When Quality Is a Matter of Taste, Use Reliability Indexes

The Kappa and intraclass techniques work well for subjective measurements.

by
David Futrell

Most measurement processes in industry rely on gauges, weighing instruments, micrometers, or other devices that make fairly direct physical measurements of a product characteristic. A variety of measurement methodologies have been developed that deal effectively with these types of measures.[1] There are, however, many situations in which the measurements are subjective classifications or ratings made by people. Examples of this are quite common in both industry and sports:

- Classification of faulty fabric by defect types
- Classification of garments as "good" or "bad"
- Quality of bourbon rated on a 1-to-100 scale
- Sorting of computer touch screens into "sticks" and "doesn't stick"
- Classification of the presence or absence of certain characteristics of wine
- Ratings of automobile handling characteristics on a 1-to-5 scale
- Assessment center performance ratings on a 1-to-9 scale
- Ratings of Olympic platform divers on a 1-to-10 scale

What these situations share is quality criteria that are either difficult or impossible to define. For example, there is no way to obtain a true score for the quality of bourbon or wine, yet trained tasters can rate and classify these with remarkable agreement.[2] Other ratings, such as the classification of garments by defect type, might have a true criterion, but one needs a way to tell whether all raters are using the same operational definitions for classification. In these situations, a different approach to assessing the quality of the measurement system is required.

To assess the extent to which classifications or ratings are made on a consistent basis, several units must be classified more than once and by more than one rater or judge. If there is substantial agreement among the raters, there is the possibility, although no guarantee, that the ratings are accurate. If there is poor agreement among the raters, however, the usefulness of the ratings is extremely limited. How can one expect to find out what qualities are associated with a set of ratings if these raters don't even agree among themselves?

Psychologists and statisticians, particularly those in industry, have been confronted with these types of measurement problems for years and have developed several reliability indexes to deal with these situations. Although there are many of these techniques, which vary subtly according to their applications, there are two families of techniques that can be used to deal effectively with most situations encountered in industry. For classifications of nonquantitative (attribute) data, Kappa techniques are most appropriate. For ratings made on some type of scale (e.g., 1 to 10), a set of methods based on the intraclass correlation is the correct choice.

Kappa techniques: purely nominal classification

If a measurement system is needed to classify objects in a nonquantitative manner, the Kappa techniques are appropriate.[3] Examples include classifying objects as good or bad, classifying fabric flaws according to defects (such as scrimp, color mismatch, or misprint), and differentiating noises (such as squeak, clank, or thump). If it is possible to order, rank, or scale objects in some way, such as "really bad," "pretty bad," "pretty good," and "perfect," the intraclass techniques (covered in the next section) should be used. The reason lies in the consequences of misclassification. Kappa techniques treat all misclassifications equally. The intraclass techniques do not: For example, the consequence of misclassifying a perfect object as "really bad" is much more serious than classifying a perfect object as "pretty good."

Since Kappa techniques are used on categorical data, they do not assume that ratings are equally distributed across the possible range; in other words, some categories can be used much more frequently than others. Kappa techniques require only that the units be independent, that the judges or raters make their classifications independently, and that the categories be mutually exclusive and exhaustive.

To assess agreement for a nominal scale, one needs to know only two pieces of information:

- The proportion of units classified in which the judges agreed, or ($P_{observed}$)
- The proportion of units for which one would expect agreement by chance, or (P_{chance})

Kappa (K) is defined as the proportion of agreement between raters after agreement by chance has been removed. The formula for Kappa is:

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

Consider these data:

Part	Judge A	Judge B
1	Good	Good
2	Good	Good
3	Good	Good
4	Good	Bad
5	Good	Good
6	Bad	Bad
7	Good	Good
8	Good	Good
9	Good	Good
10	Bad	Bad
11	Good	Good
12	Bad	Bad

First, the data are converted into proportions and then put into a contingency table as shown in Table 1. $P_{observed}$ is the sum of the probabilities on the diagonal (0.667 + 0.25) = 0.917. To get P_{chance}, the summary probabilities for each classification for each judge are multiplied and then summed: (0.667 x 0.75) + (0.333 x 0.25) = 0.5835. Kappa, then, is:

$$K = \frac{0.917 - 0.5835}{1 - 0.5835} = 0.8$$

There is obvious similarity between this statistic and chi-square. Chi-square can, in fact, be computed along with Kappa for any square contingency table. The statistics, however, have different purposes: Chi-square determines whether a relationship exists between categorical variables, while Kappa assesses the degree of the relationship.

