Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit

Radu-Daniel Vatavu
University Stefan cel Mare of Suceava
Suceava 720229, Romania
vatavu@eed.usv.ro

Jacob O. Wobbrock
Information School | DUB Group
University of Washington
Seattle, WA 98195-2840 USA
wobbrock@uw.edu

ABSTRACT

We address in this work the process of agreement rate analysis for characterizing the level of consensus between participants' proposals elicited during guessability studies. Two new measures, *i.e.*, *disagreement rate* for referents and *coagreement rate* between referents, are proposed to accompany the widely-used agreement rate formula of Wobbrock *et al.* [37] when reporting participants' consensus for symbolic input. A statistical significance test for comparing the agreement rates of $k \geq 2$ referents is presented in analogy with Cochran's success/failure *Q* test [5], for which we express the test statistic in terms of agreement and coagreement rates. We deliver a toolkit to assist practitioners to compute agreement, disagreement, and coagreement rates, and run statistical tests for agreement rates at p=.05, .01, and .001 levels of significance. We validate our theoretical development of agreement rate analysis in relation with several previously published elicitation studies. For example, when we present the probability distribution function of the agreement rate measure, we also use it (1) to explain the magnitude of agreement rates previously reported in the literature, and (2) to propose qualitative interpretations for agreement rates, in analogy with Cohen's guidelines for effect sizes [6]. We also re-examine previously published elicitation data from the perspective of the agreement rate test statistic, and highlight new findings on the effect of referents over agreement rates, unattainable prior to this work. We hope that our contributions will advance the current knowledge in agreement rate analysis, providing researchers and practitioners with new techniques and tools to help them understand user-elicited data at deeper levels of detail and sophistication.

Author Keywords

Guessability study, agreement rate, methodology, statistical test, user-defined gestures, disagreement, coagreement.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*evaluation/methodology, theory and methods*.

CHI 2015, April 18–23 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04...$15.00
http://dx.doi.org/10.1145/2702123.2702223

INTRODUCTION

Understanding users' preferences for interacting with computing devices empowers designers and practitioners with valuable knowledge about the predominant and popular patterns of interaction. Participatory design has a long standing as a useful set of practices to collect such knowledge by involving users into the early stages of the design process [2,13]. The symbolic input elicitation methodology of Wobbrock *et al.* [37] for conducting guessability studies is one example of a practice that has emerged from participatory design. The elicitation methodology has especially found applications for gesture set design, for which it has been widely adopted to study various gesture acquisition technologies [20,22,26,27,39], input devices [1,14,15,29,31], application domains [8,16,23,24, 32,33,34], and user groups [18,33]. These studies reported valuable insights about participants' mental models of the interaction, and compiled design recommendations informed by the observed consensus among participants.

The level of participants' consensus has been measured and reported in the literature with the agreement rate formula introduced by Wobbrock *et al.* [37]. Agreement rates compute normalized values in the $[0..1]$ interval that are reflective of user consensus, *e.g.*, A=.625 denotes the overall agreement reached by 20 participants, out of which 15 suggested proposals of one form and 5 of another (see this example discussed in [37] (p. 1871) and eq. 1 showing the sum of square ratios formula for agreement rate). Since they were introduced, agreement rates have been adopted by the community and reported in many studies [1,8,14,15,16,18,20,22,23,24,26,27, 29,31,32,33,34,39]. However, there has been no attempt up to date to examine in detail the properties of the agreement rate measure, *e.g.*, what does its probability distribution function look like?, how likely is it to observe an agreement rate of .625?, or what is the relationship between agreement and disagreement? Also, there has been no attempt to strengthen agreement analysis reporting with statistical significance tests, *e.g.*, is there a significant statistical difference between the agreement rate values .625 and .595 computed from 20 participants? [37] (p. 1871). Given the wide adoption of the elicitation methodology, we believe it is high time to investigate such aspects in detail. Consequently, we are concerned in this work with formalizing agreement analysis by providing theoretical argumentation, new measures, and a statistical test for comparing agreement rates, for which we show their usefulness on previously published elicitation data from the literature.

