British Journal of Mathematical and Statistical Psychology (2010), 63, 341–360

The British Psychological Society

www.bpsjournals.co.uk

The a_d coefficient as a descriptive measure of the within-group agreement of ratings

Ludwig Kreuzpointner[1]*, Patricia Simon[1] and Fabian J. Theis[2]

[1]Universität Regensburg, Germany

[2]Helmholtz Zentrum München, Germany

The a_d coefficient was developed to measure the within-group agreement of ratings. The underlying theory as well as the construction of the coefficient are explained. The a_d coefficient ranges from 0 to 1, regardless of the number of scale points, raters, or items. With some limitations the measure of the within-group agreement of different groups and groups from different studies is directly comparable. For statistical significance testing, the binomial distribution is introduced as a model of the ratings' random distribution given the true score of a group construct. This method enables a decision about essential agreement and not only about a significant difference from 0 or a chosen critical value. The a_d coefficient identifies a single true score within a group. It is not provided for multiple true score settings. The comparison of the a_d coefficient with other agreement indices shows that the new coefficient is in line with their outcomes, but does not result in infinite or inappropriate values.

1. Introduction

In recent years, multi-level theory has been increasingly used for the analysis of organizational processes since organizations are hierarchically structured systems with different levels of analysis. Due to the nested structure of organizations, it is assumed that the ongoing processes on one level influence the processes on another level. If this aspect is not considered, statistical artefacts and contradictory results may easily be obtained (Bliese, 2000). In multi-level theory, different models have been proposed for the analysis of the relations within and across levels (Chan, 1998; Klein & Kozlowski, 2000; Kozlowski & Klein, 2000; Moritz & Watson, 1998). If the analysis is based on a so-called bottom-up composition model, it is required to test the within-group agreement of the individual ratings when aggregating the data at the group level. Given a high within-group agreement, the mean of the individual data can be used as a measure of the group-level construct such as group cohesiveness, group norms, or climate

*Correspondence should be addressed to Ludwig Kreuzpointner, Institut für Psychologie, Universität Regensburg, 93040 Regensburg, Germany (e-mail: ludwig.kreuzpointner@psychologie.uni-regensburg.de).

DOI:10.1348/000711009X465647

(Cohen, Doveh, & Eick, 2001; Klein & Kozlowski, 2000; Kozlowski & Hattrup, 1992; Lance, Butts, & Michels, 2006; Lindell & Brandt, 1997; Moritz & Watson, 1998). Lüdtke and Robitzsch (2009) list some further research areas using within-group agreement: aggregating group perceptions, for example fairness, in industrial and organizational psychology (Masterson, 2001); measuring the characteristics of a group by asking its members, in small-group research; and pooling students' answers to a class-level measurement, in educational research.

For the estimation of the agreement among those ratings the r_{WG} coefficient is often proposed (Bliese, 2000; Chan, 1998; Klein & Kozlowski, 2000; Moritz & Watson, 1998), which was originally developed by Finn (1970) and further studied by James, Demaree, and Wolf (1984, 1993). The within-group agreement coefficient r_{WG} is defined in the case of a single item as $r_{WG(1)} = 1 - s^2_{x_j}/\sigma^2_E$, where $s^2_{x_j}$ corresponds to the sample variance in the ratings of the K judges on a single item X_j. σ^2_E refers to the expected variance (E) due to random response mostly based on a uniform distribution (U) (Lüdtke & Robitzsch, 2009). The expected variance in the uniform case is determined by the equation $\sigma^2_E = (A^2 - 1)/12$, where A corresponds to the number of scale points, which means that on a five-point scale $A = 5$. The term $s^2_{x_j}/\sigma^2_E$ reveals the proportion of error variance in the ratings and so $1 - s^2_{x_j}/\sigma^2_E$ represents the proportion of non-error variance (Finn, 1970; James *et al.*, 1984). The r_{WG} coefficient represents an attempt to remove the variance expected by chance from the observed variance. One less the proportion of variance in the ratings corrected for chance corresponds to the degree of agreement in the ratings. The coefficient may be extended to multiple items, referred to as $r_{WG(J)}$, based on the same principles. A weakness of the $r_{WG(1)}$ coefficient is that it becomes negative when $s^2_{x_j} > \sigma^2_E$ $\left(r_{WG(J)}\right.$ becomes negative if $\left.\sigma^2_E < s^2_{x_j} < \frac{J}{J-1}\sigma^2_E\right)$. James *et al.* (1984) proposed setting negative r_{WG} to 0. According to Lüdtke and Robitzsch (2009, p. 3) this method could cover 'the fact that a target has multiple true scores'. Cohen *et al.* (2001) analysed the statistical properties of r_{WG}. They pointed out the problems by testing the statistical significance when setting negative values to 0 with the consequence of 'a large proportion (.90) of zero values' (p. 301) in the one-item case.

Lindell and Brandt (1997) proposed replacing the uniform distribution with maximum dissensus as reference distribution so that the coefficient lies within the proper interval [0,1], referred to as r_{WG_MV}. There are several parallels between Lindell and Brandt's work and the a_d coefficient, presented in the present paper. Subtle distinctions result from using distances instead of the variance. Using the sample variance (with denominator $n - 1$) may again lead in extreme cases to negative r_{WG_MV}. The fact that the formula for the maximum variance does not exactly fit with the formula for sample variance is due to the small differences in the values of a_d and r_{WG_MV}.

As an alternative, Schmidt and Hunter (1989) proposed estimating the within-group agreement with the standard deviation of ratings or the standard error of the mean. But without a fixed reference point in these procedures, they do not satisfy the aim of obtaining a measure delivering information about the extent of agreement. The coefficient a_{WG} proposed by Brown and Hauenstein (2005) uses the maximum variance given the mean of the ratings. The values of a_{WG} lie between -1 (maximum disagreement) and $+1$ (maximum agreement). There is only a problem modelling agreement when many ratings have the maximum or minimum value on the scale.

2. The a_d coefficient for the estimation of the within-group inter-rater agreement

To simplify matters, our theoretical remarks are explicated for the case of three raters rating one object on a single item. These remarks can be readily extended to any number of raters and items. The question is how similar ratings can be in order to conclude that they agree to a substantial extent. Therefore, the degree of inter-rater agreement is a function of the deviations in the judgements. Based on the assumptions of classical test theory the observed score x can be decomposed into a true score t and an error score e; the following three equations are obtained for the raters' judgements: for rater 1, $x = t + e_x$; for rater 2, $y = t + e_y$; for rater 3, $z = t + e_z$. If the degree of inter-rater agreement is understood as a function of deviations in the judgements, the observed scores of the raters can be subtracted from one another:

$$\text{Rater 1} - \text{Rater 2}: \quad x - y = (t + e_x) - (t + e_y),$$

$$\text{Rater 1} - \text{Rater 3}: \quad x - z = (t + e_x) - (t + e_z),$$

$$\text{Rater 2} - \text{Rater 3}: \quad y - z = (t + e_y) - (t + e_z).$$

Since the same object underlies each rater's judgment and therefore the same true score t, the obtained deviations between the x-, y-, and z-scores are only caused by the error score e, referred to as distance $d_{k,l}$:

$$\text{Rater 1} - \text{Rater 2}: \quad x - y = e_x - e_y = d_{1,2},$$

$$\text{Rater 1} - \text{Rater 3}: \quad x - z = e_x - e_z = d_{1,3},$$

$$\text{Rater 2} - \text{Rater 3}: \quad y - z = e_y - e_z = d_{2,3}.$$

The distances $d_{k,l}$ between the observed scores reveal the degree of disagreement in the ratings. The greater the deviations are, the lower the raters' agreement is; as a consequence, the ratings' objectivity declines (Simon & Kreuzpointner, 2008). If one were to add up only these deviations, the positive and negative deviations could yield a value of zero. To prevent this and to be able to apply the coefficient in the case of multiple items, the squared Euclidean distances $d^2_{k,l}$ are used. As greater deviations carry more weight, a stricter criterion of objectivity results through squaring. This is useful because great deviations are a sign of low agreement. By summing up the individual squared Euclidean distances $d^2_{k,l}$, a statistical value results for the extent of disagreement among the raters, referred to as d^2:

$$d^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{k'=k}^{K} (x_{kj} - x_{k'j})^2, \tag{1}$$

where K is the number of raters and J the number of items. However, without a reference point the statistical value of d^2 cannot be interpreted as being a large or a small disagreement. The fact that rating scales are bounded now becomes an advantage. Due to the restriction of a scale, the maximum possible disagreement among the raters is known. Maximum disagreement results from the difference between the maximum value b of a scale and the minimum value a of a scale which is squared once again: $d^2_{\max} = (b - a)^2$. Given more than two raters or more than one item, the calculation of the maximum possible disagreement on a scale must be adjusted.

