APPENDIX: When Experts Disagree

March 23, 2018

A1 Agreement and Interrater Reliability Measures

Beyond political science, a large literature in organizational psychology and medicine seeks to understand how to best assess agreement and reliability in responses to particular survey items. The same issues that crop up in political science expert surveys arise in any analysis in which K raters score N targets on J items. This literature makes a significant distinction between reliability and agreement. Kozlowski and Hattrup (1992, pp. 162–63) describe reliability "as an index of consistency; it references proportional consistency of variance among raters [...] and is correlational in nature [...]. In contrast, agreement references the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings." In other words, reliability is concerned with the equivalence of *relative* ratings of experts across items, whereas agreement refers to the absolute consensus among raters on one or more items (LeBreton and Senter, 2008, p. 816). Thus, there could be relatively high reliability among raters, but low agreement. For example, all raters may rate party A higher than party B and party C, but they may use the scale differently. On an eleven-point 0–10 left-right scale, perhaps Rater 1 assigns a 10, 8, and 6 to parties A, B and C respectively; Rater 2 assigns them scores of 8, 6, and 4; and Rater 3 rates them 4, 2, and 0. There would be perfect reliability among these scores, but little agreement. It is worth noting that reliability can only be assessed when the same raters rate multiple targets (e.g., parties), and thus can only be assessed at the item level. Agreement, in contrast, can be

assessed at the target-item level.

We discussed that the literature on interrater agreement has deemed the standard error to be an inappropriate measure of agreement, but it also suggests that the standard deviation is not much better. The standard deviation is a measure of dispersion, rather than agreement. There are two primary drawbacks to the standard deviation as a measure of agreement (Kozlowski and Hattrup, 1992). First, the standard deviation is scale-dependent — items assessed on a Likert scale ranging from 0–10 will likely have a smaller standard deviation than those assessed on a 0–100 scale — such that we can only compare standard deviations of items that are measured on the same scale. Second, it does not account for within-group agreement that could occur due to chance. The most common measure of agreement, the r_{wg} (Finn, 1970; James, Demaree and Wolf, 1984), does both by examining the dispersion of responses with reference to a null distribution.[1] It is calculated as

$$r_{wg} = 1 - \frac{S_x^2}{\sigma_E^2}, \tag{1}$$

where S_x^2 is the observed variance of expert response on the item x, and σ_E^2 is the expected variance when there is a complete lack of agreement among experts.[2] The measure ranges

[1]Within political science, van der Eijk (2001) has proposed a different, scale independent measure of agreement, but his measure does not make use of a null distribution. Other measures for assessing agreement commonly used when coding data (e.g. content analysis), such as Krippendorff's α (Krippendorff, 2011), assess the level agreement among coders when rating N units (e.g. parties), but do not allow researchers to assess agreement at the unit level. In this literature, the concern is more about the performance of coders rather than what disagreement tells us about the items being coded.

[2]One could also calculate agreement across multiple items, using the $r_{wg}(j)$ measure, if the items were essentially parallel, meaning that they measure the same construct. Given that most items in political science surveys tap into different dimensions, this measure is less appropriate for our purposes.

from 0 (no agreement) to 1 (perfect agreement) and can be interpreted as the proportional reduction in error variance.[3] Of course, r_{wg} requires researchers to choose an appropriate null distribution. In practice, researchers typically use a rectangular or uniform distribution, estimated as $\frac{A^2-1}{12}$, where A is the number of response options. But any number of distributions could be used, and ideally one's results would be robust to the choice of the null distribution (Meyer et al., 2014). Below, we calculate agreement scores for the CHES survey using the rectangular distribution as well as the triangular distribution as the reference distribution.[4]

While the r_{wg} measure captures agreement, it does not take into account differences in how raters may use the scale. For that, we need a measure of reliability, which examines how experts place multiple targets, the most common of which is the intraclass correlation coefficient (ICC) (LeBreton and Senter, 2008). This measure has the advantage of capturing both the consensus among as well as the consistency across judges, but it cannot be used to evaluate agreement on particular items. In effect, if we knew that judges used the scales in the same manner, we could arrive at the same answer by simply aggregating the r_{wg} scores

[3]In rare instances, it could be negative, meaning there is more observed variance than we would expect according to the assumption of the null distribution. In these instances, it is usually truncated at zero.