The maximum value for Kappa is +1, which indicates perfect agreement. The lower limit for Kappa is more complicated and can range from –1 to zero, depending on the marginal distributions. A Kappa of zero means the agreement is the same as would be expected by chance. In the example just given, K = 0.8. The lower limit for an acceptable Kappa value (or any other reliability coefficient) varies depending on many factors, but as a general rule, if it is lower than 0.7, the measurement system needs attention. The problems are almost always caused by either an ambiguous operational definition or a poorly trained rater. Reliability coefficients above 0.9 are considered excellent, and there is rarely a need to try to improve beyond this level.

Kappa might be tested for significance, with the null hypothesis being that K = 0. To test for significance, one must first compute σ_{K_0}, which is:

$$\sigma_{K_0} = \sqrt{\frac{P_{chance}}{N(1 - P_{chance})}}$$

Thus, for the example previously given, σ_{K_0} is:

$$\sqrt{\frac{0.5835}{12(0.4165)}} = 0.342$$

K is then divided by σ_{K_0} to obtain a critical ratio to be compared to the normal curve. Again referring to the example, the critical ratio is:

$$\frac{0.8}{0.342} = 2.33$$

This indicates that K is significant at $p < 0.05$.

Testing K for significance, however, is usually trivial since one expects to find values much greater than zero. There might be situations, however, in which the significance test can serve as an absolute lower threshold.

The given example is the simplest case for Kappa. It can also be applied in cases of multiple raters and multiple categories, although the calculations become more complex.[4] Consider the

Table 1. Contingency Table

		Judge A: Good	Judge A: Bad	
Judge B	Good	0.667	0.000	0.667
	Bad	0.080	0.250	0.333
		0.750	0.250	

Table 2. Five Raters Evaluate 10 Carpet Seams

		Gap too large	Gap too small	Seam frayed	Seam uneven	Seam perfect	$\sum_{i=1}^{5} x_{ij}^2$
Sample	1	0	0	1	0	4	17
	2	2	0	1	0	2	9
	3	3	0	0	2	0	13
	4	0	0	0	0	5	25
	5	0	2	3	0	0	13
	6	4	0	0	0	1	17
	7	0	4	1	0	0	17
	8	0	0	0	5	0	25
	9	0	0	0	0	5	25
	10	3	2	0	0	0	13
		12	8	6	7	17	174
	$\bar{p}$	0.24	0.16	0.12	0.14	0.34	
	$\bar{q}$	0.76	0.84	0.88	0.86	0.66	

example in Table 2, in which five raters independently evaluate 10 different carpet seams and classify them according to their defects. The number in each cell is the number of raters who classified the product into each category. Since there are five raters, the rows all have a total of 5.

For this situation, two types of Kappa coefficients can be calculated. First, one can compute an overall Kappa, which is an assessment of rater agreement across all categories. Second, one can compute individual Kappa values for each category. This will reveal the categories in which raters have trouble agreeing.

The formula for the overall Kappa with multiple raters is:

$$K_{overall} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j \bar{q}_j}$$

where n = number of units, m = number of raters, k = number of categories, $\bar{p}$ = ratings within category/(n x m), and $\bar{q} = 1 - \bar{p}$

This is a rather intimidating formula, but if one has an example to work from, it is not difficult to compute:

$$K_{overall} = 1 - \frac{(10 \times 5^2) - 174}{10 \times 5 \times 4[(0.24 \times 0.76) + (0.16 \times 0.84) + (0.12 \times 0.88) + (0.14 \times 0.86) + (0.34 \times 0.66)]}$$

$$= 1 - (76/153.44)$$

$$= 0.5$$

The individual category Kappa is even messier. To make it a little easier, the numerator and denominator can be calculated separately:

$$K_{category} = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\bar{p}_j \bar{q}_j}$$