For the adjustment, the following two equations result depending on whether the number of raters K is even or odd:

$$d^2_{\max} = J \cdot (b - a)^2 \cdot (K^2/4), \quad \text{if } K \text{ is even}, \tag{2}$$

$$d^2_{\max} = J \cdot (b - a)^2 \cdot ((K^2 - 1)/4), \quad \text{if } K \text{ is odd}. \tag{3}$$

The mathematical proof for Equations (2) and (3) is described in Appendix A1. With these two statistical values, the sum of the squared Euclidean distances between the raters d^2 and the maximum disagreement $d^2_{\max}$, a coefficient for the measurement of the extent of inter-rater agreement can be constructed. Dividing d^2 by $d^2_{\max}$ delivers information about the extent of disagreement in the ratings. Hence, 1 as the value of perfect agreement minus the extent of disagreement delivers information about the extent of inter-rater or within-group agreement:

$$a_d := 1 - d^2/d^2_{\max}. \tag{4}$$

According to its function as a descriptive measure of the agreement among ratings, this coefficient is called the agreement coefficient. The subscript d symbolizes that the coefficient is calculated from distances. The a_d coefficient is an extension of the $\ddot{U}$ coefficient suggested by Fricke (1972) to measure the objectivity within mastery testing. Fricke derived and used his index only for dichotomous data. Moreover, we have derived precise results for $d^2_{\max}$ for odd and even numbers of raters, not taken into account by Fricke.

3. Features of the a_d coefficient

The a_d coefficient ranges from 0 to 1. This is because the ratio of $d^2/d^2_{\max}$ can only vary within the interval [0,1] since d^2 as a deviation measure cannot be negative and is limited by the maximum value $d^2_{\max}$ (see Appendix A1). Given maximum agreement, $a_d = 1$ since in this case $d^2 = 0$. Given maximum disagreement, $d^2 = d^2_{\max}$ and therefore $a_d = 1 - d^2_{\max}/d^2_{\max} = 1 - 1 = 0$. Hence, the range of a_d is limited between 0 and 1 regardless of the number of scale points, raters, or items (see Appendix A3).

A further feature of the a_d coefficient is that the scores are interval-scaled. Given equal numbers of raters, items, and scale points, the difference of .05 between two a_d values of .80 and .85 reflects the same difference in agreement as the difference between two a_d values of .90 and .95. This is true because the a_d coefficient is based on distances, which are by definition at least interval-scaled (see Appendix A4). Additionally, under the condition of the same number of items and raters, two a_d coefficients based on different numbers of scale points can also be compared. The effect of the scale points on the magnitude of the coefficient is eliminated by the ratio of $d^2/d^2_{\max}$ since the sum of the squared Euclidean distances d^2 as well as the maximum possible distances $d^2_{\max}$ depend in the same manner on the number of scale points (see Appendix A2). However, this is only true without restrictions when continuous scales (such as visual or other analogue scales) are used. Given discrete scales, an identical transformation of different scales is not always possible. In this case, whole numbers only result when the upper bound of the broader scale represents an integer multiple of the upper bound of the narrower scale. Therefore, in the case of discrete

scales, one can only speak of an approximate independence of the number of scale points (compare the critical values in Tables B1 and B2 or B3 and B4).

It is important to mention that the a_d coefficient can also be expressed in terms of the sum of squares (SS) since the squared Euclidean distances divided by the number of raters is equal to the SS. Accordingly, d^2 could be substituted by the SS between the ratings and $d^2_{\max}$ by the maximum SS. In cases of more than one item, d^2 is equal to the sum of SS for each item and $d^2_{\max}$ corresponds to J times the maximum SS of one item. Both procedures lead to the same values. However, squared Euclidean distances were deliberately chosen for three reasons. First, the Euclidean distances better suit the theoretical derivation of the coefficient from the concept of objectivity. Second, the automatic calculation of $d^2_{\max}$ is slightly easier to implement. Third, the description in terms of SS spuriously implies a chi-square test. This test requires a normal distribution of the ratings which is not given (cf. Dunlap, Burke, & Smith-Crowe, 2003). To avoid this fallacy, the a_d coefficient builds on Euclidean distances which are, from our point of view, clearer and intuitively more comprehensible, even if the SS is much more familiar.

One issue which initially seems questionable is the fact that with the a_d coefficient the extent of disagreement is estimated on the basis of maximum disagreement. Maximum disagreement implies that multiple true scores can be given within a group. But for the calculation and application of the a_d coefficient it is assumed that every group has only one single true score (like Brown and Hauenstein's (2005) a_{WG} coefficient). In contrast to a_{WG}, a_d does not include the estimation of the true score value within the construction of the coefficient, but takes it into account while testing the statistical significance (see Section 4). So the case of two (or more) true scores is not a technical issue but rather a theoretical one. The solution lies in the stated aim of measuring the amount of agreement of one group. When there are two true scores in a group, which will lead to a low to very low value of a_d, we suggest interpreting this setting as two subgroups within a group when using the a_d coefficient for calculating the within-group agreement. A low value of a_d implies no significant agreement within the group. When there are two true scores within a group there is no agreement of all members. A low value of a_d could only serve as an indication of more than one true score within the group. But it is not (yet) intended to identify them.

4. Testing the statistical significance of the a_d coefficient

The first assumption is that, according to composition models, a group-level construct only exists when there is a perceptual consensus in the members' ratings. Based on this, the mean of the ratings is used as an estimate of the true score of a group construct. A violation of this assumption would lead to a rejection of the hypothesis of within-group agreement, as we will see later. The binomial distribution as a model for the distribution of the ratings on a scale provides a basis for the determination of statistical significance. The binomial distribution is constructed in such a manner that the ratings around a group's estimated true score are more likely than those farther away from the true score. For this to be realized, the individual judgements are treated as the sum of how often each binary variable occurs in n trials. In other words, each judgement is interpreted as the sum over n random experiments of a binary random variable. The number of trials n corresponds to the decremented number of scale points ($A - 1$),

since the scale point one represents zero occurrences in the binomial distribution. The probability p of a hit determines the expected value of the binomial distribution, which corresponds to the also decremented estimated true value of the group construct (see Figure 1). The probability p can easily be calculated as $p = (\bar{X} - 1)/(A - 1)$, since the expected value of a binomial distribution is $E(X) = pn$, with $n = \mathrm{A} - 1$ and $E(X) = \bar{X} - 1$. Given a set of samples, p is therefore directly given by the mean, and this is indeed also the best possible estimate of p in the maximum-likelihood sense. Thus, the larger the mean of the ratings of one group, the larger is the hit probability p of the binomial distribution used to model the distribution of the ratings for this group.

Figure 1 illustrates the procedure. An estimated true score of $t = 2.8$ corresponds to an average hit rate of $E(X) = 2.8 - 1 = 1.8$. Using a seven-point scale, this condition results when the hit probability is $p = 1.8/6 = .3$. The range of the scale and the hit probability together define the binomial distribution $B(.3, 6)$, which determines the probability of the seven possible results or ratings (see Figure 1a). Figure 1 presents examples of histograms of an infinite number of Bernoulli experiments with specified conditions. Thus, the binomial distribution provides the likelihood of the ratings under the condition of the true score which is necessary for the calculation of the null distribution of the a_d coefficient.

By drawing 10,000 samples from such a binomial distribution $B(p, n)$, the null distribution of a_d can be determined using the Monte Carlo technique. This distribution is the basis for testing H_0, the hypothesis that the ratings agree only by chance given the mean rating as an estimate of a group's true score. The following example illustrates the method: if within a group of seven members with ratings {2, 3, 3, 3, 3, 4, 4} the mean of $\bar{X} = 3.14$ is observed on a seven-point scale, for instance, measuring the construct of group cohesion, then this value is used for estimating a group's true cohesion score. The mean implies $p = .4$ since $p = (3.14 - 1)/(7 - 1) = .36 \approx .4$. Using random sampling based on the binomial distribution $B(.4, 6)$, the null distribution of a_d can be determined. The critical value of a_d with an alpha level of 5% corresponds to the 95% quantile of the resulting null distribution, which in this case equals .95 (see Table B3 in Appendix B). As $a_d = .954$ is higher than this critical value, the a_d coefficient obtained is statistically significant. Of course, it is possible to calculate the distribution and the critical value of a_d more exactly by using $p = .36$.

