

4. John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.
5. Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *CHI 2017*.
6. Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
7. David M Corey, William P Dunlap, and Michael J Burke. 1998. Averaging correlations: Expected values and bias in combined Pearson r s and Fisher's z transformations. *The Journal of general psychology* 125, 3 (1998), 245–261.
8. Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J. Dolan. 2006. Frames, Biases, and Rational Decision-Making in the Human Brain. *Science* 313, 5787 (2006), 684–687.
<http://science.sciencemag.org/content/313/5787/684>
9. Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
10. M. Drouhard, N. C. Chen, J. Suh, R. Kocielnik, V. Pena-Araya, K. Cen, Xiangyi Zheng, and C. R. Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. 220–229.
11. Anca Dumitrache. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *European Semantic Web Conference*. Springer, 701–710.
12. Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic Why the adjustments are insufficient. *Psychological science* 17, 4 (2006), 311–318.
13. Frank Fischer, Johannes Bruhn, Cornelia Gräsel, and Heinz Mandl. 2002. Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction* 12, 2 (2002), 213–232.
14. Ronald A Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 4 (1915), 507–521.
15. Ronald A Fisher. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1 (1921), 3–32.
16. D Randy Garrison, Martha Cleveland-Innes, Marguerite Koole, and James Kappelman. 2006. Revisiting methodological issues in transcript analysis: Negotiated coding and reliability. *The Internet and Higher Education* 9, 1 (2006), 1–8.
17. Richard L Gorsuch and Curtis S Lehmann. 2010. Correlation coefficients: Mean bias and confidence interval distortions. *Journal of Methods and Measurement in the Social Sciences* 1, 2 (2010), 52–65.
18. Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48.
19. Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.
20. Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods* 16, 3 (2004), 307–331.
21. Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *CSCW 2016*.
22. Katie Kuksenok, Michael Brooks, John J Robinson, Daniel Perry, Megan K Torkildson, and Cecilia Aragon. 2012. Automating large-scale annotation for analysis of social media content. In *IEEE Workshop on Interactive Visual Text Analytics for Analysis of Social Media*.
23. Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *UIST 2014*.
24. Margaret D LeCompte. 2000. Analyzing Qualitative Data. *Theory into practice* 39, 3 (2000), 146–154.
25. Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.
26. Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20, 1 (2012), 78–91.
27. Dan J Putka, Huy Le, Rodney A McCloy, and Tirso Diaz. 2008. Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology* 93, 5 (2008), 959.
28. Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers*. Sage.
29. Anselm L Strauss. 1987. *Qualitative Analysis for Social Scientists*. Cambridge University Press.
30. Jasy Liew Suet Yan, Nancy McCracken, Shichun Zhou, and Kevin Crowston. 2014. Optimizing features in active machine learning for complex qualitative content analysis. *ACL 2014* 44 (2014).