For example, the numerator for the "gap too large" category is:

$$[0 \times (5-0)] + [2 \times (5-2)] + [3 \times (5-3)] + [0 \times (5-0)] + [0 \times (5-0)] + [4 \times (5-4)] + [0 \times (5-0)] + [0 \times (5-0)] + [0 \times (5-0)] + [3 \times (5-3)] = 22$$

The denominator for the "gap too large" category is:

$$10 \times 5 \times (5-1) \times 0.24 \times 0.76 = 36.48$$

$$K_{category} = 1 - \frac{22}{36.48} = 0.4$$

The numerators and denominators for all of the categories are shown in Table 3. The individual Kappa values range from 0.15 to 0.75. This means that agreement among raters is good for "seam uneven," marginal for "seam perfect," and unacceptably low for the remaining categories. Improving this measuring system would likely require changing operational definitions for each defect type, retraining the raters, or both.

Table 3. Numerators, Denominators, and Kappa Values for Seam Categories

Category	Numerator	Denominator		Kappa
Gap too large	22	36.48	1 – (22/36.48)	0.40
Gap too small	16	26.88	1 – (16/26.88)	0.40
Seam frayed	18	21.12	1 – (18/21.12)	0.15
Seam uneven	6	24.08	1 – (6/24.08)	0.75
Seam perfect	14	44.88	1 – (14/44.88)	0.69

Table 4. Judges' Ratings of Fabrics' Print Quality

	Judge 1	Judge 2	Judge 3	SUM	SUM^2
Fabric 1	5	7	7	19	361
Fabric 2	4	3	2	9	81
Fabric 3	4	2	3	9	81
Fabric 4	6	7	8	21	441
Fabric 5	5	5	5	15	225
SUM	24	24	25	73	1,189
SUM^2	576	576	625	1,777	

Sum of all squared ratings: 405.00
Average of all ratings: 4.87
Sum x average (73 x 4.87): 355.27

Sums of Squares:		DF*	Mean squares	Component
Judges (1,777/5 – 355.27):	0.13	2	0.07	JMS
Between fabrics (1,189/3 – 355.27):	41.07	4	10.27	BMS
Total (405 – 355.27):	49.73	14	3.55	
Within fabrics (49.73 – 41.07):	8.67	10	0.87	WMS
Error (49.733 – 41.07 – 0.13):	8.53	8	1.07	EMS

*** Degrees of freedom calculations**
Judges = number of judges – 1 = 2
Between fabrics = number of fabrics – 1 = 4
Total = (number of raters x number of fabrics) – 1 = 4
Within fabrics = DF total – DF between fabrics = 10
Error = (number of raters – 1) x (number of fabrics – 1) = 8

Key:
BMS = Between mean square
DF = Degrees of freedom
EMS = Error mean square
ICC = Intraclass correlation
JMS = Judges' mean square
WMS = Within mean square

Intraclass correlation: assessing the reliability of quantitative ratings

If the data are ratings made on some type of scale, intraclass correlation is a better measure than Kappa.[5] There are actually several variations of the intraclass correlation, but all are set up the same way; they only differ in their treatment of the components of variation. The following methods lend themselves readily to calculation by hand, but all of the terms can be easily computed using analysis of variance set up as a rater-by-category, full-factorial design.

Consider the example in Table 4. Here the judges made independent ratings of the print quality of fabrics, using a 1-to-9 scale in which 1 = poor and 9 = excellent. Once the variance components (actually mean squares) were calculated, the remaining analysis was simple.

There are six different forms of the intraclass correlation. Each form is appropriate for different situations, which will be described in terms of the example in Table 4:

Situation 1: Each printed fabric is rated by a different set of three judges, who are randomly selected from a larger population of judges. Here is how to estimate the reliability of each judge's ratings:

$$ICC = \frac{BMS - WMS}{BMS + (k-1)WMS} = \frac{10.27 - 0.87}{10.27 + 2(0.87)} = 0.78$$

Situation 2: Each printed fabric is rated by a different set of three judges, who are randomly selected from a larger population of judges. Here is how to estimate the reliability of the judges' averaged ratings.

$$ICC = \frac{BMS - WMS}{BMS} = \frac{10.27 - 0.87}{10.27} = 0.92$$

Situation 3: A random sample of three judges is selected from a larger population of judges, and this set of judges rates all five fabrics. Here is how to estimate the reliability of each judge's ratings:

