Human Computation Must Be Reproducible

Praveen Paritosh
Google
345 Spear St,
San Francisco, CA 94105.
pkp@google.com

ABSTRACT

Human computation is the technique of performing a computational process by outsourcing some of the difficult-to-automate steps to humans. In the social and behavioral sciences, when using humans as measuring instruments, *reproducibility* guides the design and evaluation of experiments. We argue that human computation has similar properties, and that the results of human computation must be reproducible, in the least, in order to be informative. We might additionally require the results of human computation to have high *validity* or high *utility*, but the results must be reproducible in order to measure the validity or utility to a degree better than chance. Additionally, a focus on reproducibility has implications for design of task and instructions, as well as for the communication of the results. It is humbling how often the initial understanding of the task and guidelines turns out to lack reproducibility. We suggest ensuring, measuring and communicating reproducibility of human computation tasks.

1. INTRODUCTION

Some examples of tasks using human computation are: labeling images [Nowak and Ruger, 2010], conducting user studies [Kittur, Chi, and Suh, 2008], annotating natural language corpora [Snow, O'Connor, Jurafsky and Ng, 2008], annotating images for computer vision research [Sorokin and Forsyth, 2008], search engine evaluation [Alonso, Rose and Stewart, 2008; Alonso, Kazai and Mizzaro, 2011], content moderation [Ipeirotis, Provost and Wang, 2010], entity reconciliation [Kochhar, Mazzocchi and Paritosh, 2010], conducting behavioral studies [Suri and Mason, 2010; Horton, Rand and Zeckhauser, 2010].

These tasks involve presenting a question, e.g.,"Is this image offensive?," to one or more humans, whose answers are aggregated to produce a resolution, a suggested answer for the original question. The humans might be paid contributors. Examples of paid workforces include Amazon Mechanical Turk [www.mturk.com] and oDesk [www.odesk.com]. An example of a community of volunteer contributors is Foldit [www.fold.it], where human computation augments machine computation for predicting protein structure [Cooper et al., 2010]. Another example involving volunteer contributors is games with a purpose [von Ahn, 2006], and the upcoming Duolingo [www.duolingo.com], where the contributors are translate previously untranslated web corpora while learning a new language.

The results of human computation can be characterized by accuracy [Oleson et al., 2011], information theoretic measures of quality [Ipeirotis, Provost and Wang, 2010], utility [Dai, Mausam and Weld, 2010], among others. In order for us to have confidence in any such criteria, the results must be reproducible, i.e., not a result of chance agreement or irreproducible human idiosyncrasies, but a reflection of the underlying properties of the questions and task instructions, on which others could agree as well. Reproducibility is the degree to which a process can be replicated by different human contributors working under varying conditions, at different locations, or using different but functionally equivalent measuring instruments. A total lack of reproducibility implies that the given results could have been obtained merely by chance agreement. If the results are not differentiable from chance, there is little information content in them. Using human computation in such a scenario is wasteful of an expensive resource, as chance is cheap to simulate.

A much stronger claim than reproducibility is *validity.* For a measurement instrument, e.g., a vernier caliper, a standardized test, or, a human coder, the reproducibility is the extent to which a measurement gives consistent results, and the validity is the extent to which the tool measures what it claims to measure. In contrast to reproducibility, validity concerns truths. Validity requires comparing the results of the study to evidence obtained independently of that effort. Reproducibility provides assurances that particular research results can be duplicated, that no (or only a negligible amount of) extraneous noise has entered the process and polluted the data or perturbed the research results, validity provides assurances that claims emerging from the research are borne out in fact.

We might want the results of human computation to have high validity, high utility, low cost, among other desirable characteristics. However, the results must be reproducible in order for us to measure the validity or utility to a degree better than chance.