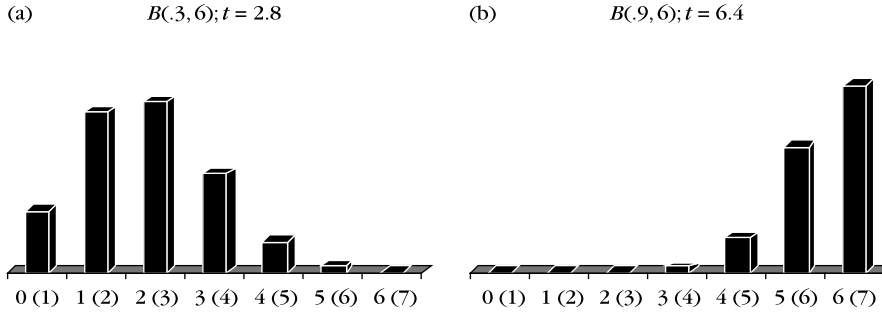

Figure 1. The binomial distribution as a model for the distribution of the ratings of a group on a scale (two examples). *Note*. The numbers without brackets correspond to the binomial distribution. The numbers in brackets correspond to the modelled rating scale. The ordinate represents the probability of the results under the given conditions.

In Appendix B, the critical values of the a_d coefficient are listed for the often used five-point and seven-point scales. The number of group members (raters) ranges from $k = 3$ to 12 since a small group is often understood as being composed of 3 to 12 persons (Wiendieck, 1994). The number of items ranges from 1 to 10. The critical values correspond to an experimental accuracy of two decimal digits. For investigations with more items, larger scales, or exact calculations for a specific setting a program to calculate the critical values is available from the authors (http://www-cgi.uni-regensburg.de/~krl02854/a-coef/). Based on these tables, the statistical significance of a given a_d value can easily be determined. First, it is necessary to calculate the mean of the ratings, which is used to estimate the true value t and then the underlying p. Afterwards, the desired significance level for testing a_d must be selected. Tables B1 and B3 provide the critical values for the significance test with an alpha level of 5%; Tables B2 and B4 provide those with one of 1%. The next step consists of identifying the scale used, the number of raters k, and the number of items j. In the case of six raters on a five-point scale, the critical value with a level of 5% is .94 for a single item where $p = .7$, and .92 for five items where $p = .2$. Moreover, the tables in Appendix B also list the critical values of the a_d coefficient under the assumption of a (discrete) uniform distribution of ratings (see column U). This enables wider application of the a_d coefficient beyond multi-level theory if it is assumed that the ratings are not similar.

With an underlying binomial distribution and the expectation that the data agree considerably more than this, the significance test of the a_d coefficient provides a clear criterion for the question of how high the agreement has to be in order to treat it as a substantial within-group agreement. In the construction of the significance test for the a_d coefficient we took Fowler's (1985) proposal into account. The null hypothesis was specified insofar as it corresponded to the distribution of the ratings expected by chance under the assumption of a group's true score. This specification ensures that statistically significant a_d values are also practically significant. We know that this criterion is a strict one, but we think it is more appropriate than testing non-zero associations. Cohen *et al.* (2001) compare this problem with the significance test of correlations which only indicate that a statistically significant non-zero association is obtained which is not necessarily practically significant.

Independent of the clear interpretation according to the statistical significance test, the a_d values are more difficult to interpret in comparison, for instance, with the product moment correlation coefficient. At first sight, with only a low agreement, the values of the a_d coefficient are normally relatively high. This is due to the fact that the sum of the distances in the ratings d^2 merely corresponds to the maximum possible disagreement $d^2_{\max}$. Therefore often values in the upper range of the scale from 0 to 1 will be obtained, which seem to be more difficult to interpret. For this reason, Burke, Finkelstein, and Dusig (1999) criticized the usage of maximum dissensus as reference point. On the other hand, it has to be considered that it is only an effect of usage and the statistical significance test of the a_d coefficient prevents misinterpretation.

5. Examples of the calculation of the a_d coefficient

5.1. One item

Six judgements {5, 5, 4, 4, 3, 2} on a five-point scale were obtained for the measurement of group cohesion. First, the sum of the squared Euclidean distances

d^2 has to be calculated using equation (1):

$$d^2 = (5-5)^2 + (5-4)^2 + (5-4)^2 + (5-3)^2 + (5-2)^2 + (5-4)^2 + (5-4)^2$$
$$+ (5-4)^2 + (5-3)^2 + (5-2)^2 + \cdots + (3-2)^2 = 41.$$

Second, $d^2_{\max}$ must be calculated. Since the number of judges K is even, equation (2) must be used:

$$d^2_{\max} = J \cdot (b-a)^2 \cdot \frac{1}{4} K^2 = 1 \cdot (5-1)^2 \cdot \frac{1}{4} 6^2 = 144.$$

Based on the two values of d^2 and $d^2_{\max}$, the agreement coefficient a_d can be calculated using equation (4):

$$a_d = 1 - d^2/d^2_{\max} = 1 - 41/144 = .72.$$

With $K = 6$ judges on a five-point scale and a p of .7 for the binomial distribution, the critical value of a_d is .94 with an alpha level of 5% (see Table B1). Since the a_d coefficient obtained is less than the critical value, the null hypothesis cannot be rejected. This means that the members do not show substantial agreement in their cohesion ratings.

5.2. Multiple items

Suppose we have a data matrix of three judges rating on five items for the measurement of group climate with seven alternatives (see Table 1). Then d^2 is given by the sum of the squared Euclidean distances of the three answer vectors of the judges according to equation (1):

Table 1. Data matrix of three judges rating on five items each with seven alternatives

Judges	Item 1	Item 2	Item 3	Item 4	Item 5	*M*
A	1	2	2	2	1	1.6
B	2	2	1	2	2	1.8
C	2	3	2	3	2	2.4

$$d^2 = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 = \underbrace{\left[(1-2)^2 + (1-2)^2 + (2-2)^2\right]}_{=d_1^2}$$
$$+ \underbrace{\left[(2-2)^2 + (2-3)^2 + (2-3)^2\right]}_{=d_2^2} + \cdots + \underbrace{\left[(1-2)^2 + (1-2)^2 + (2-2)^2\right]}_{=d_5^2} = 10.$$

Since the number of judges K is odd, equation (3) is used for the calculation of the maximum squared Euclidean distances $d^2_{\max}$:

$$d^2_{\max} = J \cdot (b-a)^2 \cdot \frac{1}{4}(K^2 - 1) = 5 \cdot (7-1)^2 \cdot \frac{1}{4}(3^2 - 1) = 5 \cdot 36 \cdot \frac{8}{4} = 360.$$

In this case, $a_d = 1 - 10/360 = .972$. The p value of the binomial distribution for the statistical significance test is .2 since $p = (1.93 - 1)/(7 - 1) = .155$. With $K = 3$ judges on a seven-point scale and $J = 5$, a critical value of .97 results with an alpha level

of 5% (see Table B3). Therefore, there is a statistically significant within-group agreement. For an easy calculation of the a_d coefficient, a software program has been written which is available from the authors (http://www-cgi.uni-regensburg.de/~krl02854/a-coef/). The coefficient can also be calculated with SPSS which provides a calculation of the squared Euclidean distances in the program of hierarchical cluster analysis. d^2 is equal to the sum of one symmetrical half of the Euclidean distance matrix provided by SPSS. $d^2_{\max}$ can then easily be calculated by hand.

6. A comparison of the a_d coefficient with other agreement indexes

Since there is no objective criterion for the evaluation of the validity of an agreement index, the comparison can only be conducted with regard to considerations of plausibility. The a_d coefficient is compared to the r_{WG} coefficient and its alternatives discussed in the literature. The alternatives proposed by Schmidt and Hunter (1989) as well as the average deviation index proposed by Burke *et al.* (1999) are not included, because these coefficients are not interpretable with regard to the extent of agreement due to a missing fixed reference point. Three arbitrary extreme examples were chosen as comparison material showing at best the features of the different coefficients. Table 2 presents the raw data of these examples which are based on a five-point scale. The coefficients are presented for each item separately and for all items of an example overall.