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n}$$

$$= \frac{10.27 - 1.07}{10.27 + 2(1.07) + 3(0.07 - 1.07)/5}$$

$$= 0.78$$

Situation 4: A random sample of three judges is selected from a larger population of judges, and this set of judges rates all five printed fabrics. Here is how to estimate the reliability of the judges' averaged ratings:

$$ICC = \frac{BMS - EMS}{BMS + (JMS - EMS)/n} = \frac{10.27 - 1.07}{10.27 + (0.07 - 1.07)/5} = 0.91$$

Situation 5: All five fabrics are rated by three judges, who are the only judges of interest (i.e., there is not a larger population). Here is how to estimate the reliability of each judge's ratings:

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS} = \frac{10.27 - 1.07}{10.27 + (3-1)1.07} = 0.74$$

Situation 6: All five fabrics are rated by three judges, who are the only judges of interest (i.e., there is not a larger population). Here is how to estimate the reliability of the judges' averaged ratings:

$$ICC = \frac{BMS - EMS}{BMS} = \frac{10.27 - 1.07}{10.27} = 0.9$$

There is no way to obtain a true score for the quality of bourbon or wine, yet trained tasters can rate and classify these with remarkable agreement.

While it is possible to create data sets that produce wildly different values for the intraclass correlation across the six situations, this rarely happens in real measurement situations. The main factor in determining the reliability of ratings is whether the ratings are from a single judge (as in situations 1, 3, and 5) or whether the ratings are averaged across a set of judges (as in situations 2, 4, and 6). Situations 5 and 6 are most common in industry because the raters (or inspectors) are usually dedicated, not selected, from some larger population.

Interpreting intraclass correlation is much like interpreting Kappa: 0.7 should be regarded as a lower bound of acceptability, and anything higher than 0.9 is quite good. If values below 0.7 are obtained, the measurement system needs to be evaluated. Like Kappa, the problems will generally be attributable to poor operational definitions or poorly trained judges.

The case of the hot sauces

Sometimes users have difficulty deciding whether their measurement system requires Kappa or intraclass correlation. The following example might be helpful.

Wilson and Justin are visiting New Orleans and are overwhelmed by the varieties of local hot sauces available. As they taste a few varieties, they notice that they seem to agree about how hot each one is. Since Wilson aspires to become a psychometrician, he designs a study to measure their agreement. They randomly purchase 10 bottles of hot sauce and independently classify them into four categories: mild (M), hot (H), very hot (VH), and makes me suffer (MMS).

Here is how Wilson and Justin rated each sauce:

Sauce	Wilson	Justin
1	M	M
2	M	H
3	MMS	VH
4	VH	MMS
5	H	VH
6	VH	VH
7	H	M
8	H	H
9	MMS	VH
10	M	H

Table 5. Hot Sauce Ratings by Category

		Justin				
		M	H	VH	MMS	Total
Wilson	M	1	2	0	0	3
	H	1	1	1	0	3
	VH	0	0	1	1	2
	MMS	0	0	2	0	2
	Total	2	3	4	1	10

Table 6. Hot Sauce Proportions by Category

		Justin				
		M	H	VH	MMS	Total
Wilson	M	0.1	0.2	0.0	0.0	0.3
	H	0.1	0.1	0.1	0.0	0.3
	VH	0.0	0.0	0.1	0.1	0.2
	MMS	0.0	0.0	0.2	0.0	0.2
	Total	0.2	0.3	0.4	0.1	1.0

If the data are ratings made on some type of scale, intraclass correlation is a better measure than Kappa.

Table 5 contains the data in a tabular format by category, and Table 6 converts these data into proportions. From these tables, Wilson calculated Kappa:

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

$P_{observed} = (0.1 + 0.1 + 0.1 + 0) = 0.3$

$P_{chance} = (0.3 \times 0.2) + (0.3 \times 0.3) + (0.2 \times 0.4) + (0.2 \times 0.1) = 0.25$

$$K = \frac{0.3 - 0.25}{1 - 0.25} = 0.067$$

Wilson was confused by this very low value; it seemed that he and Justin had, for the most part, agreed about the hotness of the sauces, but the statistics showed that their agreement was barely above chance level. Wilson crumpled his calculations, gave up psychometrics, and decided to stay in New Orleans and open a restaurant with Justin.