More than a statistic, a focus on reproducibility offers valuable insights regarding the design of the task and the guidelines for the human contributors, as well as the communication of the results. The output of human computation is thus akin to the result of a scientific experiment, and it can only be considered meaningful if it is reproducible — that is, the same results could be replicated in an independent exercise. This requires clearly communicating the task instructions, and the criterion of selecting the human contributors, ensuring that they work independently, and reporting an appropriate measure of reproducibility. Much of this is well established in the methodology of content analysis in the social and behavioral sciences [Armstrong, Gosling, Weinman

CrowdSearch 2012 workshop at WWW 2012, Lyon, France

and Marteau, 1997; Hayes and Krippendorff, 2007], being required of any publishable result involving human contributors. In Section 2 and 3, we argue that human computation resembles the coding tasks of behavioral sciences. However, in the human computation and crowdsourcing research community, reproducibility is not commonly reported.

We have collected millions of human judgments regarding entities and facts in Freebase [Kochhar, Mazzocchi and Paritosh, 2010; Paritosh and Taylor, 2012]. We have found reproducibility to be a useful guide for task and guideline design. It is humbling how often the initial understanding of the task and guidelines turns out to lack reproducibility. In section 4, we describe some of the widely used measures of reproducibility. We suggest ensuring, measuring and communicating reproducibility of human computation tasks.

In the next section, we describe the sources of variability human computation, which highlight the role of reproducibility.

2. SOURCES OF VARIABILITY IN HUMAN COMPUTATION

There are many sources of variability in human computation that are not present in machine computation. Given that human computation is used to solve problems that are beyond the reach of machine computation, by definition, these problems are incompletely specified. Variability arises due to incomplete specification of the task. This is convolved with the fact that the guidelines are subject to differing interpretations by different human contributors. Some characteristics of human computation tasks are:

- *Task guidelines are incomplete:* A task can span a wide set of domains, not all of which are anticipated at the beginning of the task. This leads to incompleteness in guidelines, as well as varying levels of performance depending upon the contributor's expertise in that domain. Consider the task of establishing relevance of an arbitrary search query [Alonso, Kazai and Mizzaro, 2011].
- *Task guidelines are not precise:* Consider, for example, the task of declaring if an image is unsuitable for a social network website. Not only is it hard to write down all the factors that go into making an image offensive, it is hard to communicate those factors to human contributors with vastly different predispositions. The guidelines usually rely upon shared common sense knowledge and cultural knowledge.
- *Validity data is expensive or unavailable:* An oracle that provide the true answer for any given question is usually unavailable. Sometimes for a small subset of *gold questions*, we have answers from another independent source. This can be useful in making estimates of validity, subject to the degree that the gold questions are representative of the set of questions. These gold questions could be very useful for training and feedback to the human contributors, however, we have to be ensure their representatitiveness in order to make warranted claims regarding validity.

Each of the above might be true to a different degree for different human computation tasks. These factors are similar to the concerns of behavioral and social scientists in using humans as measuring instruments.

3. CONTENT ANALYSIS AND CODING IN THE BEHAVIORAL SCIENCES

In the social sciences, content analysis is a methodology for studying the content of communication [Berelson, 1952; Krippendorff, 2004]. Coding of subjective information is a significant source of empirical data in social and behavioral sciences, as they allow techniques of quantitative research to be applied to complex phenomena. These data are typically generated by trained human observers who record or transcribe textual, pictorial or audible matter in terms suitable for analysis. This task is called coding, which involves assigning categorical, ordinal or quantitative responses to units of communication.

An early example of a content analysis based study is "Do newspapers now give the news?" [Speed, 1893], which tried to show that the coverage of religious, scientific and literary matters was dropped in favor of gossip, sports and scandals between 1881 and 1893, by New York newspapers. Conclusions from such data can only be trusted if the reading of the textual data as well as of the research results are replicable elsewhere, that the coders demonstrably agree on what they are talking about. Hence, the coders need to demonstrate the trustworthiness of their data by measuring their reproducibility. To perform reproducibility tests, additional data are needed: by duplicating the research under various conditions. Reproducibility is established by independent agreement between different but functionally equal measuring devices, for example, by using several coders with diverse personalities. The reproducibility of coding has been used for comparing consistency of medical diagnosis [e.g., Koran, 1975], for drawing conclusions from meta-analysis of research findings [e.g., Morley et al., 1999], for testing industrial reliability [Meeker and Escobar, 1998], for establishing the usefulness of a clinical scale [Hughes et al., 1982].

