maximised. In the absence of an external criterion, judges' weights are defined in a recursive manner, as the weights that maximise the over-all correspondence of the individual judgements with their weighted sum. Implicitly, the judgements are split up into a common part plus a noise part, orthogonal to each other by virtue of the PCA model. Empirically, the judgement matrix may contain more than one principal component, if noise is shared by sub-groups of judges. For practical purposes, however, such group components cannot define quality any more than idiosyncrasies can, and are thus likewise disregarded.

A judge's score – which would add to or detract from his or her reputation as a competent judge – is to be based on the correspondence between the individual judgement and the common component. I take it that judges would be motivated to maximise their score, not so much because of any consequences attached to it, but rather because it symbolises a proper conception of their task. Note that scoring based on correspondence with the unweighted mean, though resulting in closely comparable scores in most cases, would be inferior from that symbolic point of view: Judges should not be asked to predict some average opinion, but to identify with the most representative judges, a distinction that is comparable

with the difference between direct and representative democracy.

Ideally, individual judgements should be scored and weighted against the common component of the relevant population of judges. In practice, a sample in the shape of the panel or committee in question will have to suffice. To be acceptable, that solution presupposes optimal representativeness of the sample of judges, therefore: balanced composition, independent judging, and abstention rules in case of personal involvement. A prerequisite of a different order consists of securing informed consent by the individual judges regarding the way the committee or panel is composed. In more general terms, the situation should be arranged so that judges trust each other's evaluations, even if they happen to disagree.

A question that lingers at the background is how the population of relevant judges is to be conceived. At the universalistic extreme, it comprises all humans of all times including the past and the future; at the other end of the spectrum, a handful of experts. However, any urge to take sides in an egalitarian *versus* elitist debate on quality may be moderated by the prospect of weighting judgements, including abstentions (null weighting): In judging the quality of Mozart's 40th symphony, for example, many members of the world's population would abstain; others would receive negligible or negative weights, provided enough other judges were to tune in to the relevant elite's taste. (Formally speaking, egalitarianism is one limiting case of weighted judgement, as egalitarianism presupposes unit weighting; the other is elitism, in which weights are unity for the elite and zero for the rest of the population). Of course, there is no guarantee against populist outcomes: At the turn of the Millennium, the German public decided upon Konrad Adenauer as the greatest German of all times, to the bewilderment of most foreigners (and German elites). In the same vein, the Dutch chose Pim Fortuijn. In these cases, however, the problem evidently arose from restricted (intra-national) sampling, so technically speaking the procedure was elitist rather than egalitarian. More properly elitist aberrations may be observed in expert panels of all kinds as a result of certain social-psychological processes leading to shared preconceptions.

Developing weighting procedures

In the most elementary case, pass-fail judgements have to be made. Such judgements may be represented as +1 (positive) and –1 (negative). For the validating and weighting of an individual judgement against the panel or committee average, an appropriate coefficient of correspondence is needed. The simplest index of the correspondence between [–1…+1] variables is their mean product, named Likeness coefficient *L* by Hofstee and Ten Berge (2004); other coefficients of association for scales with a fixed midpoint take the same shape for the [+1, –1] scale (see Appendix 1). In the single case, *L* is simply the product of the two values; here, an individual judgement and the committee average. Evidently, the validities for the judgements in the binary case differ only in sign: If, for example, 4 judgements are positive and 1 is negative, the unweighted mean is (4–1)/5 = 0.6, and the validities are +0.6 and –0.6, respectively. Upon weighting proportional to validity, with the sum of the absolute weights set at unity to keep the weighted mean on the same scale, the respective weights in the example become +0.2 and –0.2. The weighted mean becomes +1; generally with [+1, –1] data, the weighted mean can only be +1 or –1. Likewise, a judge's score consisting of the product of that judge's rating and the weighted sum can only be +1 or –1 (see Appendix 2).