Fortunately, a psychometrician happened to pass by the picnic table where Wilson had left his crumpled calculations and began to check his work. He observed that Kappa had been calculated perfectly, but that this was not the best statistic for the job. Since the hot sauces can be placed in order of hotness, disagreeing about some categories is not as serious as disagreeing about others. For example, there were no situations where Justin called a sauce "mild" and Wilson called it "makes me suffer"; when they disagreed, they only disagreed by one category.

The psychometrician then recomputed the results using an intraclass correlation. First, he converted the categories into ratings, assigning them as mild = 1, hot = 2, very hot = 3, and makes me suffer = 4. Then he converted the information into the format in Table 7.

As Table 7 shows, the intraclass correlation yields a value of

Table 7. Psychometrician's Intraclass Correlation for the Hot Sauces

	Wilson	Justin	SUM	SUM^2
1	1	1	2	4
2	1	2	3	9
3	4	3	7	49
4	3	4	7	49
5	2	3	5	25
6	3	3	6	36
7	2	1	3	9
8	2	2	4	16
9	4	3	7	49
10	1	2	3	9
SUM	23	24	47	255
SUM^2	529	576	1,105	

Sum of all squared ratings:	131.00
Average of all ratings:	2.35
Sum x average (47 x 2.35):	110.45

Sums of squares:		DF	Mean squares	Component
For raters:	0.05	1	0.05	**JMS**
For between sauces:	17.05	9	1.89	**BMS**
For total:	20.55	19	1.08	
For within sauces:	3.50	10	0.35	**WMS**
For error:	3.45	9	0.38	**EMS**

Situation	ICC
1	0.69
2	0.82
3	0.68
4	0.81
5	0.66
6	0.80

Key:
BMS = Between mean square
DF = Degrees of freedom
EMS = Error mean square
ICC = Intraclass correlation
JMS = Judges' mean square
WMS = Within mean square

0.69 for individual ratings and 0.82 for averaged ratings. While the 0.69 value is not great, it is much better than the Kappa value of 0.067 that Wilson obtained.

The reason why the intraclass correlation is a more appropriate measure is simple: These ratings are quantitative, while the classifications that Wilson used were not. For Wilson, labeling a sample as "hot" when Justin called it "very hot" was just as serious an error as labeling a sample "mild" when Justin called it "makes me suffer." All misclassifications are treated equally. The intraclass method, however, uses the information about the relative hotness of the sauces and is sensitive to how serious a misclassification is. The intraclass correlation is preferable when the data can be ordered and when the perception of the distance between these ordered categories is roughly equal.

References

1. L.B. Barrentine, *Concepts for R&R Studies* (Milwaukee, WI: ASQC Quality Press, 1991).

2. D.R. Peryam, "Quality Control in the Production of Blended Whiskey," *Quality Engineering*, Vol. 5, No. 2, 1992, pp. 347-357.

3. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, 1960, pp. 37-46.

4. H.E.A. Tinsley and D.J. Weiss, "Interrater Reliability and Agreement of Subjective Judgments," *Journal of Counseling Psychology*, Vol. 22, No. 4, 1973, pp. 358-375.

5. M.A. Lahey, R.G. Downey, and F.E. Saal, "Intraclass Correlations: There's More There Than Meets the Eye," *Psychological Bulletin*, Vol. 93, No. 3, 1983, pp. 586-595.

Bibliography

Ebel, R.L., "Estimation of the Reliability of Ratings," *Psychometrika*, Vol. 16, 1951, pp. 407-427.

Shrout, P.E., and J.L. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin*, Vol. 86, No. 2, 1979, pp. 420-428.

Wheeler, D.J., and R.W. Lyday, *Evaluating the Measurement Process* (Knoxville, TN: Statistical Process Controls, Inc., 1984).

David Futrell is a consultant at QualPro in Knoxville, TN. He received a doctorate in industrial/organizational psychology from the University of Tennessee at Knoxville. Futrell is a member of ASQC.