3.1 Relationship between Reproducibility and Validity

- *Lack of reproducibility limits the chance of validity:* If the coding results are a product of chance, it may well include a valid account of what was observed, but researchers would not be able to identify that account to a degree better than chance. Thus, the more unreliable a procedure, the less likely it is to result in data that lead to valid conclusions.
- *Reproducibility does not guarantee validity:* Two observers of the same event who hold the same conceptual system, prejudice, or, interest may well agree on what they see but still be objectively wrong, based on some external criterion. Thus a reliable process may or may not lead to valid outcomes.

In some cases, validity data might be so hard to obtain that one has to contend with reproducibility. In tasks such as interpretation and transcription of complex textual matter, suitable accuracy standards are not easy to find. Because interpretations can only be compared to interpretations, attempts to measure validity presuppose the privileging of some interpretations over others, and this puts any claims regarding validity on epistemologically shaky grounds. In some tasks like psychiatric diagnosis, even reproducibility is hard to attain for some questions. Aboraya et al.

[2006] review the reproducibility of psychiatric diagnosis. Lack of reproducibility has been reported for judgments of schizophrenia and affective disorder [Goodman et al., 1984], calling such diagnosis into question.

3.2 Relationship with Chance Agreement

In this section, we look at some properties of chance agreement, and its relationship to reproducibility. Given two coding schemes for the same phenomenon, the one with fewer categories will have higher chance agreement. For example, in reCAPTCHA [von Ahn et al., 2008], two independent humans are shown an image containing text and asked to transcribe it. Assuming that there is no collusion, chance agreement, i.e., two different humans typing in the same word/phrase by chance is very small. However, in a task in which there are only two possible answers, e.g., true and false, the probability of chance agreement between two answers is 0.5.

If a disproportionate amount of data falls under one category, then the expected chance agreement is very high, so in order to demonstrate high reproducibility, even higher observed agreement is required [Feinstein and Cicchetti 1990; Di Eugenio and Glass 2004].

Consider a task of rating a proposition as true or false. Let p be the probability of the proposition being true. An implementation of chance agreement is the following: toss a biased coin with the same odds, i.e., p is the probability that it turns heads, and declare the proposition to be true when the coin lands heads. Now, we can simulate n judgments by tossing the coin n times. Let us look at the properties of unanimous agreement between two judgments. The likelihood of a chance agreement on true is p^2. Independent of this agreement, the probability of the proposition being true is p, therefore, the accuracy of this chance agreement on true is p^3. By design, these judgments do not contain any information other than the a priori distribution across answers. Such data has close to zero reproducibility, however, sometimes it can show up in surprising ways when looked through the lens of accuracy.

For example, consider the task of an airport agent declaring a bag as safe or unsafe for boarding on the plane. A bag can be unsafe if it contains toxic or explosive materials that could threaten the safety of the flight. Most bags are safe. Let us say that one in a thousand bags is potentially unsafe. Random coding would allow two agents to jointly assign "safe" 99.8% of the time, and since 99.9% of the bags are safe, this agreement would be accurate 99.7% of the time! This leads to the surprising result that when data are highly skewed, the coders may agree on a high proportion of items while producing annotations that are accurate, but of low reproducibility. When one category is very common, high accuracy and high agreement can also result from indiscriminate coding. The test for reproducibility in such cases is the ability to agree on the rare categories. In the airport bag classification problem, while chance accuracy on safe bags is high, chance accuracy on unsafe bags is extremely low, 10^{-7}%. In practice, the cost of errors vary: mistakenly classifying a safe bag as unsafe causes far less damage than classifying an unsafe bag as safe.

In this case, it is dangerous to consider an averaged accuracy score, as different errors do not count equal: a chance process that does not add any information has an average accuracy which is higher than 99.7%, most of which is reflecting the original bias in the distribution of safe and unsafe bags. A misguided interpretation of accuracy or a poor estimate of accuracy can be less informative than reproducibility.