The contributions of this work are as follows: (1) we introduce two new measures for evaluating *disagreement rate* for referents and *coagreement rate* between referents that accompany the widely-adopted *agreement rate* measure of Wobbrock *et al.* [37] for reporting participants' consensus for symbolic input; (2) a statistical significance test for comparing agreement rates for two or multiple referents derived in analogy with Cochran's Q test statistic [5] and following the χ^2 distribution; (3) an analysis of the probability distribution of the agreement rate measure, and qualitative interpretations for agreement rates, in analogy with Cohen's guidelines for effect sizes [6]; (4) a toolkit to compute agreement rates and report the statistical significance of the effect of referents over agreement rates at p=.05, .01, and .001 levels of significance; and (5) a re-examination of several published datasets collected during elicitation studies that shows the benefits of using statistical significance tests for comparing agreement rates. We hope that these contributions will advance the current knowledge in agreement rate analysis for user elicitation studies, and will prove useful to researchers and practitioners that are in search of techniques and tools to help them understand user-elicited data at deeper levels of detail and sophistication.

RELATED WORK

We review in this section previous work concerned with conducting elicitation studies and running agreement rate analysis, and we look at development of statistical techniques and tools in the Human-Computer Interaction community.

Elicitation studies

Wobbrock *et al.* [37] introduced a general methodology for maximizing the guessability of symbolic input, which was originally evaluated on the EdgeWrite alphabets. Say a practitioner wants to design a toolbar icon for an uncommon command in a spreadsheet program he calls "Shift." He asks 20 participants to draw an icon representing this command. The command itself is called a "referent," and the drawn icons are "proposals" for that referent. The designer can judge which proposals are equivalent and which are different. How much agreement is represented among the proposals is the purpose of the agreement rate calculation. Of course, real elicitation studies tend to be concerned with more than one referent, *e.g.*, eliciting proposals for every letter of the alphabet [37].

The guessability methodology consists in computing agreement rates defined as the sum of squares of the percents of participants preferring various symbols for various referents [37] (p. 1871). Since it was introduced, agreement rate analysis gained popularity, as it has been applied to evaluate consensus between users' gesture preferences for various application domains [3,8,16,23,24,26,27,29,31,32,33,34,39]. For example, Wobbrock *et al.* [39] evaluated user agreement for single-handed and bimanual gesture interaction on tabletops. Ruiz *et al.* [27] used agreement rates to characterize users' preferences for motion gestures on mobile devices. Vatavu [32,33] and Vatavu and Zaiti [34] applied the gesture elicitation methodology to reveal viewers' gesture preferences for controlling the TV set using the remote, freehand and whole body Kinect gestures, and fine finger gestures captured by the Leap Motion controller. Piumsomboon *et al.* [24] investigated users' preferences for hand gestures to be used for augmented reality interfaces. Obaid *et al.* [23] applied the gesture elicitation methodology to derive a set of commands to control the navigation of a robot. Buchanan *et al.* [3] explored users' preferences for manipulating 3-D objects on multi-touch screens, and Liang *et al.* [16] were interested in object manipulation at a distance. Seyed *et al.* [29] and Kurdyukova *et al.* [15] elicited gestures for multi-display environments, and Kray *et al.* [14] looked at gestures that span multiple devices.

These studies reported gesture sets for various application domains and gesture acquisition technologies, as well as qualitative data (*e.g.*, users' evaluations of ease of execution and fit-to-function of proposed gestures) and insights into users' conceptual models about gesture interaction. In some cases, these studies revealed surprising results, such as users preferring different gestures than those designed by experienced designers [22], or cultural and technical experience influences on users' gesture proposals [18,27,32,39]. In fact, Morris *et al.* [21] showed in a recent work that elicitation studies are often biased by users' experience with technology, such as Windows-like graphical user interfaces (*i.e.*, the legacy bias), and suggested ways to reduce this bias.

Alternative measures to evaluate agreement

The practice of running guessability studies also led to alternative ways to evaluate agreement between participants' elicited proposals. For example, Findlater *et al.* [8] proposed a variation for Wobbrock *et al.*'s original agreement rate measure [37] that evaluates to 0 when there is no agreement at all. Morris [20] introduced two new metrics to better capture the degree of agreement between participants for experimental designs that elicit multiple proposals for the same referent from the same participant: max-consensus (*i.e.*, the percent of participants suggesting the most popular proposal for a given referent) and consensus-distinct ratio (*i.e.*, the percent of distinct proposals for a given referent). Vatavu and Zaiti [34] used Kendall's W coefficient of concordance[1] in conjunction with agreement rates, and reported similar values for a gesture elicitation study involving TV control (*i.e.*, mean agreement rate was .200 and W=.254). Chong and Gellersen [4] defined the popularity of user-defined techniques for associating wireless devices as a function of the percent of participants suggesting the technique and the number of times it occurred (p. 1564). Their measure of popularity takes values in the unit interval, *e.g.*, 1 denotes the maximum level of popularity. Vatavu [33] defined the confidence value of a referent as the maximum percent of participants that were in agreement for that referent.