[4]The formula for the latter is as follows (James, Demaree and Wolf, 1984):

$$\sigma^2_{ET} = \begin{cases} \frac{(A-1)(A+3)}{24} & \text{for A odd, and} \\ \frac{A^2+2A-2}{24} & \text{for A even} \end{cases} \tag{2}$$

Using the rectangular distribution as reference implies that experts are essentially selecting party policy positions at random, whereas the triangular reference distribution implies a central tendency bias. While the rectangular distribution is the standard go-to reference distribution in the applied literature, the assumption of essentially random placement is mostly unrealistic. As such, agreement — the proportional reduction in error variance — tends to be overstated. The triangular distribution, which reflects a central tendency bias, is more realistic. The choice of reference distribution ultimately needs to be theoretically motivated (Meyer et al., 2014).

over the targets on each item. For the purposes of our analysis, we focus on single items and therefore set aside the issue of reliability across multiple items.[5] In the next section, we describe Monte Carlo simulations to explore the issues of drastically different expert distributions, how to aggregate responses in the face of disagreement, and the consequences of these choices.[6]

A2 Agreement in the CHES Expert Survey

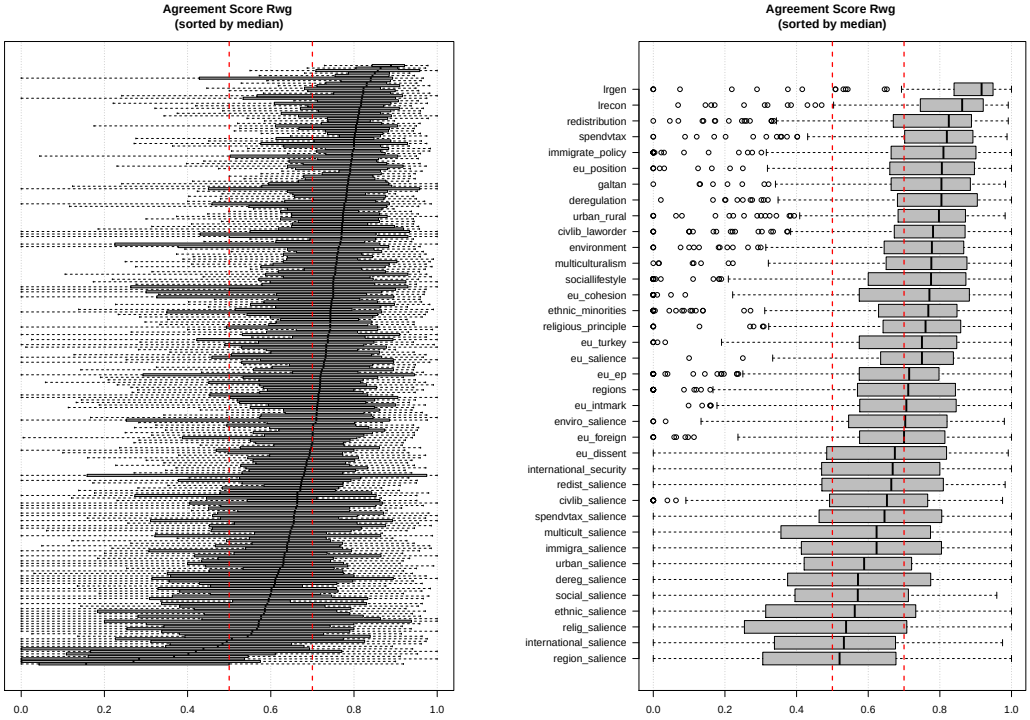
We apply our measures of agreement to different waves of the CHES expert surveys. We first calculate the r_{wg} coefficient, which captures agreement only. The advantage of the r_{wg} coefficient is that, since we are only concerned with agreement and not reliability, it can be applied to the party-dimension level. In Figures A1, we use box plots to display the distribution of r_{wg} coefficients by dimension across parties and by party across dimensions. The plots use a rectangular reference distribution and refer to the 2010 CHES wave. There are no hard and fast rules as to what constitutes an acceptable level of agreement, but the extant literature often considers scores in excess of 0.7 to be indicative of strong agreement and scores below 0.5 of weak agreement. In Figure A1, these two cut-off values are demarcated with vertical dashed lines.