First, an example is presented in which all group members agree in their ratings. Since the ratings correspond to the upper bound of the five-point scale, a_{WG} is not defined, while the other coefficients deliver the value of 1 for perfect agreement. In the second item of this example, the fourth rater does not agree with the others and chooses a value of 4. While the values of all coefficients are more or less reasonable, the value of the a_{WG} coefficient of .60 seems odd as there is only one disagreement. According to the test statistic developed by Dunlap *et al.* (2003), r_{WG} indicates that there is no statistically significant agreement as well as the a_d coefficient (see Appendix B1, $K = 4, J = 1, p = .9$) due to the low number of raters. A comparison of the raw data of the second and third items shows that there is lower agreement with regard to the third one while a_{WG} delivers a higher value for this item, reflecting its problems with sets of data including values on the limit of the scale. While in this case a_d with respect to all items yields almost the same value as $r_{\mathrm{WG}(J)}$, this example shows a special characteristic of the former which is due to its feature as interval scale: The a_d coefficient for all items equals the mean of a_d for every single item, which is true in general (see Appendix A5).

The first item of the second example represents a case of maximum possible disagreement, indicated by the a_d coefficient with a value of zero and by the a_{WG} coefficient with a value of -1. r_{WG} and $r_{\mathrm{WG\_MV}}$ exceeds the supposed lower bound of 0. The second item of this example makes clear that the a_d coefficient also delivers interpretable values with regard to the extent of disagreement which is slightly lower in this case. Such differences are not detectable with the r_{WG} coefficient as it is recommended to replace negative values with zero. With regard to both items in this example, the coefficients belonging to the r_{WG} family deliver aberrant scores. How should a $r_{\mathrm{WG}(J)}$ value of 13.2 be interpreted? It seems to be an odd range of values which should all be set to zero. Lindell, Brandt, and Whitney (1999) claimed that $r^*_{\mathrm{WG}(J)}$ would range between $[-1, 1]$ in the case of a five-point scale but in this – admittedly extreme – example the score exceeds the lower bound. Similarly, $r_{\mathrm{WG\_MV}}$ does not show the purported features.

Table 2. A comparison of the a_d coefficient with other agreement indexes

Item	R1	R2	R3	R4	M	d^2	d^2_{max}	s^2	a_d	r_{WG}	r_{WG_MV}	a_{WG}
1	5	5	5	5	5.00	0	64	0.00	1.00	1.00	1.00	n.d.
2	5	5	5	4	4.75	3	64	0.25	0.95	0.88	.94	0.60
3	5	5	4	4	4.50	4	64	0.33	0.94	0.83	.92	0.71

Overall coefficients: a_d=.96, $r_{WG(J)}$=.97, $r_{WG\_MV(J)}$=.98, $r^*_{WG(J)} = 90$, $a_{WG(J)}$=n.d.

Item	R1	R2	R3	R4	R5	R6	R7	M	d^2	d^2_{max}	s^2	a_d	r_{WG}	r_{WG_MV}	a_{WG}
1	1	1	1	1	5	5	5	2.71	192	192	4.57	0.00	−1.29	−0.14	−1.00
2	1	1	1	4	5	5	5	3.14	174	192	4.14	0.09	−1.07	−0.04	−0.78

Overall coefficients: $a_d = .05$, $r_{WG(J)} = 13.2$, $r_{WG\_MV(J)} = -.20$, $r^*_{WG(J)} = -1.18$, $a_{WG(J)} = -.89$

Item	R1	R2	R3	R4	R5	R6	R7	R8	M	d^2	d^2_{max}	s^2	a_d	r_{WG}	r_{WG_MV}	a_{WG}
1	4	4	5	3	4	3	3	5	3.88	39	256	0.70	0.85	0.65	0.83	0.62
2	1	1	4	5	5	2	5	5	3.50	192	256	3.43	0.25	−0.71	0.14	−0.60
3	1	1	1	1	1	5	5	5	2.50	240	256	4.29	0.06	−1.14	−0.07	−1.00

Overall coefficients: a_d=.39, $r_{WG(J)}$=−6.14, $r_{WG\_MV(J)}$=.56, $r^*_{WG(J)}$=−.40, $a_{WG(J)}$=−.33

Note. R, rater; n.d., not defined; $r^*_{WG(1)} = r_{WG}$.

For the first item in the third example the raw data indicate low agreement; a_d yields a relatively high value of .85, but the test shows that this value is not statistically significant. On the other hand, the low r_{WG} value of .65 achieves statistical significance. However, this value is below the theoretical discussed cut-off value for practical significance of .70 (Dunlap *et al.*, 2003; Lance *et al.*, 2006), showing that the statistical significance test for the r_{WG} coefficient provides no clear decision rule. Similarly to the examples before, r_{WG} yields negative values for items 2 and 3 which leads to the large negative value of -6.14 for all three items. On the other hand, the a_d coefficient for all items of .39 takes into account that the raters slightly agreed with respect to item 1.

7. Discussion

In composition models, a perceptual convergence among group members is required in order to aggregate the data at the group level. Therefore, an index is needed to measure the agreement among ratings. With the a_d coefficient a new index of within-group agreement is available which addresses some of the deficiencies of other indices. This coefficient fulfils the requirement of obtaining a comparable measure for the agreement of different groups and groups from different studies. It ranges from 0 to 1, regardless of the number of scale points, items, or raters. Its significance test provides an unequivocal criterion to answer the question of how high the within-group agreement must be in order to speak of a construct at the group level. This is due to the underlying binomial distribution, which represents the distribution of the members' ratings around the estimated true score. This test ensures that only values with practical significance yield statistical significance. If there is more than one true score within a group, the result of a_d is expected to be so low that no agreement could be assumed and so the hypothesis of within-group agreement has to be rejected.

The tables in Appendix B show that the critical values are relatively high. This means that only when slight deviations in the ratings are obtained a group construct does actually exist. This is justified by the assumption that the mean of the ratings corresponds to a group's construct true score, even more so as the mean is used for further analysis in the context of composition models. Even though the coefficient was originally developed for the validation of group-level constructs in multi-level theory, it can also be applied in other contexts where the question of the extent of agreement on ratings arises. If it is not expected that the individual ratings are similar, the uniform distribution can also be used for a statistical significance test. Although the critical values cannot be determined by an analytical approach, the Monte Carlo technique delivers satisfying adaptations. The comparison of the a_d coefficient with other agreement indices showed that it provides the best interpretable scores.

Acknowledgements

The authors thank the anonymous reviewers for their detailed comments, which helped improve the paper considerably.

References

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability. Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey Bass.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the r_{wg} indices. *Organizational Research Methods*, *8*(2), 165–184.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*(1), 49–68.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234–246.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, *6*(3), 297–310.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate testing of statistical significance for r_{WG} and average deviation interrater agreement index. *Journal of Applied Psychology*, *88*(2), 356–362.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, *30*, 71–76.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology*, *70*(1), 215–218.

Fricke, R. (1972). Testgütekriterien bei lehrzielorientierten Tests [Quality criteria for mastery testing]. *Zeitschrift für Erziehungswissenschaftliche Forschung*, *6*, 150–175.

James, R. L., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85–98.

James, R. L., Demaree, R. G., & Wolf, G. (1993). r_{WG}: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*(2), 306–309.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*(3), 211–236.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*(2), 161–167.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations. Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco: Jossey Bass.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement*, *21*(3), 271–278.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, *23*(2), 127–135.

Lüdtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods*, *12*(3), 461–487.

Masterson, S. (2001). A trickle-down model of organizational justice: Relating employees' and customers' perceptions of and reactions to fairness. *Journal of Applied Psychology*, *86*(4), 594–604.

Moritz, S. E., & Watson, C. B. (1998). Levels of analysis issues in group psychology: Using efficacy as an example of a multilevel model. *Group Dynamics: Theory, Research and Practice*, *2*(4), 285–298.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, *74*(2), 368–370.

Simon, P., & Kreuzpointner, L. (2008). Die Verwässerung des Reliabilitätskonzepts der klassischen Testtheorie im Falle von Ratingskalen [The confusion of the test theoretical criteria of objectivity, reliability, and validity in the case of rating data]. In W. Sarges & D. Scheffer (Eds.), *Innovative Ansätze für die Eignungsdiagnostik* (pp. 321–331). Göttingen: Hogrefe. (For an English-language manuscript submitted for publication, see http://www-cgi.uni-regensburg.de/~krl02854/confusion.pdf).