Thus for [+1, –1] judgements, the weighting procedure takes the shape of the simple majority rule; in fact, its development may be viewed as a way to model that rule – broadly applied in parliamentary –democratic contexts – by which decisions taken by the smallest possible majority have a status equal to unanimous decisions. This is quite different from taking the unweighted mean, by which the aggregate judgement is diluted through dissenting votes (and abstentions), and candidates may pass or fail more or less. The feed-forward effect on the judge is also radically different: To receive a maximum score, the individual judge should try to hit the majority judgement; conversely, under the unweighted mean rule judges are tempted to oppose the majority if that fits their personal or group preferences. Under the unanimity or conjunctive rule, that temptation is even maximised.

Ratings and rankings

In many judgement tasks, binary judgements do not suffice. Ratings and rankings are much more frequently used, mostly to prepare for a comparative selection following the judgement phase. One would therefore think of generalising the weighting procedure to a continuous [–1…+1] scale, that is, validating and weighting continuous-scale individual judgements against their weighted average (see Appendix 3). However, the generalised procedure does not appear to work. The most important reason is that judges could maximise their expected score by giving extreme (+1 or –1) ratings (see Appendix 4). Consequently, the scale would more or less automatically degenerate back into a dichotomy in practice.

An effective and much simpler solution consists of taking the gradations or ranks as cumulative pass-fail thresholds. As an example, take a five-point scale from excellent to poor, coded A to E, with A>B>C>D>E. The scale has four thresholds; a B candidate fails the A/B threshold

and passes the B/C and lower thresholds, an E candidate fails all thresholds, and so on. So a judgement may be represented as a vector of [–1,+1]scores on threshold variables, as follows:

	Threshold:			
	E/D	D/C	C/B	B/A
A	+1	+1	+1	+1
B	+1	+1	+1	–1
C	+1	+1	–1	–1
D	+1	–1	–1	–1
E	–1	–1	–1	–1

Take the following case, in which seven judges would have evaluated a candidate, with judgements E, B, D, D, B, B, and A; in dummy-variable form, with the judgements as rows:

	E/D	D/C	C/B	B/A
E	–1	–1	–1	–1
B	+1	+1	+1	–1
D	+1	–1	–1	–1
D	+1	–1	–1	–1
B	+1	+1	+1	–1
B	+1	+1	+1	–1
A	+1	+1	+1	+1
Score	+1	+1	+1	–1

The bottom row shows the aggregate judgements of the candidate according to the weighted-average (majority) rule. The candidate passes the threshold for a B qualification, as the majority of the ratings are at or above B; he or she falls short of an A qualification. The cumulative-thresholds procedure amounts to finding the median rating per candidate. No quantification is involved, as in taking the unweighted-mean rating, which would be below the B/C threshold in this case. The latter result might have come about because the first judge decided to 'bury' the candidate; the median rule ignores such an extremity in judgement.

With rankings, the median rule applies identically. The example could have come from a ranking of five candidates by seven judges, with one candidate receiving:

5	2	4	4	2	2	1

Here, the candidate passes the threshold for the second position.

Obviously, the median rule may produce tied rankings between candidates, necessitating supplementary procedures that may bring about complications. From the point of view of voting theory, still other reservations may be brought forward (see, e.g., Gehrlein & Lepelley, 2003). Here, however, the emphasis is on deriving the rule from the premises of a judgement theory, which are different from those of preferential voting or social choice. From personal experience in committees, I expect the median rule to be quite feasible in practice; but its main reason of existence relies in operationalising an intersubjective conception of evaluative judgement, discouraging purposely deviant positionings.