The first plot, Figure A1a, displays a box plot for each party in the CHES survey. Due

[5]As previously discussed, the political science literature has provided alternative approaches to dealing with scale perception issues (e.g., Aldrich and McKelvey, 1977; King and Wand, 2007; Lo, Proksch and Gschwend, 2014; Bakker, Edwards, Jolly, Polk, Rovny and Steenbergen, 2014; Bakker, Jolly, Polk and Poole, 2014; Pemstein, Tzelgov and Wang, 2015). Here we are more concerned with the conditions under which aggregation makes sense, at all, and how truncation bias on Likert scales affects different methods for aggregation.

[6]The existing literature does not discuss problems of aggregation beyond diagnosing levels of variance in responses. But see Beal and Dawson (2007) who discuss how aggregation of truncated Likert scales affects measures of intraclass correlation.

Figure A1: *Measure of agreement (r_{wg}) by party and policy dimension (using rectangular reference distribution).*



(a) r_{wg} by party

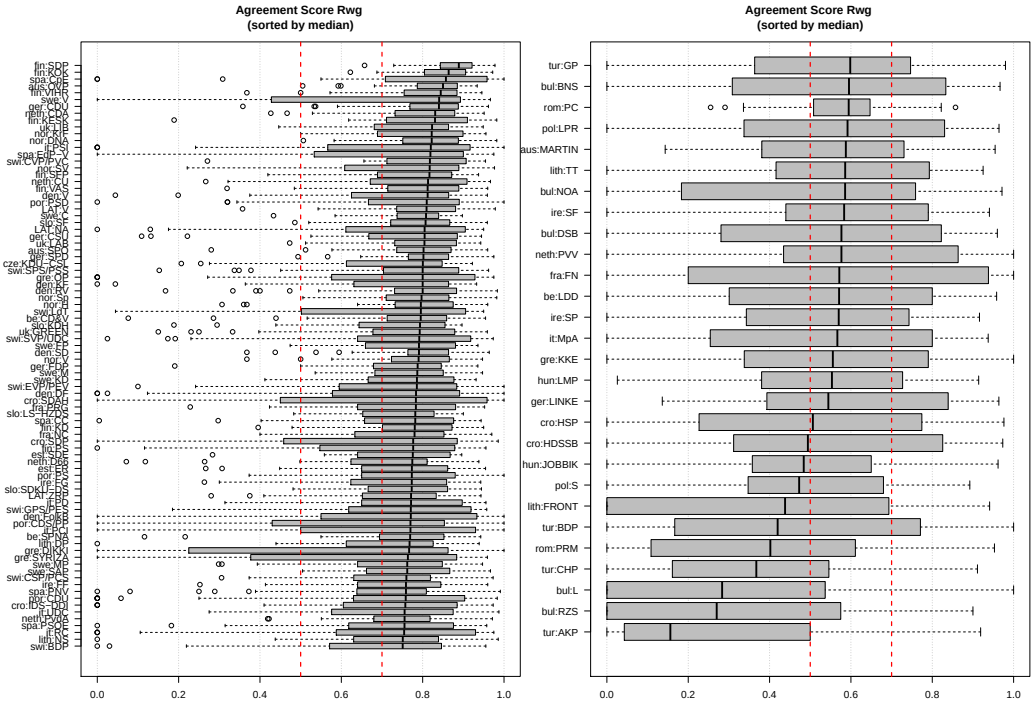
(b) r_{wg} by dimension

to the large number of parties, it is difficult to assess how any particular party is measured. Thus, we suppress the party names on the y-axis and do not plot outlying points in this figure. The plot shows the overall distribution of the levels of agreement across all parties. There are many parties for which the median r_{wg} over the dimensions is quite a bit better than the 0.7 cut-off for strong agreement when we assume a rectangular reference distribution, which gives the best-case scenario for finding high levels of agreement among experts. It is also worth noting that even for the parties on which experts agree quite often, there are many dimensions with agreement scores in the moderate or even poor range. There are also many party-dimensions for which there is only moderate or poor agreement. Figure A1b presents box plots by dimension. We note that there is particularly low agreement on the salience items.