3.3 Reproducibility and Experiment Design

The focus on reproducibility has implications on design of task and instruction materials. Krippendorff [2004] argues that any study using observed agreement as a measure of reproducibility must satisfy the following requirements:

- It must employ an exhaustively formulated, clear, and usable guidelines;
- It must use clearly specified criteria concerning the choice of contributors,so as others may use such criteria to reproduce the data;
- It must ensure that the contributors that generate the data used to measure reproducibility work independently of each other. Only if such independence is assured can covert consensus be ruled out the observed agreement be explained in terms of the given guidelines and the task.

The last point cannot be stressed enough. There are potential benefits from multiple contributors collaborating, but data generated in this manner neither ensure reproducibility nor reveal its extent. In groups like these, humans are known to negotiate and to yield to each other in quid pro quo exchanges, with prestigious group members dominating the outcome [see for example, Esser, 1998]. This makes the results of collaborative human computation a reflection of the social structure of the group, which is nearly impossible to communicate to other researchers and replicate. The data generated by collaborative work are akin to data generated by a single observer, while reproducibility requires at least two independent observers. To substantiate the contention that collaborative coding is superior to coding by separate individuals, a researcher would have to compare the data generated by at least two such groups and two individuals, each working independently.

A model in which the coders work independently, but consult each other when unanticipated problems arise, is also problematic. A key source of these unanticipated problems is the fact that the writers of the coding instructions did not anticipate all the possible ways of expressing the relevant matter. Ideally, these instructions should include every applicable rule on which agreement is being measured. However, discussing emerging problems could create re-interpretation of the existing instructions in ways that are a function of the group and not communicable to others. In addition, as the instructions become reinterpreted, the process loses its stability: data generated early in the process use instructions that differ from those later.

In addition to the above, Craggs and McGee Wood [2005] discourage researchers from testing their coding instructions on data from more than one domain. Given that the reproducibility of the coding instructions depends to a great extent on how complications are dealt with, and that every domain displays different complications, the sample should contain sufficient examples from all domains which have to be annotated according to the instructions.

Even the best coding instructions might not specify all possible complications. Besides the set of desired answers, the coders should also be allowed to *skip* a question. If the coders cannot prove any of the other answers is correct, they skip that question. For any other answer, the instructions define an a priori model of agreement on that answer, while skip represents the unanticipated properties of questions and coders. For instance, some questions might be too difficult for certain coders. Providing the human contributors with an option to skip is a nod to the openness of the task, and can be used to explore the poorly defined parts of the task that were not anticipated at the outset. Additionally, the skip votes can be removed from the analysis for computing reproducibility, as we do not have expectation of agreement on them [Krippendorff, 2012, personal communication].

4. MEASURING REPRODUCIBILITY

In measurement theory, *reliability* is the more general guarantee that the data obtained are independent of the measuring event, instrument or person. There are three different kinds of reliability:

Stability: measures the degree to which a process is unchanging over time. It is measured by agreement between multiple trials of the same measuring or coding process. This is also called test-retest condition, in which one observer does a task, and after some time, repeats the task again. This measures intra-observer reliability. A similar notion is *internal consistency* [Cronbach, 1951], which is the degree to which the answers on the same task are consistent. Surveys are designed so that the subsets of similar questions are known a priori, and measures for internal consistency metrics are based on correlation between these answers.

Reproducibility: measures the degree to which a process can be replicated by different analysts working under varying conditions, at different locations, or, using different but functionally equivalent measuring instruments. Reproducible data, by definition, are data that remain constant throughout variations in the measuring process [Kaplan and Goldsen, 1965].

Accuracy: measures the degree to which the process produces valid results. To measure accuracy, we have to compare the performance of contributors with the performance of a procedure that is known to be correct. In order to generate estimates of accuracy, we need accuracy data, i.e., valid answers to a representative sample of the questions. Estimating accuracy gets harder in cases where the heterogeneity of the task is poorly understood.

The next section focuses on reproducibility.