The implied way of scoring the individual judgements against the panel outcome is to give $1/R$ points for each correct prediction regarding a threshold, with R the number of thresholds (i.e. R=4), and deduct $1/R$ points for each failure. For example, judges who assigned a D rating obtain +¼ for the E/D threshold, as the final judgement (B) implies that the candidate is indeed above it; they score –¼ for the D/C and C/B thresholds as their predictions were wrong; for the B/A threshold, the score for their –1 judgement is +¼ as the panel judgement for that threshold is also –1. So their score is +¼–¼–¼+¼ = 0. The B judgements, of course, get full marks, or +1. The general expression for the judge's score is $1 - 2D/R$, with D the absolute distance between two ratings, and R the range (or the maximum D, which equals the number of thresholds). This index appears to be the special, $N = 1$, case of g' (Hofstee and Zegers, 1991; see also Appendix, 1). The invitation to the judge is clear: Try to minimise the mean absolute distance to the others, that is, try to hit the median rating. The g' coefficient for $N > 1$ is appropriate for analysing matrices of judgement data.

Conclusion

The analysis implies an alternative rationale for using the median, here conceived as an element in an ordered sequence of thresholds. The median rule emerged from a reasoning by which data are weighted according to their correspondence with a weighted mean. That reasoning, in turn, was based on the central proposition that judges are committed to aim for representativeness.

The analysis leading to the majority and median rank or rating rule is situation-specific. It does not lead to a recommendation for consumers or members of trial juries to opt for the middle. In official judgements of quality, however, professional actors may be expected to accept the set of rules as a proper expression of the task structure.

Next to pragmatic goals, the purpose of this paper is to model the concept of intersubjectivity in a hopefully coherent manner. Taking a com-

mon point of view is basically different from maximising individual utility, which serves as the take-off point for branches of economics such as voting theory. Intersubjectivity is intimately associated with notions such as citizenship, responsible professionalism, and representative democracy, which deserve study and development in their own right. Finding relevant conceptual tools should contribute to promoting intersubjectivity.

Acknowledgments

I am greatly indebted to Jos M. F. ten Berge, Henk A. L. Kiers, and Ivo Molenaar for their comments on earlier drafts.

References

Armstrong, J. S. (2001). Combining Forecasts. In J. S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp. 1-19. Norwell, Ma: Kluwer.

Arrow, K. J. (1951, 1963). *Social Choice and Individual Values*. New York: Wiley.

Basset, G. W., & Persky, J. (1994). Rating Skating. *Journal of the American Statistical Association, 89,* 1075-1079.

De Finetti, B. (1965). Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item. *British Journal of Mathematical and Statistical Psychology, 18,* 87-123.

Gehrlein, W. V., & Lepelley, D. (2003). On some limitations of the median voting rule. *Public Choice, 117,* 177-190.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of its Properties. *Biometrics, 27*, 857-871.

Hofstee, W. K. B. (1984). Methodological Decision Rules as Research Policies: A Betting Reconstruction of Empirical Research. *Acta Psychologica, 56*, 93-109.

Hofstee, W. K. B., & Ten Berge, J. M. F. (2004). Personality in Proportion: A Bipolar Proportional Scale for Personality Assessment and its Consequences for Trait Structure. *Journal of Personality Assessment, 83*, 120-127.

Hofstee, W. K. B., & Zegers, F. E. (1991). Idiographic Correlation: Modelling Judgments of Agreement between School Grades. *Tijdschrift voor Onderwijsresearch, 16*, 331-336.

Holley, J. W., & Guilford, J. P. (1964). A Note on the G-Index of Agreement. *Educational and Psychological Measurement, 24*, 749-753.

Horst, P. (1964). *Factor Analysis of Data Matrices.* New York: Holt.

Hurley, W. J., & Lior, D. U. (2002). Combining Expert Judgement: On the Performance of Trimmed Mean Vote Aggregation Procedures in the Presence of Strategic Voting. *European Journal of Operational Research, 140*, 142-147.

Keynes, J. M. (1973). *The General Theory of Employment, Interest and Money.* London: MacMillan.

Lord, F. M., & Novick, M. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology, 15*, 201-293.

Tucker, L. R. (1951). A Method for Synthesis of Factor Analytic Studies (Personnel Research Section Report No. 984). Washington, D.C.: Department of the Army.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A Family of Association Coefficients for Metric Scales. *Psychometrika, 50*, 17-24.