Next, in Figure A3 we examine the best performing parties (i.e., those with a median

$r_{wg} > 0.75$) and the worst performing parties (i.e., those with a median $r_{wg} < 0.60$). Most of the parties with the highest levels of agreement are found in Northern and Western Europe. However, some parties in post-communist countries also show high levels of agreement, namely some Latvian, Czech, and Slovenian parties. The parties with the lowest levels of agreement among experts are found in Southern and Eastern Europe. There is virtually no agreement on the positions of the Turkish parties on many dimensions. But even here, there is strong agreement on certain items for certain parties. For example, on the CHES *galtan* dimension (Green/Alternative/Libertarian to Traditional/Authoritarian/Nationalist), the ruling AK party, the MHP, the DYP and the Greens all display high agreement, while the CHP and BDP display virtually no agreement at all. Thus, while there appears to be agreement with respect to the center-right Islamist parties AKP and DYP, the ultranationalist MHP, and the Greens, there is little agreement on the secular CHP, which ruled Turkey for much of the post-war era, or the Kurdish BDP.

Figure A2: r_{wg} *by party — best and worst.*



(a) Median $r_{wg} > 0.75$ (b) Median $r_{wg} < 0.60$

The agreement scores using a rectangular reference distribution should already give us pause in many cases. But this reference distribution suggests high levels of agreement relative to other reference distributions as it assumes an essentially random placement by experts. A triangle reference distribution, which reflects a potential central tendency bias among experts, yields even less encouraging results compared to the rectangular distribution. The average agreement score for parties using this distribution is 0.55. Even more discouraging is the fact that between 10% and approximately 17% of the party-issue scales yield zero agreement amongst experts across the four waves of the expert survey we examine (1999, 2002, 2006 and 2010).

Figure A3: *Parties with highest and lowest expert agreement (CHES waves 2002, 2006, 2010) using triangular reference distribution.*