Contributions to statistical analysis in HCI research

In this work, we also describe a statistical significance test for evaluating the effect of referents on agreement rates. The significance test was derived from Cochran's Q test for categorical data evaluated in terms of the success or failure of treatments [5]. Our concern for providing tools to analyze the statistical significance of experimental data is not new in the Human-Computer Interaction field of study. In fact, HCI

[1] Kendall's coefficient of concordance [12] is a normalization of the Friedman statistic used to assess the agreement between multiple raters with a number ranging between 0 (no agreement at all) and 1 (perfect agreement).

researchers have made important statistics contributions in this direction, empowering practitioners with the right tools to analyze their data resulted from complex or unconventional experimental designs [11,17,38]. For example, Wobbrock *et al.* [38] introduced the Aligned Rank Transform to assist practitioners for detecting interaction effects in nonparametric data resulted from conducting multi-factor experiments. Kaptein *et al.* [11] pointed to nonparametric techniques for analyzing data collected with Likert scales, and they provided an online tool for analyzing 2×2 mixed subject designs. Kaptein and Robertson [10] were concerned with the HCI community adopting a thorough consideration of effect size magnitudes when analyzing experimental data. Martens [17] introduced Illmo, a software application that implements log-likelihood modeling to help practitioners analyze their data in an interactive way.

AGREEMENT, DISAGREEMENT, AND COAGREEMENT

Agreement rate

The definition of an *agreement rate* for a given referent r for which feedback has been elicited from multiple participants during a guessability study was introduced by Wobbrock *et al.* [37] (p. 1871) as the following sum of square ratios:

$$A(r) = \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 \tag{1}$$

where P is the set of all proposals for referent r, $|P|$ the size of the set, and P_i subsets of identical proposals from P.

However, Wobbrock *et al.* did not provide any justification for the specific mathematical formula chosen to define agreement rate in equation 1, other than a note referring to the capability of this formula to intuitively characterize differences in agreement between various partitions of P: "*for example, in 20 proposals for referent r, if 15/20 are of one form and 5/20 are of another, there should be higher agreement than if 15/20 are of one form, 3/20 are of another, and 2/20 are of a third. Equation 1 captures this.*" [37] (p. 1871).

In the following, we provide a mathematical argumentation for the agreement rate formula introduced by Wobbrock *et al.* [37] (eq. 1), and we show that two correcting factors need to be applied to its current definition. Inspired by the modified calculation formula of Findlater *et al.* [8] (p. 2680), we adopt the same definition for agreement rate as the number of pairs of participants in agreement with each other divided by the total number of pairs of participants that could be in agreement:

$$\mathcal{AR}(r) = \frac{\displaystyle\sum_{P_i \subseteq P} \frac{1}{2}|P_i|\,(|P_i|-1)}{\displaystyle\frac{1}{2}|P|\,(|P|-1)} \tag{2}$$

Note the different notation $\mathcal{AR}$ (Agreement Rate) that we use in eq. 2 to differentiate from Wobbrock *et al.*'s formula [37] (eq. 1).

☞ EXAMPLE. Let's assume a number of 20 participants, from which $|P|$=20 proposals were collected for a given referent r, out of which 15/20 are of one form and 5/20 of another, *i.e.*, $|P_1|$=15 and $|P_2|$=5. The number of pairs of participants in agreement with each other is $\frac{15\cdot 14}{2} + \frac{5\cdot 4}{2}$, while the total number of pairs that could have been in agreement is $\frac{20\cdot 19}{2}$. By dividing the two values, we obtain the agreement rate $\mathcal{AR}(r) = \frac{115}{190} = .605$. By comparison, the original calculation from Wobbrock *et al.* [37] would yield $\left(\frac{15}{20}\right)^2 + \left(\frac{5}{20}\right)^2 = .625$.