4.1 Reproducibility

There are two different aspects of reproducibility: inter-rater reliability and inter-method reliability. Inter-rater reliability focuses on the reproducibility by agreement between independent raters, and inter-method reliability focuses on the reliability of different measuring devices. For example, in survey and test design, *parallel forms reliability* is used to create multiple equivalent tests, of which more than one are administered to the same human. We focus on inter-rater reliability as the measure of reproducibility typically applicable to human computation tasks, where we generate judgments from multiple humans per question. The simplest form of inter-rater reliability is percent agreement, however it is not suitable as a measure of reproducibility as it does not correct for chance agreement.

For extensive survey of measures of reproducibility, refer to Popping [1988], Artstein and Poesio [2007]. The different coefficients of reproducibility differ in the assumptions they make about the properties of coders, judgments and units. Scott's π [1955] is applicable to two raters and assumes that the raters have the same distribution of responses, where Cohen's κ [1960; 1968] allows for a a separate distribution of chance behavior per coder. Fleiss' κ [1971] is a generalization of Scott's π for an arbitrary number of raters. All of these coefficients of reproducibility correct for chance agreement similarly. First, they find how much agreement is expected by chance: let us call this vallue A_e. The data from the coding is a measure of the observed agreement, A_o. Various inter-rater reliabilities measure the proportion of the possible agreement beyond chance that was actually observed.

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

Krippendorff's α [1970; 2004] is a generalization of many of these coefficients. It is a generalization of the above metrics for an arbitrary number of raters, not all of whom have to answer every question. Krippendorff's α has the following desirable characteristics:

- It is applicable to an arbitrary number of contributors and invariant to the permutation and selective participation of contributors. It corrects itself for varying amounts of reproducibility data.
- It constitutes a numerical scale between at least two points with sensible reproducibility interpretations, 0 representing absence of agreement, and 1 indicating perfect agreement.
- It is applicable to several scales of measurement: ordinal, nominal, interval, ratio, and more.

Alpha's general form is:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where D_o is the observed disagreement:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} . \delta^2_{ck}$$

and D_e is the disagreement one would expect when the answers are attributable to chance rather than to the properties of the questions:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c . n_k . \delta^2_{ck}$$

The δ^2_{ck} term is the distance metric for the scale of the answerspace. For a nominal scale,

$$\delta^2_{ck} = \begin{cases} 0 \text{ if } c = k \\ 1 \text{ if } c \neq k \end{cases}$$

4.2 Statistical Significance

The goal of measuring reproducibility is to ensure that the data does not deviate too much from perfect agreement, not that the data is different from chance agreement. In the definition of α, chance agreement is one of the two anchors for the agreement scale, the other, more important, reference point being that of perfect agreement. As the distribution of α is unknown, confidence intervals on α are obtained from empirical distribution generated by bootstrapping — that is, by drawing a large number of subsamples from the reproducibility data, computing α for each. This gives us a probability distribution of hypothetical α values that could occur within the constraints of the observed data. This can be used to calculate the probability of failing to reach the smallest acceptable reproducibility α_{min}, $q|\alpha < \alpha_{min}$, or a two tailed confidence interval for chosen level of significance.

4.3 Sampling Considerations

To generate an estimate of the reproducibility of a population, we need to generate a representative sample of the population. The sampling needs to ensure that we have enough units from the rare categories of questions in the data. Assuming that α is normally distributed, Bloch and Kraemer [1989] provide a suggestion for minimum number of questions from each category to be included in the sample, N_c, by,

$$N_c = z^2 \left(\frac{(1+\alpha_{min})(3-\alpha_{min})}{4p_c(1-p_c)(1-\alpha_{min})} - \alpha_{min} \right)$$

Where,

- p_c is the smallest estimated proportion values of the category c in the population,
- $alpha_{min}$ is the smallest acceptable reproducibility below which data will have to be rejected as unreliable, and
- z is the desired level of statistical significance, represented by the corresponding z value for one-tailed tests

This is a simplification, as it assumes α is normally distributed, and binary data, and does not account for the number of raters. A general description of sampling requirement is an open problem.