Wiendieck, G. (1994). *Arbeits- und Organisationspsychologie* [*Work and organisational psychology*]. Berlin: Quintessenz.

Received 25 May 2007; revised version received 8 May 2009

Appendix A: Proofs

A1. Calculation of d^2_{max}

By definition,

$$d^2 = d^2(X) = \sum_{k=1}^{K}\sum_{k'=k}^{K}\sum_{j=1}^{J}(x_{kj} - x_{k'j})^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{k'=k}^{K}(x_{kj} - x_{k'j})^2 = \sum_{j=1}^{J} s_j,$$

with the squared sum over the columns

$$s_j = \sum_{k=1}^{K}\sum_{k'=k}^{K}(x_{kj} - x_{k'j})^2$$

and the observation matrix $X = (x_{kj})_{k=1,j=1}^{K,J}$. Here J denotes the number of items and K the number of raters. Assume that the values of X are from the interval $[a, b]$. In the following, we will assume continuous variables; however, it is easy to see that the claims also hold for discrete variables.

First we show that for even K,

$$d^2_{\max} = J(b-a)^2\frac{K^2}{4},$$

where $d^2_{\max} = \max_X d^2(X)$ for X with values from $[a, b]$. For this, note that d^2 is maximal if and only if all s_j are maximal. Hence, we only have to maximize s_j. But this is maximal if and only if X has $K/2$ times item a and $K/2$ times item b in each column. Since the problem is symmetric, we may assume that $s_j = a$ for $j \leq K/2$ and $s_j = b$ for $j > K/2$. Then

$$s_j = \sum_{k=1}^{K}\sum_{k'=k}^{K}(x_{kj} - x_{k'j})^2 = \sum_{k=k/2}^{K}\sum_{k'=1}^{k/2}(x_{kj} - x_{k'j})^2 = \sum_{k=k/2}^{K}\sum_{k'=1}^{k/2}(b-a)^2 = (b-a)^2\frac{K^2}{4},$$

which establishes the claim.

Similarly, if K is odd, s_j is maximal if we have $(K-1)/2$ times item a and $(K+1)/2$ times item b or vice versa in each column of X. Again, we may assume that $s_j = a$ for $j \leq (K-1)/2$ and $s_j = b$ for $j > (K-1)/2$. This guarantees that

$$\begin{aligned} s_j = \sum_{k=1}^{K}\sum_{k'=k}^{K}(x_{kj} - x_{k'j})^2 &= \sum_{k=(K+1)/2}^{K}\sum_{k'=1}^{(k+1)/2}(x_{kj} - x_{k'j})^2 = \sum_{k=(K+1)/2}^{K}\sum_{k'=1}^{(k+1)/2}(b-a)^2 \\ &= (b-a)^2\frac{(K-1)(K+1)}{4} = (b-a)^2\frac{K^2-1}{4}, \end{aligned}$$

as claimed.

A2. Invariance under scaling

$a_d = a(X)$ is independent of the interval boundaries $a < b$, i.e. invariant under scaling. For this, assume another interval $a' < b'$ is given, and let $X' := (((b' - a')/(b - a)) \times (x_{jk} - a) + a')_{ij}$ be the variable X rescaled onto $[a', b']$. Then $a(X) = a(X')$ as claimed, because

$$d^2(X') = \sum_{k=1}^{K} \sum_{k'=k}^{K} \sum_{j=1}^{J} \left(\frac{b' - a'}{b - a}(x_{kj} - a) + a' - \frac{b' - a'}{b - a}(x_{k'j} - a) + a' \right)^2$$

$$= \left(\frac{b' - a'}{b - a} \right)^2 \sum_{k=1}^{K} \sum_{k'=k}^{K} \sum_{j=1}^{J} (x_{kj} - x_{k'j})^2 = \left(\frac{b' - a'}{b - a} \right)^2 d^2(X)$$

and, by definition,

$$d^2_{\max}(X') = \left(\frac{b' - a'}{b - a} \right)^2 d^2_{\max}(X).$$

A3. Scaling with respect to item count **J** *and number of raters* **K**

$1 - a(X)$ is J-normal-interval-scaled, i.e. if X and X' denote two observation matrices of sizes $K \times J$ and $K \times J'$, respectively with values from $[a, b]$, and if (X, X') denotes the $K \times (J + J')$ matrix generated by juxtaposition of X and X', then $(J + J') \times [1 - a(X, X')] = J[1 - a(X)] + J'[1 - a(X')]$ or $(J + J')a(X, X') = Ja(X) + J'a(X')$. In other words, this formula illustrates how to calculate the a_d coefficient of an experiment from the a_d coefficients of two subexperiments in terms of item number J. This follows directly from

$$1 - a(X, X') = \frac{d^2(X, X')}{d^2_{\max}(X, X')} = \frac{\sum_{j=1}^{J+J'} s_j(X, X')}{d^2_{\max}(X, X')} = \frac{\sum_{j=1}^{J} s_j(X)}{\frac{J+J'}{J} d^2_{\max}(X)} + \frac{\sum_{j=1}^{J'} s_j(X')}{\frac{J+J'}{J'} d^2_{\max}(X')}$$

$$= \frac{1}{J + J'} \{ J[1 - a(X)] + J'[1 - a(X')] \}.$$

In the special case of $X = X'$, the above formula implies that

$$2J[1 - a(X, X)] = J[1 - a(X)] + J'[1 - a(X)],$$

so $a(X, X) = a(X)$, which means that the a_d coefficient is invariant under 'doubling' the experiment.

The same holds with respect to the number of raters K. Let us assume even K for simplicity, and let $\begin{pmatrix} X \\ X \end{pmatrix}$ denote the matrix generated from X by doubling the number of raters. This $(K + K) \times J$ matrix has the same a_d coefficient as X, because

$$s_j \begin{pmatrix} X \\ X \end{pmatrix} = \sum_{k=1}^{K+K} \sum_{k'=k}^{K+K} (x_{kj} - x_{k'j})^2 = s_j(X) + s_j(X) + \sum_{k=1}^{K} \sum_{k'=1}^{K} (x_{kj} - x_{k'j})^2 = 4s_j(X).$$

Now, using the above formulas for $d^2_{\max}$, we get

$$1 - a\begin{pmatrix} X \\ X \end{pmatrix} = \frac{d^2\begin{pmatrix} X \\ X \end{pmatrix}}{d^2_{\max}\begin{pmatrix} X \\ X \end{pmatrix}} = \frac{\sum_{j=1}^{J} s_j\begin{pmatrix} X \\ X \end{pmatrix}}{4d^2_{\max}(X)} = \frac{\sum_{j=1}^{J} 4s_j(X)}{4d^2_{\max}(X)} = 1 - a(X).$$

So, $a\begin{pmatrix} X \\ X \end{pmatrix} = a(X)$, which confirms the k-scalability of the index.

If two observation sets with different numbers of raters are compared, the a_d coefficients can easily be transformed into each other as follows (again assuming even K and K' for simplicity):

The squared column sums of the total $(K + K') \times J$ observation matrix $\begin{pmatrix} X \\ X' \end{pmatrix}$ can now be calculated as

$$s_j\begin{pmatrix} X \\ X' \end{pmatrix} = \sum_{k=1}^{K+K'} \sum_{k'=k}^{K+K'} (x_{kj} - x_{k'j})^2 = s_j(X) + s_j(X') + \sum_{k=1}^{K} \sum_{k'=1}^{K'} (x_{kj} - x'_{k'j})^2.$$

If we define the cross sum of X and X' to be $d^2(X\|X') = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{k'=1}^{K'} (x_{kj} - x'_{k'j})^2$, then clearly $d^2\begin{pmatrix} X \\ X' \end{pmatrix} = d^2(X) + d^2(X') + 2d^2(X\|X')$ and therefore

$$1 - a\begin{pmatrix} X \\ X' \end{pmatrix} = \frac{d^2\begin{pmatrix} X \\ X' \end{pmatrix}}{d^2_{\max}\begin{pmatrix} X \\ X' \end{pmatrix}} = \frac{d^2(X)}{\frac{(k+k')^2}{k^2} d^2_{\max}(X)} + \frac{d^2(X')}{\frac{(k+k')^2}{k'^2} d^2_{\max}(X')} + 2\frac{d^2(X\|X')}{\frac{(k+k')^2}{k^2} d^2_{\max}(X)}$$

$$= \frac{1}{(k+k')^2}\left\{k^2\left[1 - a(X)\right] + k'^2\left[1 - a(X')\right] + 2k^2 d(X\|X')/d^2_{\max}(X)\right\}$$

so

$$(k+k')^2\left(1 - a\begin{pmatrix} X \\ X' \end{pmatrix}\right) = k^2\left[1 - a(X)\right] + k'^2\left[1 - a(X')\right] + 2k^2 d(X\|X')/d^2_{\max}(X),$$

or, for the sake of symmetry,

$$(k+k')^2\left(1 - a\begin{pmatrix} X \\ X' \end{pmatrix}\right)$$

$$= k^2\left[1 - a(X)\right] + k'^2\left[1 - a(X')\right] + k^2 d(X\|X')/d^2_{\max}(X) + k'^2 d(X\|X')/d^2_{\max}(X').$$

The fact that $a(X, X) = a(X)$ and $a\begin{pmatrix} X \\ X \end{pmatrix} = a(X)$ shows that the a_d coefficient represents the agreement inherent in the data and that all a_d coefficients can be compared with one another regardless of the number of items or raters.