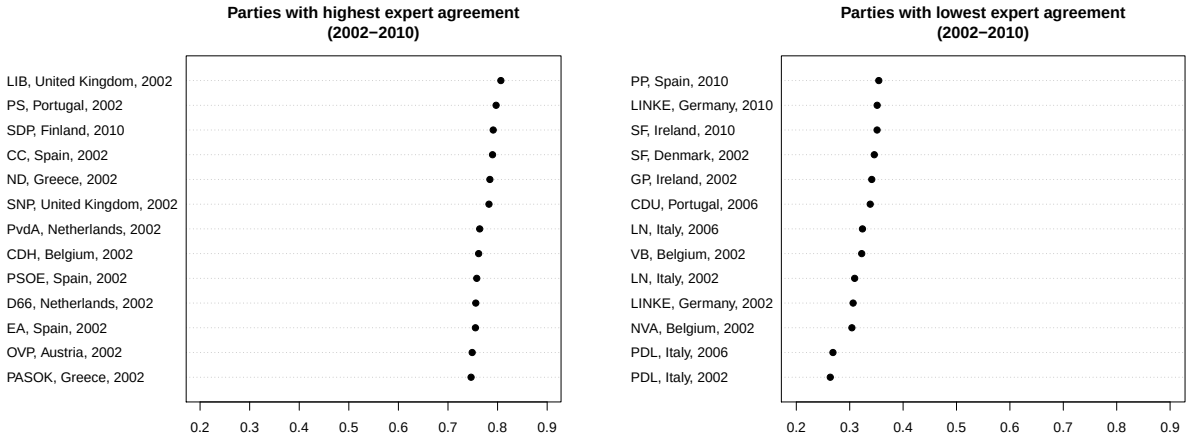


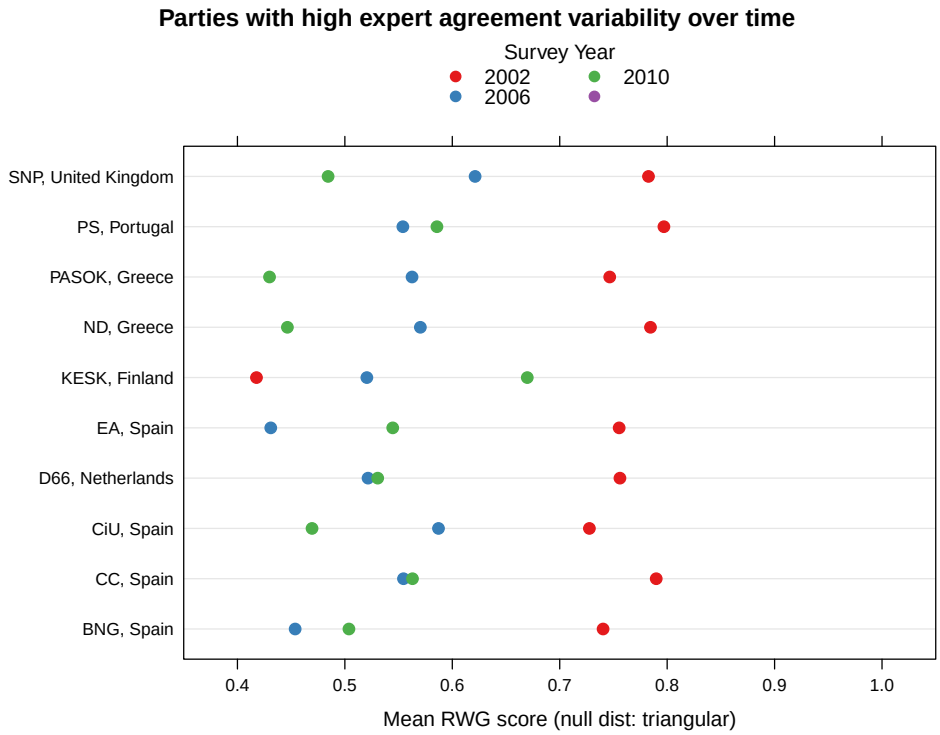
Figure A3 shows the best and worst measured parties in terms of the agreement score using a triangular reference distribution across the expert survey waves between 2002 and 2010. At the high end, experts achieve agreement on the order of 0.75, whereas at the low end agreement ranges between around 0.25 and 0.35. What is more disconcerting, however, is that the lists of parties on the low as well as the high end of agreement are not necessarily intuitive. For example, the German *Die Linke* is one of the parties with high expert disagreement across the different waves of the expert survey. Experts have no problem placing this party on the left-right dimension, the redistribution dimension, the religious principle dimension, or the spending versus taxation dimension. However, on other dimensions there is practically no expert agreement whatsoever, including the position of the party on multiculturalism, the new politics dimension (*galtan*), or on the environment. Moreover, salience is uniformly poorly measured across all dimensions.

Finally, we look at the variation in agreement within parties and dimensions across waves. Figure A4a shows the parties with the highest expert agreement variability across waves. The fact that some major parties, such as the Portuguese PS, show quite a bit of variation across time does not instil a lot of confidence in expert assessments. The variability across

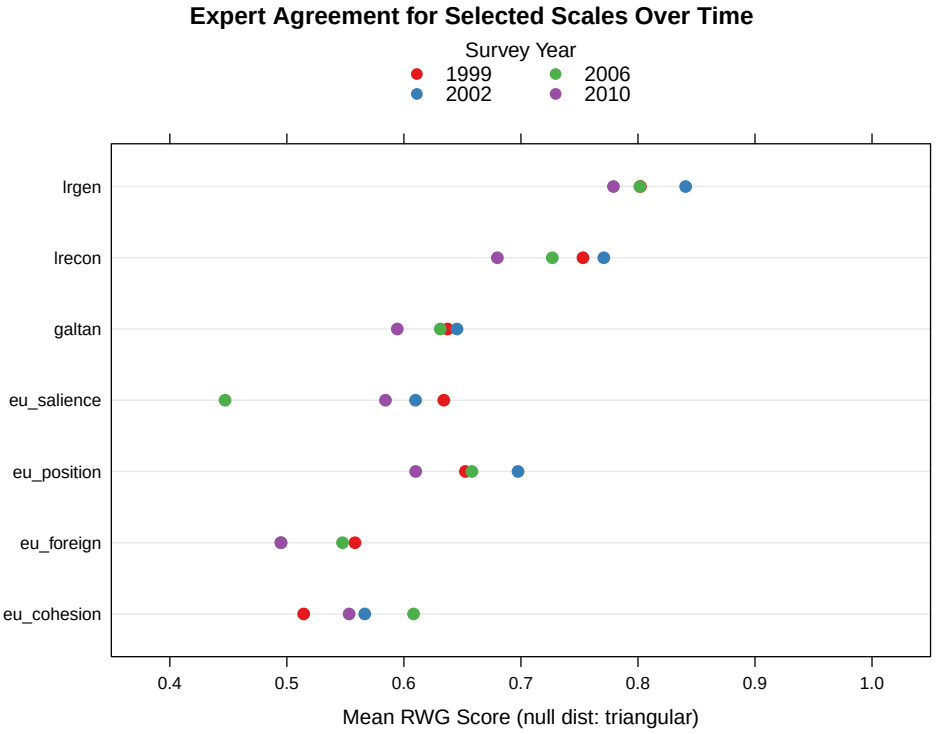
dimensions is also worrisome (see Figure A4b). Across the four waves since 1999, even the left-right economic dimension exhibits non-trivial variability in agreement. The problem only gets worse when we turn to the dimensions related to the EU or the new politics dimension (*galtan*).