4.4 Acceptable Levels of Reproducibility

Fleiss [1981] and Krippendorff [2004] present guidelines for what should acceptable values of reproducibility based on surveying the empirical research using these measures. Krippendorff suggests,

- Rely on variables with reproducibility above $\alpha = 0.800$. Additionally don't accept data if the confidence interval reaches below the smallest acceptable reproducibility, $\alpha_{min} = 0.667$, or, ensure that the probability, q, of the failure to have less than smallest acceptable reproducibility $alpha_{min}$ is reasonably small, e.g., $q < 0.05$.
- Consider variables with reproducibility between $\alpha = 0.667$ and $\alpha = 0.800$ only for drawing tentative conclusions.

These are suggestions, and the choice of thresholds of acceptability depend upon the validity requirements imposed on the research results. It is perilous to "game" α by violating the requirements of reproducibility: for example, by removing a subset of data post-facto to increase α. Partitioning data by agreement measured after the experiment will not lead to valid conclusions.

4.5 Other Measures of Quality

Ipeiritos, Provost and Wang [2010] present an information theoretic *quality score*, which measures the quality of a contributor in terms of comparing their score to a spammer who is trying to advantage of chance accuracy. In that regard, it is a similar metric to Krippendorff's alpha, and additionally models a notion of cost.

$$QualityScore = 1 - \frac{ExpCost(Contributor)}{ExpCost(Spammer)}$$

Turkontrol [Dai, Mausam and Weld, 2010], uses both a model of utility and quality using decision-theoretic control to make trade-offs between quality and utility for workflow control.

Le, Edmonds, Hester and Biewald [2010] develop a gold standard based quality assurance framework that provides direct feedback to the workers and targets specific worker errors. This approach requires extensive manually generated collection of gold data. Oleson et al. [2011], further develop this approach to include *pyrite*, which are programmatically generated gold questions on which contributors are likely to make an error, for example by mutating data so that it is no longer valid. These are very useful metrics for training, feedback and protection against spammers, but these *do not* reveal the accuracy of the results. The gold questions, by design, are not representative of the original set of questions. These lead to wide error bars on the accuracy estimates, and it might be valuable to measure reproducibility of results.

5. CONCLUSIONS

We describe reproducibility as a necessary but not sufficient requirement for results of human computation. We might additionally require the results to have high validity or high utility, but our ability to measure validity or utility with confidence is limited if the data are not reproducible. Additionally, a focus on reproducibility has implications for design of task and instructions, as well as for the communication of the results. It is humbling how often the initial understanding of the task and guidelines turns out to lack reproducibility. We suggest ensuring, measuring and communicating reproducibility of human computation tasks.

6. ACKNOWLEDGMENTS

The author would like to thank Jamie Taylor, Reilly Hayes, Stefano Mazzocchi, Mike Shwe, Ben Hutchinson, Al Marks, Tamsyn Waterhouse, Peng Dai, Ozymandias Haynes, Ed Chi and Panos Ipeirotis for insightful discussions on this topic.

7. REFERENCES

[1] Aboraya, A., Rankin, E., France, C., El-Missiry, A., John, C. The Reliability of Psychiatric Diagnosis Revisited, Psychiatry (Edgmont). 2006 January; 3(1): 41-50.

[2] Alonso, O., Rose, D. E., Stewart, B.. Crowdsourcing for relevance evaluation. SIGIR Forum, 42(2):9-15, 2008.

[3] Alonso, O., Kazai, G. and Mizzaro, S., 2011, Crowdsourcing for search engine evaluation, Springer 2011.

[4] Armstrong, D., Gosling, A., Weinman, J. and Marteau, T., 1997, The place of inter-rater reliability in qualitative research: An empirical study, Sociology, vol 31, no. 3, 597-606.

[5] Artstein, R. and Poesio, M. 2007. Inter-coder agreement for computational linguistics. Computational Linguistics.

[6] Bennett, E. M., R. Alpert, and A. C. Goldstein. 1954. Communications through limited questioning. Public Opinion Quarterly, 18(3):303-308.

[7] Berelson, B., 1952. Content analysis in communication research, Free Press, New York.

[8] Bloch, Daniel A. and Helena Chmura Kraemer. 1989. 2 x 2 kappa coefficients: Measures of agreement or association. Biometrics, 45(1):269-287

[9] Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J., 2008, Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In the Proceedings of the 28th ACM SIGMOD International Conference on Management of Data, Vancouver.