A4. Interval scaling of a_d

a_d is interval scaled with constant K and J and identical scale points, i.e. if X, X', and Y denote three observation matrices of size $K \times J$, and $a(X) - a(Y) = a(Y) - a(X')$, then the difference in agreement between X and Y implies the same as the difference in agreement between Y and X', and is half the difference between X and X' regardless of the value of the a_d, because

$$a(X) - a(Y) = a(Y) - a(X'),$$

$$1 - \frac{d_X^2}{d_{\max}^2} - 1 + \frac{d_Y^2}{d_{\max}^2} = 1 - \frac{d_Y^2}{d_{\max}^2} - 1 + \frac{d_{X'}^2}{d_{\max}^2},$$

$$d_Y^2 - d_X^2 = d_{X'}^2 - d_Y^2,$$

$$d_Y^2 = \frac{d_{X'}^2 + d_X^2}{2}.$$

As $d_{\max}^2$ is equal for each matrix, their agreement depends only on the squared Euclidean distances (d_X^2, $d_{X'}^2$, d_Y^2), which are interval scaled.

A5. a_d *for* J *items equals the mean of the* a_d *of the single items*

The mean of the a_d of the single items can be transformed into the calculation of a_d for J items:

$$\frac{1}{J}\sum_{j=1}^{J} a_d(j) = \frac{1}{J}\left(1 - \frac{d^2(1)}{d_{\max}^2(1)} + 1 - \frac{d^2(2)}{d_{\max}^2(2)} + \cdots + 1 - \frac{d^2(J)}{d_{\max}^2(J)}\right).$$

Since all items are based on the same scale and rated by every rater, $d_{\max}^2(j)$ of each item is the same. For $j = 1$ we can substitute $d^2(j)$ with s_j (see Appendix A1). So

$$\frac{1}{J}\sum_{j=1}^{J} a_d(j) = \frac{1}{J}\left(J - \frac{\sum_{j=1}^{J} s_j}{d_{\max}^2(j)}\right) = 1 - \frac{\sum_{j=1}^{J} s_j}{J d_{\max}^2(j)} = a_d.$$

Appendix B: Critical values for a_d when testing the statistical significance

Table B1. Upper bound for the 95th percentile of the a_d coefficient with $A = 5$

			p or $1 - p$								p or $1 - p$				
K	J	U	.1	.2	.3	.4	.5	K	J	U	.1	.2	.3	.4	.5
3	1	1.00	1.00	1.00	1.00	1.00	1.00	8	1	.88	.97	.95	.94	.94	.94
	2	.94	1.00	.97	.97	.97	.97		2	.83	.97	.94	.91	.89	.89
	3	.92	.98	.96	.96	.96	.94		3	.82	.96	.92	.89	.88	.87
	4	.91	.98	.95	.94	.94	.94		4	.80	.96	.92	.88	.87	.86
	5	.89	.98	.95	.94	.93	.93		5	.79	.95	.91	.88	.86	.85
	6	.88	.98	.95	.93	.92	.92		6	.79	.95	.91	.87	.85	.85
	7	.87	.97	.95	.92	.91	.91		7	.78	.95	.90	.87	.85	.84
	8	.86	.97	.94	.92	.91	.90		8	.78	.95	.90	.87	.85	.84
	9	.86	.97	.94	.92	.90	.90		9	.78	.95	.90	.86	.84	.84
	10	.86	.97	.94	.91	.90	.89		10	.77	.95	.90	.86	.84	.83
4	1	.95	1.00	1.00	.95	.95	.95	9	1	.86	.97	.94	.94	.92	.92
	2	.91	.98	.97	.95	.95	.95		2	.82	.97	.93	.90	.89	.88
	3	.89	.98	.95	.94	.93	.92		3	.80	.96	.91	.89	.87	.86
	4	.87	.98	.95	.93	.92	.91		4	.79	.95	.91	.88	.86	.85
	5	.86	.97	.94	.92	.91	.91		5	.78	.95	.91	.87	.85	.84
	6	.85	.97	.94	.91	.90	.90		6	.78	.95	.90	.87	.85	.84
	7	.85	.97	.93	.91	.90	.89		7	.77	.95	.90	.86	.84	.83
	8	.84	.96	.93	.91	.89	.89		8	.77	.95	.90	.86	.84	.83
	9	.84	.96	.93	.90	.89	.88		9	.77	.94	.89	.86	.83	.83
	10	.83	.96	.93	.90	.88	.88		10	.76	.94	.89	.85	.83	.82
5	1	.94	1.00	.96	.96	.96	.96	10	1	.85	.98	.95	.93	.91	.91
	2	.89	.98	.95	.94	.93	.92		2	.81	.96	.92	.90	.88	.88
	3	.85	.97	.94	.92	.91	.90		3	.80	.96	.91	.88	.86	.86
	4	.84	.97	.93	.91	.90	.89		4	.79	.95	.91	.87	.85	.85
	5	.83	.96	.93	.90	.89	.88		5	.78	.95	.90	.87	.85	.84
	6	.82	.96	.92	.90	.88	.88		6	.77	.95	.90	.86	.84	.83
	7	.82	.96	.92	.89	.88	.87		7	.77	.95	.90	.86	.84	.83
	8	.81	.96	.92	.89	.87	.86		8	.77	.94	.89	.86	.83	.83
	9	.81	.96	.91	.88	.87	.86		9	.76	.94	.89	.85	.83	.82
	10	.80	.95	.91	.88	.86	.86		10	.76	.94	.89	.85	.83	.82
6	1	.92	1.00	.97	.94	.94	.94	11	1	.85	.98	.94	.92	.90	.90
	2	.87	.97	.94	.93	.91	.91		2	.81	.96	.92	.89	.87	.87
	3	.84	.97	.94	.91	.90	.89		3	.79	.95	.91	.88	.86	.85
	4	.83	.96	.93	.90	.89	.88		4	.78	.95	.90	.87	.85	.84
	5	.82	.96	.92	.89	.88	.87		5	.77	.95	.90	.86	.84	.83
	6	.81	.96	.92	.89	.87	.87		6	.77	.95	.89	.86	.84	.83
	7	.81	.96	.91	.88	.87	.86		7	.76	.94	.89	.85	.83	.82
	8	.80	.95	.91	.88	.86	.86		8	.76	.94	.89	.85	.83	.82
	9	.80	.95	.91	.88	.86	.85		9	.76	.94	.89	.85	.83	.82
	10	.80	.95	.91	.88	.86	.85		10	.75	.94	.89	.85	.82	.82
7	1	.90	1.00	.95	.95	.94	.94	12	1	.84	.97	.94	.91	.90	.90
	2	.84	.97	.94	.92	.90	.90		2	.80	.96	.92	.89	.87	.86
	3	.82	.96	.93	.90	.89	.88		3	.79	.95	.91	.87	.85	.85
	4	.81	.96	.92	.89	.87	.87		4	.77	.95	.90	.87	.85	.84
	5	.80	.96	.91	.88	.86	.86		5	.77	.95	.90	.86	.84	.83
	6	.80	.95	.91	.88	.86	.85		6	.76	.94	.89	.86	.83	.83
	7	.79	.95	.91	.87	.85	.85		7	.76	.94	.89	.85	.83	.82
	8	.79	.95	.90	.87	.85	.84		8	.76	.94	.89	.85	.83	.82
	9	.78	.95	.90	.87	.85	.84		9	.75	.94	.89	.85	.82	.82
	10	.78	.95	.90	.86	.84	.84		10	.75	.94	.88	.85	.82	.81

Note. K, number of raters; J, number of items; p, probability from the binomial distribution; U, critical values according to the uniform distribution.