The definition of eq. 2 was introduced by Findlater *et al.* [8] in their touch-screen keyboards study, but the authors did not provide the connection with Wobbrock *et al.*'s initial definition of agreement rate A [37]. In the following, we fill the gap between the two papers and show how $\mathcal{AR}(r)$ is connected to $A(r)$. We also define two new measures of agreement, *i.e.*, *disagreement* and *coagreement* that we use later in the paper to introduce a statistical significance test for agreement rate and to re-examine published data from user elicitation studies.

After successive stages of simplification of eq. 2, we obtain:

$$\begin{aligned}\mathcal{AR}(r) &= \frac{1}{|P|\,(|P|-1)} \sum_{P_i \subseteq P} \left(|P_i|^2 - |P_i|\right) \\ &= \frac{1}{|P|\,(|P|-1)} \left(\sum_{P_i \subseteq P} |P_i|^2 - \sum_{P_i \subseteq P} |P_i| \right)\end{aligned}$$

and, knowing that $\sum_{P_i \subseteq P} |P_i| = |P|$, we obtain:

$$\mathcal{AR}(r) = \frac{1}{|P|\,(|P|-1)} \sum_{P_i \subseteq P} |P_i|^2 - \frac{1}{|P|-1}$$

We continue by placing $|P|^2$ at the denominator of the values $|P_i|^2$ under the sum $\sum_{P_i \subseteq P}$ in order to arrive at a formula resembling the one introduced by Wobbrock *et al.* [37] (eq. 1):

$$\boxed{\mathcal{AR}(r) = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 - \frac{1}{|P|-1}} \tag{3}$$

What we find is that eq. 3 is similar to the formula proposed by Wobbrock *et al.* [37] (eq. 1), except for two correcting factors ($\frac{|P|}{|P|-1}$ and $-\frac{1}{|P|-1}$) that depend on the number of participants or, equivalently, the number of elicited proposals $|P|$. The two correcting factors are related to the number of degrees of freedom for computing the agreement rate, *i.e.*, because the sum of all ratios $|P_i|/|P|$ equals 1, the number of observations $|P_i|/|P|$ that are free to vary is one less than the number of distinct proposals. In the following, due to the many studies that have already used $A(r)$ to report agreement between participants [1,8,14,15,16,18,20,22,23,24,26,27,29, 31,32,33,34,39], we discuss the relationship between $A(r)$ and the new definition of agreement rate $\mathcal{AR}(r)$ with the two correcting factors. The following properties characterize the differences and relationship between the two definitions:

Property #1: $\mathcal{AR}(r) \in [0..1]$**, while** $A(r) \in [1/|P|..1]$.
$\mathcal{AR}$ takes values in the entire unit interval, with 0 denoting total disagreement between participants, and 1 absolute agree-

ment. Conversely, A never equals zero, having the minimum value $1/|P|$. The justification for this was that each proposal trivially agrees with itself, so zero agreement is not possible [37] (p. 1871) (except when $|P| \to \infty$, which is a purely theoretical situation). While that may be defensible conceptually, it does not create a desirable property, which is that $\min_r \{A(r)\}$ varies with $|P|$. For example, if each of $|P|$=20 participants has a distinct proposal, $\mathcal{AR}(r)$ will evaluate at 0, and $A(r)$ at .05. If the number of participants increases to 40 and they are still all in disagreement, $\mathcal{AR}(r)$ continues to evaluate at 0, while $A(r)$=.025. From this perspective, $\mathcal{AR}$ is more consistent in reporting disagreement than A.

Property #2: $\mathcal{AR}$ conserves the relative ordering of referents delivered by A, but is less optimistic.

Equation 3 describes a linear relationship between $\mathcal{AR}$ and A and, consequently, the relative ordering of referents delivered by A is left unchanged by $\mathcal{AR}$, *i.e.*, if $A(r_1)<A(r_2)$ then also $\mathcal{AR}(r_1)<\mathcal{AR}(r_2)$ for any two referents r_1 and r_2. This property is extremely important in the context of the existing body of work reporting agreement rates [1,8,14,15,16,18,20,22,23, 24,26,27,29,31,32,33,34,39], as it conserves all the referents' ordering previously reported in the literature. However, $\mathcal{AR}$ delivers less optimistic agreement rates than A:

$$\mathcal{AR}(r) \leq A(r) \quad \text{for any referent}^2 \tag{4}$$

which can be easily verified from eq. 3 knowing that $A(r) \leq 1$, which makes $|P| \cdot A(r) - 1 \leq (|P|-1) \cdot A(r)$. For example, if 15 out of all the 20 proposals elicited for r are of one form, and 5 of another, $\mathcal{AR}(r)$ evaluates at .605, while $A(r)$ computes .625. If we multiply the number of proposals by a factor of 2, *i.e.*, 30/40 proposals of one form and 10/40 of another, $\mathcal{A}(r)$ remains .625, but $\mathcal{AR}(r)$ increases to .615 showing the effect of multiple participants. It is conceptually defensible that $\mathcal{AR}(r)$ should increase under these conditions given that the same agreement ratios over more participants means greater quantities of actual agreement was achieved amidst greater possibility for disagreement. The magnitudes of the two measures become close for large $|P|$ values as $\lim_{|P|\to\infty} \frac{\mathcal{AR}(r)}{A(r)}=1$ (knowing that $\lim_{|P|\to\infty} \frac{|P|}{|P|-1}=1$ and $\lim_{|P|\to\infty} \frac{1}{|P|-1}=0$).

Disagreement rate

Using the same reasoning path, we define the *disagreement rate* between participants for a given referent r as the number of pairs of participants that are in disagreement divided by the total number of pairs of participants that could be in disagreement (*i.e.*, the case of total disagreement, where all participants' proposals for a given referent are different):

$$\mathcal{DR}(r) = \frac{\frac{1}{2} \sum_{P_i \subseteq P} |P_i| \left(|P| - |P_i|\right)}{\frac{1}{2}|P|\left(|P|-1\right)} \tag{5}$$

and, after successive simplifications, we arrive at:

$$\boxed{\mathcal{DR}(r) = -\frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|}\right)^2 + \frac{|P|}{|P|-1}} \tag{6}$$

☞ EXAMPLE. Following the previous example, the number of pairs of participants in disagreement with each other is $\frac{1}{2}\left(15\cdot(20-15)+5\cdot(20-5)\right)$, while the total number of pairs of participants that could have been in disagreement is $\frac{20\cdot 19}{2}$. When we divide the two values, we obtain the disagreement rate for referent r as $\mathcal{DR}(r) = \frac{75}{190} = .395$. Note how the value of the disagreement rate is complementary with respect to 1.0 to the agreement rate of referent r calculated previously, *i.e.*, $\mathcal{AR}(r) = .605$.

Coagreement rate

Prior work has established agreement rates only for individual referents in isolation. However, it would be useful to know how much agreement is shared between two referents r_1 and r_2. To this end, we define the *coagreement rate* of two referents r_1 and r_2 as the number of pairs of participants that are in agreement for both r_1 and r_2 divided by the total number of pairs of participants that could have been in agreement:

$$\boxed{\mathcal{CR}(r_1, r_2) = \frac{\sum_{i=1}^{n} \delta_{i,1} \cdot \delta_{i,2}}{n}}, n = \frac{1}{2}|P|\left(|P|-1\right) \tag{7}$$

where $\delta_{i,1}$ takes the value of 1 if the i-th pair of participants are in agreement for referent r_1 and 0 otherwise[3], and the same applies to $\delta_{i,2}$ and referent r_2. For notation convenience, we use the variable n to denote the number of pairs of participants. Table 1 shows in a tabular form the agreement indicators $\delta_{i,1}$ and $\delta_{i,2}$ for referents r_1 and r_2 for all pairs of participants.

☞ EXAMPLE. Let's assume that referent r_1 received 3 proposals of one form (♣) and 2 proposals of another (♦) from $|P|$=5 participants, while referent r_2 received 3 proposals of one form (♠), 1 proposal of another (♥), and 1 proposal of yet another form (✿), as shown in Table 2, left. The number of pairs of participants for which we evaluate agreement is $\frac{5\cdot 4}{2} = 10$, see Table 2, right. The agreement rates of the two referents are $\mathcal{AR}(r_1) = 4/10 = .400$ and $\mathcal{AR}(r_2) = 3/10 = .300$, while the coagreement between the two referents is $\mathcal{CR}(r_1, r_2) = 1/10 = .100$.

The coagreement rate can be generalized to $k>2$ referents:

$$\boxed{\mathcal{CR}(r_1, r_2, ..., r_k) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} \delta_{i,j}}, \; n = \frac{1}{2}|P|\left(|P|-1\right) \tag{8}$$

to characterize the degree to which pairs of participants are simultaneously in agreement for referents $\{r_1, r_2, ..., r_k\}$, $2 \leq k \leq |P|$. We refer to this measure as the *k-coagreement rate*.