[10] Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37-46.

[11] Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4):213-220.

[12] Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., and Popovic, Z. Predicting protein structures with a multiplayer online game. Nature, 466:7307, 756-760

[13] Cronbach, L. J., 1951, Coefficient alpha and the internal structure of tests. Psychometrica, 16, 297-334.

[14] Craggs, R. and McGee Wood, M. 2004. A two-dimensional annotation scheme for emotion in dialogue. In Proc.of AAAI Spring Symmposium on Exploring Attitude and Affect in Text, Stanford.

[15] Dai, P., Mausam, Weld, D. S. Decision-Theoretic Control of Crowd-Sourced Workflows. AAAI 2010.

[16] Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. Computational Linguistics, 30(1):95-101.

[17] Esser, J.K., 1998, Alive and Well after 25 Years: A Review of Groupthink Research, Organizational Behavior and Human Decision Processes, Volume 73, Issues 2-3, 116-141.

[18] Feinstein, Alvan R. and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6):543-549.

[19] Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378-382.

[20] Goodman AB, Rahav M, Popper M, Ginath Y, Pearl E. The reliability of psychiatric diagnosis in Israel's Psychiatric Case Register. Acta Psychiatr Scand. 1984 May;69(5):391-7.

[21] Hayes, A. F., and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1):77-89.

[22] Hughes CD, Berg L, Danziger L, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Then British Journal of Psychiatry 1982;140:56.

[23] Ipeirotis, P.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In KDD-HCOMP '10.

[24] Krippendorff, K. 1970. Bivariate agreement coefficients for reliability of data. Sociological Methodology, 2:139-150

[25] Krippendorff, K. 2004. Content Analysis: An Introduction to Its Methodology, Second edition. Sage, Thousand Oaks.

[26] Kochhar, S., Mazzocchi, S., and Paritosh, P., 2010, The Anatomy of a Large-Scale Human Computation Engine, In Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010, Washington D.C.

[27] Koran, L. M., 1975, The reliability of clinical methods, data and judgments (parts 1 and 2), New England Journal of Medicine, 293(13/14).

[28] Kittur, A., Chi, E. H. and Suh, B. Crowdsourcing user studies with Mechanical Turk. In Proceedings of the Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florence.

[29] Mason, W., and Siddharth. 2010. Conducting Behavioral Research on Amazon's Mechanical Turk, Working Paper, Social Science Research Network.

[30] Meeker, W. and Escobar, L., 1998, Statistical Methods for Reliability Data, Wiley.

[31] Morley, S. Eccleston, C. and Williams, A., 1999, Systematic review and meta-analysis of randomized controlled trials of cognitive behavior therapy and behavior therapy for chronic pain in adults, excluding headache, Pain, 80, 1-13.

[32] Nowak, S. and Ruger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In Multimedia Information Retrieval, pages 557-566.

[33] Oleson, D., Hester, V., Sorokin, A., Laughlin, G., Le, J., and Biewald, L. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. In HCOMP '11: Proceedings of the Third AAAI Human Computation Workshop.

[34] Paritosh, P. and Taylor, J., 2012, Freebase: Curating Knowledge at Scale, To appear in the 24th Conference on Innovative Applications of Artificial Intelligence, IAAI 2012.

[35] Popping, R., 1988, On agreement indices for nominal data, Sociometric Research: Data Collection and Scaling, 90-105.

[36] Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 19(3):321-325.

[37] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. 2008. Cheap and fast, but is it good?: evaluating non-expert annotations for natural language tasks. In EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[38] Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. In First International Workshop on Internet Vision, CVPR 08.

[39] Speed, J. G., 1893, Do newspapers now give the news, Forum, August, 1893.

[40] von Ahn, L. Games With A Purpose. IEEE Computer Magazine, June 2006. Pages 96-98

[41] von Ahn, L., and Dabbish, L. Labeling Images with a Computer Game. ACM Conference on Human Factors in Computing Systems, CHI 2004. Pages 319-326.

[42] von Ahn, L., Maurer, B., McMillen, C., Abraham, D. and Blum, M. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, September 12, 2008. pp 1465-1468.