Table B2. Upper bound for the 99th percentile of the a_d coefficient with $A = 5$

			p or $1 - p$								p or $1 - p$				
K	J	U	.1	.2	.3	.4	.5	K	J	U	.1	.2	.3	.4	.5
3	1	1.00	1.00	1.00	1.00	1.00	1.00	8	1	.94	1.00	.97	.95	.95	.95
	2	.97	1.00	1.00	1.00	.97	.97		2	.88	.98	.95	.94	.92	.92
	3	.96	1.00	.98	.98	.96	.96		3	.85	.97	.94	.92	.90	.90
	4	.94	.99	.97	.97	.95	.95		4	.83	.97	.93	.91	.89	.89
	5	.93	.99	.96	.96	.95	.95		5	.82	.96	.93	.90	.88	.88
	6	.92	.98	.96	.95	.95	.94		6	.81	.96	.92	.89	.88	.87
	7	.90	.98	.96	.95	.94	.94		7	.81	.96	.92	.89	.87	.86
	8	.90	.98	.95	.94	.93	.93		8	.80	.96	.91	.88	.87	.86
	9	.89	.98	.95	.94	.92	.92		9	.80	.95	.91	.88	.86	.86
	10	.89	.97	.95	.93	.92	.92		10	.79	.95	.91	.88	.86	.85
4	1	1.00	1.00	1.00	1.00	1.00	1.00	9	1	.92	1.00	.96	.96	.94	.94
	2	.95	1.00	.98	.98	.98	.97		2	.86	.98	.95	.93	.92	.91
	3	.93	.98	.97	.96	.95	.95		3	.84	.97	.93	.91	.90	.89
	4	.91	.99	.96	.95	.95	.94		4	.82	.96	.93	.90	.88	.88
	5	.90	.98	.96	.94	.93	.93		5	.81	.96	.92	.89	.88	.87
	6	.89	.98	.95	.93	.93	.92		6	.80	.96	.91	.88	.87	.86
	7	.88	.98	.95	.93	.92	.92		7	.80	.96	.91	.88	.86	.86
	8	.87	.97	.95	.93	.91	.91		8	.79	.95	.91	.88	.86	.85
	9	.86	.97	.94	.92	.91	.91		9	.79	.95	.91	.87	.85	.85
	10	.86	.97	.94	.92	.91	.90		10	.78	.95	.90	.87	.85	.84
5	1	.96	1.00	1.00	1.00	.96	.96	10	1	.90	1.00	.96	.95	.95	.95
	2	.93	1.00	.97	.96	.95	.95		2	.85	.97	.94	.92	.91	.91
	3	.90	.99	.96	.94	.94	.93		3	.83	.97	.93	.91	.89	.89
	4	.88	.98	.95	.93	.92	.92		4	.81	.96	.92	.89	.88	.87
	5	.86	.97	.95	.93	.91	.91		5	.80	.96	.92	.89	.87	.86
	6	.85	.97	.94	.92	.90	.90		6	.80	.96	.91	.88	.86	.86
	7	.85	.97	.93	.91	.90	.89		7	.79	.95	.91	.88	.86	.85
	8	.84	.97	.93	.91	.89	.89		8	.79	.95	.91	.87	.85	.85
	9	.83	.97	.93	.90	.89	.88		9	.78	.95	.90	.87	.85	.84
	10	.83	.96	.93	.90	.88	.88		10	.78	.95	.90	.87	.85	.84
6	1	.97	1.00	.97	.97	.97	.97	11	1	.89	.98	.95	.95	.94	.94
	2	.91	.98	.97	.95	.94	.94		2	.84	.97	.94	.91	.90	.90
	3	.88	.98	.95	.93	.92	.92		3	.82	.96	.93	.90	.88	.88
	4	.86	.97	.94	.92	.91	.91		4	.80	.96	.92	.89	.87	.86
	5	.85	.97	.94	.92	.90	.90		5	.79	.96	.91	.88	.86	.86
	6	.84	.97	.93	.91	.89	.89		6	.79	.95	.91	.88	.86	.85
	7	.83	.97	.93	.90	.89	.88		7	.78	.95	.90	.87	.85	.84
	8	.83	.96	.93	.90	.88	.88		8	.78	.95	.90	.87	.85	.84
	9	.82	.96	.92	.90	.88	.88		9	.77	.95	.90	.86	.84	.84
	10	.82	.96	.92	.89	.88	.87		10	.77	.95	.90	.86	.84	.83
7	1	.95	1.00	.97	.97	.97	.97	12	1	.88	.98	.95	.94	.94	.94
	2	.89	.98	.95	.94	.93	.93		2	.84	.97	.94	.91	.90	.89
	3	.86	.97	.94	.92	.91	.91		3	.81	.96	.92	.90	.88	.87
	4	.84	.97	.93	.91	.90	.90		4	.80	.96	.92	.89	.87	.86
	5	.83	.97	.93	.90	.89	.89		5	.79	.96	.91	.88	.86	.85
	6	.82	.96	.93	.90	.88	.88		6	.78	.95	.91	.87	.85	.85
	7	.82	.96	.92	.89	.88	.87		7	.78	.95	.90	.87	.85	.84
	8	.81	.96	.92	.89	.87	.86		8	.77	.95	.90	.87	.84	.84
	9	.80	.96	.92	.89	.87	.86		9	.77	.95	.90	.86	.84	.83
	10	.80	.96	.91	.88	.86	.86		10	.77	.95	.90	.86	.84	.83

Note. K, number of raters; J, number of items; p, probability from the binomial distribution; U, critical values according to the uniform distribution.

Table B3. Upper bound for the 95th percentile of the a_d coefficient with $A = 7$