Appendix 1

The Likeness coefficient L for variables (or single pairs of scores) X and Y is equal to $\Sigma XY/N$. Zegers and Ten Berge's (1985) Identity coefficient e is equal to $2\Sigma XY/(\Sigma X^2 + \Sigma Y^2)$; as $\Sigma X^2 = \Sigma Y^2 = N$ for [–1, +1] data, $e = L$. Similarly, Tucker's (1951) $phi = \Sigma XY/(\Sigma X^2 \Sigma Y^2)^{½} = L$ in the [–1, +1] case. The Agreement coefficient proposed by Holley and Guilford (1964) for 2×2 tables equals the proportion of agreements, found in the diagonal cells, minus the proportion of disagreements, found in the off-diagonal cells; for agreements on the [–1, +1] scale, $L = 1$, for disagreements, $L = –1$; the aggregate L equals the mean of the individual L's (this distributive property holds for L in general, as the expression shows); thus the Agreement coefficient equals L. Finally, Zegers (see, Hofstee & Zegers, 1991) presented a transformation of Gower's (1971) g coefficient, called $g' = 1 – 2\Sigma D/(NR)$, with D the discrepancy in a pair of scores and R the range of the scale. As $R = 2$ on the [–1, +1] scale, and D is either 0 (for agreements) or 2 (for disagreements), $g' = 1– 2q = p – q$, with p the proportion of disagreements and q the proportion of disagreements. So, like the agreement coefficient, $g' = L$, for [–1, +1] data.

Appendix 2

Logically, one would now wish to validate and weight the judgements against the weighted mean; in the single case, however, that iteration would lead to the same result. With more cases taken together, the iterative procedure would lead to taking the first principal component of the raw-scores matrix, with the sum of the absolute component weights set at unity (see, Hofstee & Ten Berge, 2004). However, for reasons of simplicity and because judgements may not form a matrix (as different sets of judges may operate on different cases), one may wish to profit from the fact that L and related coefficients for scales with a fixed midpoint are defined in the single case.

Appendix 3

The general expression for the weighted mean M_w in question is readily seen to be $M_w = Sign(M)M_{sq}/M_{abs}$: the sign of the algebraic mean M times the mean of the squared judgements M_{sq} divided by the mean of their absolute values M_{abs}, with $Sign(0)=0$.

Appendix 4

M_w is a function of three other means; it thus has an elegance of its own. However, its behaviour in function of gradual shifts in one of the judgements appears to be discontinuous and non-monotone (H. A. L. Kiers (personal communication, 31 March 2006), making it difficult for a judge to anticipate the effect of such shifts. What judges can anticipate, however, is that their expected score in function of their judgement X is maximal for $X = 1$ if they expect the aggregate judgement to be positive, and –1 if negative. That intuition can be proven to be correct along the following lines: A judge's score can be written as $X\,Sign(X + a)\,(X^2 + b) / (|X| + c)$, with a, b, and c the sum, the sum of squares, and the sum of the absolute values of the others' judgements, respectively; the score is positive if the signs of X and $(X + a)$ correspond, and negative if they do not. (In case other judges are so divided or undecided that $|X| > a$, so that the judge in question tilts the average judgement his or her way, the score is positive irrespective of the direction of a). For an extreme judgement $|X| = 1$, the score is $Sign(X)\,Sign(X + a)\,(1 + b) / (1 + c)$. Upon expansion, it appears that $(1 + b) / (1 + c) > X\,(X^2 + b) (|X| + c)$ for all $X<1$. So changing a judgement $1>X> –1$ into an extreme judgement yields a score gain d, with a subjective probability p, if the signs appear to correspond, and a loss – d, with probability $1 – p$, if they do not. Thus on balance, the gain on submitting an extreme judgement is $(2p – 1)$ d, which is positive with $p > 0.5$.