Figure A4: *Expert agreement across CHES waves.*

(a)



(b)



A3 Monte Carlo Simulation: Additional Parameter Information

We draw j α's from a uniform distribution that ranges from -2 to 2. Thus, experts may shift the space by up to 2 points on the 11 point scale in either direction. We draw the j β's from a uniform distribution that ranges from 0.7 to 1.3, meaning the experts may expand or contract the space by up to 30%. Drawing the parameters this way means that some experts will use the scale almost exactly as presented to them, while others will apply quite different shift and stretch parameters. The random noise component ϵ is expert-party specific, and it is drawn from a normal distribution with a mean $\mu = 0$ and a party-specific standard deviation, σ_i, which may range from 0 to 3 again drawn from a uniform distribution. Thus, experts assess some parties better than others.

References

Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(01):111–130.

Bakker, Ryan, Erica Edwards, Seth Jolly, Jonathan Polk, Jan Rovny and Marco Steenbergen. 2014. "Anchoring the experts: Using vignettes to compare party ideology across countries." *Research & Politics* 1(3):2053168014553502.

Bakker, Ryan, Seth Jolly, Jonathan Polk and Keith Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *The Journal of Politics* 76(04):1089–1101.

Beal, Daniel J. and Jeremy F. Dawson. 2007. "On the Use of Likert-Type Scales in Multilevel Data." *Organizational Research Methods* 10(4):657–672.

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

James, Lawrence R., Robert G. Demaree and Gerrit Wolf. 1984. "Assessing within-group interrater reliability with and without response bias." *Journal of Applied Psychology* 69(1):85–98.

King, Gary and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15(1):46–66.

Kozlowski, Steve W.J. and Keith Hattrup. 1992. "A Disagreement About Within-Group Agreement: Disentangling Issues of Consistency Versus Consensus." *Journal of Applied Psychology* 77(2):161–167.

Krippendorff, Klaus. 2011. "Agreement and information in the reliability of coding." *Communication Measures and Methods* 5(2):93–112.

LeBreton, James M. and Jenell L. Senter. 2008. "Answers to 20 Questions about Interrater Reliability and Interrater Agreement." *Organizational Research Methods* 11(4):815–852.

Lo, James, Sven-Oliver Proksch and Thomas Gschwend. 2014. "A common left-right scale for voters and parties in Europe." *Political Analysis* 22(2):205–223.

Meyer, Rustin D., Troy V. Mumford, Carla J. Burrus, Michael A. Campion and Lawrence R. James. 2014. "Selecting Null Distributions When Calculation Rwg: A Tutorial and Review." *Organizational Research Methods* DOI: 10.1177/1094428114526927.

Pemstein, Daniel, Eitan Tzelgov and Yi-Ting Wang. 2015. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." Paper Presented at the 2015 Annual Meeting of the Midwest Political Science Association.

van der Eijk, Cees. 2001. "Measuring Agreement in Ordered Rating Scales." *Quality and Quantity* 35(3):325–341.