			p or $1 - p$								p or $1 - p$				
K	J	U	.1	.2	.3	.4	.5	K	J	U	.1	.2	.3	.4	.5
3	1	.97	1.00	1.00	1.00	1.00	1.00	8	1	.88	.98	.97	.96	.95	.95
	2	.94	.99	.99	.99	.97	.97		2	.83	.98	.96	.94	.93	.93
	3	.92	.99	.98	.97	.96	.96		3	.81	.97	.95	.93	.92	.92
	4	.90	.99	.97	.97	.96	.96		4	.80	.97	.94	.92	.91	.91
	5	.88	.98	.97	.96	.95	.95		5	.79	.97	.94	.92	.91	.90
	6	.88	.98	.96	.95	.94	.94		6	.78	.97	.94	.92	.90	.90
	7	.87	.98	.96	.95	.94	.94		7	.78	.97	.94	.91	.90	.90
	8	.86	.98	.96	.95	.94	.94		8	.77	.96	.93	.91	.90	.89
	9	.85	.98	.96	.94	.94	.93		9	.77	.96	.93	.91	.90	.89
	10	.85	.98	.96	.94	.93	.93		10	.77	.96	.93	.91	.89	.89
4	1	.97	1.00	.98	.98	.98	.98	9	1	.86	.98	.96	.95	.95	.95
	2	.92	.99	.98	.97	.96	.96		2	.82	.97	.95	.94	.93	.92
	3	.89	.98	.97	.96	.96	.95		3	.80	.97	.94	.93	.91	.91
	4	.87	.98	.97	.95	.95	.94		4	.79	.97	.94	.92	.91	.90
	5	.86	.98	.96	.95	.94	.94		5	.78	.97	.94	.91	.90	.90
	6	.85	.98	.96	.94	.94	.93		6	.77	.96	.93	.91	.90	.89
	7	.84	.98	.96	.94	.93	.93		7	.77	.96	.93	.91	.90	.89
	8	.84	.98	.95	.94	.93	.93		8	.76	.96	.93	.91	.89	.89
	9	.83	.97	.95	.94	.93	.92		9	.76	.96	.93	.91	.89	.89
	10	.83	.97	.95	.93	.92	.92		10	.76	.96	.93	.90	.89	.88
5	1	.94	.98	.98	.98	.97	.97	10	1	.86	.98	.96	.95	.95	.94
	2	.88	.98	.97	.96	.95	.95		2	.81	.97	.95	.93	.92	.92
	3	.85	.98	.96	.95	.94	.94		3	.79	.97	.94	.92	.91	.91
	4	.84	.98	.96	.94	.93	.93		4	.78	.97	.94	.92	.90	.90
	5	.83	.98	.95	.94	.93	.92		5	.77	.97	.93	.91	.90	.90
	6	.82	.97	.95	.93	.92	.92		6	.77	.96	.93	.91	.90	.89
	7	.81	.97	.95	.93	.92	.91		7	.76	.96	.93	.91	.89	.89
	8	.81	.97	.94	.93	.91	.91		8	.76	.96	.93	.91	.89	.89
	9	.80	.97	.94	.92	.91	.91		9	.76	.96	.93	.90	.89	.88
	10	.80	.97	.94	.92	.91	.91		10	.75	.96	.93	.90	.89	.88
6	1	.91	.98	.98	.98	.97	.97	11	1	.84	.98	.96	.95	.94	.94
	2	.86	.98	.96	.95	.94	.94		2	.80	.97	.95	.93	.92	.91
	3	.84	.98	.96	.94	.93	.93		3	.78	.97	.94	.92	.91	.90
	4	.82	.98	.95	.93	.93	.92		4	.77	.97	.94	.91	.90	.90
	5	.81	.97	.95	.93	.92	.92		5	.77	.96	.93	.91	.90	.89
	6	.81	.97	.94	.93	.92	.91		6	.76	.96	.93	.91	.89	.89
	7	.80	.97	.94	.92	.91	.91		7	.76	.96	.93	.90	.89	.88
	8	.80	.97	.94	.92	.91	.91		8	.75	.96	.93	.90	.89	.88
	9	.79	.97	.94	.92	.91	.90		9	.75	.96	.93	.90	.89	.88
	10	.79	.97	.94	.92	.91	.90		10	.75	.96	.92	.90	.88	.88
7	1	.89	.99	.97	.97	.95	.95	12	1	.84	.98	.96	.94	.94	.94
	2	.84	.98	.96	.94	.94	.93		2	.80	.97	.95	.93	.91	.91
	3	.82	.98	.95	.93	.92	.92		3	.78	.97	.94	.92	.90	.90
	4	.81	.97	.95	.93	.92	.91		4	.77	.96	.93	.91	.90	.89
	5	.80	.97	.94	.92	.91	.91		5	.76	.96	.93	.91	.89	.89
	6	.79	.97	.94	.92	.91	.90		6	.76	.96	.93	.90	.89	.89
	7	.78	.97	.94	.92	.90	.90		7	.75	.96	.93	.90	.89	.88
	8	.78	.97	.94	.91	.90	.90		8	.75	.96	.93	.90	.89	.88
	9	.78	.97	.93	.91	.90	.90		9	.75	.96	.92	.90	.88	.88
	10	.77	.96	.93	.91	.90	.89		10	.75	.96	.92	.90	.88	.88

Note. K, number of raters; J, number of items; p, probability from the binomial distribution; U, critical values according to the uniform distribution.

Table B4. Upper bound for the 99th percentile of the a_d coefficient with $A = 7$

			p or $1 - p$								p or $1 - p$				
K	J	U	.1	.2	.3	.4	.5	K	J	U	.1	.2	.3	.4	.5
3	1	1.00	1.00	1.00	1.00	1.00	1.00	8	1	.92	.99	.98	.97	.97	.97
	2	.97	1.00	.99	.99	.99	.99		2	.87	.98	.97	.96	.95	.95
	3	.95	.99	.99	.98	.98	.98		3	.85	.98	.96	.95	.94	.94
	4	.94	.99	.98	.98	.97	.97		4	.83	.98	.95	.94	.93	.93
	5	.92	.99	.98	.97	.97	.97		5	.82	.97	.95	.93	.92	.92
	6	.91	.99	.98	.97	.96	.96		6	.81	.97	.95	.93	.92	.92
	7	.90	.99	.97	.96	.96	.96		7	.80	.97	.94	.93	.91	.91
	8	.90	.99	.97	.96	.95	.95		8	.80	.97	.94	.92	.91	.91
	9	.89	.98	.97	.96	.95	.95		9	.79	.97	.94	.92	.91	.91
	10	.88	.98	.97	.96	.95	.95		10	.79	.97	.94	.92	.91	.90
4	1	.98	1.00	1.00	1.00	1.00	1.00	9	1	.91	.99	.98	.97	.97	.97
	2	.95	.99	.99	.98	.98	.98		2	.86	.98	.96	.95	.94	.94
	3	.93	.99	.98	.97	.97	.97		3	.83	.98	.96	.94	.93	.93
	4	.91	.99	.98	.97	.96	.96		4	.82	.97	.95	.93	.92	.92
	5	.89	.99	.97	.96	.96	.95		5	.80	.97	.95	.93	.92	.91
	6	.88	.98	.97	.96	.95	.95		6	.80	.97	.94	.92	.91	.91
	7	.88	.98	.97	.95	.95	.95		7	.79	.97	.94	.92	.91	.91
	8	.87	.98	.96	.95	.94	.94		8	.79	.97	.94	.92	.91	.90
	9	.86	.98	.96	.95	.94	.94		9	.78	.97	.94	.92	.90	.90
	10	.86	.98	.96	.95	.94	.94		10	.78	.97	.94	.91	.90	.90
5	1	.97	1.00	1.00	.98	.98	.98	10	1	.91	.99	.98	.97	.96	.96
	2	.93	.99	.98	.98	.97	.97		2	.85	.98	.96	.95	.94	.94
	3	.90	.98	.97	.96	.96	.96		3	.82	.98	.95	.94	.93	.93
	4	.88	.98	.97	.96	.95	.95		4	.81	.97	.95	.93	.92	.92
	5	.86	.98	.96	.95	.94	.94		5	.80	.97	.94	.93	.91	.91
	6	.85	.98	.96	.95	.94	.94		6	.79	.97	.94	.92	.91	.91
	7	.84	.98	.96	.94	.93	.93		7	.79	.97	.94	.92	.91	.90
	8	.84	.98	.95	.94	.93	.93		8	.78	.97	.94	.92	.90	.90
	9	.83	.98	.95	.94	.93	.92		9	.78	.97	.94	.91	.90	.90
	10	.82	.97	.95	.93	.92	.92		10	.77	.97	.93	.91	.90	.89
6	1	.95	1.00	.98	.98	.98	.98	11	1	.89	.98	.97	.96	.96	.96
	2	.91	.99	.98	.97	.96	.96		2	.84	.98	.96	.94	.94	.93
	3	.88	.98	.97	.96	.95	.95		3	.81	.97	.95	.93	.92	.92
	4	.86	.98	.96	.95	.94	.94		4	.80	.97	.95	.93	.92	.91
	5	.85	.98	.96	.94	.94	.93		5	.79	.97	.94	.92	.91	.91
	6	.84	.98	.96	.94	.93	.93		6	.78	.97	.94	.92	.91	.90
	7	.83	.98	.95	.94	.93	.92		7	.78	.97	.94	.92	.90	.90
	8	.82	.97	.95	.93	.92	.92		8	.77	.97	.93	.91	.90	.90
	9	.82	.97	.95	.93	.92	.92		9	.77	.96	.93	.91	.90	.89
	10	.81	.97	.95	.93	.92	.92		10	.77	.96	.93	.91	.89	.89
7	1	.94	1.00	.99	.98	.98	.98	12	1	.88	.98	.97	.96	.96	.95
	2	.89	.99	.97	.96	.96	.95		2	.83	.98	.96	.94	.93	.93
	3	.86	.98	.96	.95	.94	.94		3	.81	.97	.95	.93	.92	.92
	4	.84	.98	.96	.94	.93	.93		4	.80	.97	.94	.92	.91	.91
	5	.83	.98	.95	.94	.93	.92		5	.79	.97	.94	.92	.91	.90
	6	.82	.97	.95	.93	.92	.92		6	.78	.97	.94	.92	.90	.90
	7	.81	.97	.95	.93	.92	.92		7	.77	.97	.93	.91	.90	.90
	8	.80	.97	.95	.93	.91	.91		8	.77	.96	.93	.91	.90	.89
	9	.80	.97	.94	.92	.91	.91		9	.77	.96	.93	.91	.90	.89
	10	.80	.97	.94	.92	.91	.91		10	.76	.96	.93	.91	.89	.89

Note. K, number of raters; J, number of items; p, probability from the binomial distribution; U, critical values according to the uniform distribution.