A SIGNIFICANCE TEST FOR AGREEMENT RATES

We derive in this section a statistical significance test for comparing two or multiple agreement rates $\mathcal{AR}$ calculated from

[2] With equality occuring for perfect agreement, *i.e.*, $\mathcal{AR}(r)$=$A(r)$=1.

[3] We adopt in this section and adapt to our problem the definition of Kronecker's delta, $\delta_{i,j}$ [9] (p. 240), which is a function of two variables i and j returning 1 if the two variables are equal and 0 otherwise, *i.e.*, $\delta_{i,j} = [i = j]$.

Observation		Referent		Deltas
No.	Pair	r_1	r_2	$\delta_{i,1} \cdot \delta_{i,2}$
1	$(1, 2)$	$\delta_{1,1}$	$\delta_{1,2}$	$\delta_{1,1} \cdot \delta_{1,2}$
2	$(1, 3)$	$\delta_{2,1}$	$\delta_{2,2}$	$\delta_{2,1} \cdot \delta_{2,2}$
...	...	...	...	...
n	$(\|P\| - 1, \|P\|)$	$\delta_{n,1}$	$\delta_{n,2}$	$\delta_{n,1} \cdot \delta_{n,2}$
		$n \cdot \mathcal{AR}(r_1)$	$n \cdot \mathcal{AR}(r_2)$	$n \cdot \mathcal{CR}(r_1, r_2)$

Table 1: Agreement between participants for referents r_1 and r_2 expressed using the δ notation. Note: $\delta_{i,1}, \delta_{i,2} \in \{0, 1\}$, 1 indicates agreement and 0 disagreement; $n = \frac{1}{2}|P|\,(|P| - 1)$.

Participant No.	r_1	r_2
1	♣	♥
2	♦	✿
3	♣	♠
4	♣	♠
5	♦	♠

Observation		Referent		Deltas
No.	Pair	r_1	r_2	$\delta_{i,1} \cdot \delta_{i,2}$
1	$(1, 2)$	0	0	0
2	$(1, 3)$	1	0	0
3	$(1, 4)$	1	0	0
4	$(1, 5)$	0	0	0
5	$(2, 3)$	0	0	0
6	$(2, 4)$	0	0	0
7	$(2, 5)$	1	0	0
8	$(3, 4)$	1	1	1
9	$(3, 5)$	0	1	0
10	$(4, 5)$	0	1	0
	Total	**4**	**3**	**1**

Table 2: Agreement and coagreement rate calculation example for proposals elicited from $|P|$=5 participants for two referents r_1 and r_2 (left table). The agreements for each of the 10 pairs of participants are shown in the right table as $0/1$ values.

proposals elicited from the same participants, *i.e.*, repeated measures design. Toward this goal, we formulate our problem in terms of Cochran's Q test for detecting the significance of differences between $k \geq 2$ treatments [5]. Cochran's Q is a nonparametric test for analyzing experimental designs for which the outcome is a binary variable indicating the success or failure of an observation under a given treatment. For our problem, each referent $r_{i(=1..k)}$ represents a distinct treatment, for which we evaluate its success as whether agreement between participants was reached or not. Therefore, we have $n = \frac{1}{2}|P|(|P| - 1)$ observations resulting from $|P|$ subjects having participated in the guessability study that we arrange in a tabular form following Cochran's method [5] (p. 257), see Table 3.

The null and alternative hypotheses for agreement rates are:

H$_0$: All referents have equal agreement rates.
H$_a$: There is a difference among the agreement rates of the $k \geq 2$ referents.

With the notations employed in Table 3, the statistic employed by Cochran's Q test [5] (p. 266) is:

$$k(k-1)\frac{\sum_{j=1}^{k}\left(T_j - \frac{T}{k}\right)^2}{\sum_{i=1}^{n} R_i\,(k - R_i)}$$

which we successively adapt to the specifics of our problem by expressing it in terms of agreement and coagreement rates.

Observation		Referent				
No.	Pair	r_1	r_2	...	r_k	Total
1	$(1, 1)$	$\delta_{1,1}$	$\delta_{1,2}$	...	$\delta_{1,k}$	R_1
2	$(1, 2)$	$\delta_{2,1}$	$\delta_{2,2}$	...	$\delta_{2,k}$	R_2
...	...	...	...	...	...	...
n	$(n-1, n)$	$\delta_{n,1}$	$\delta_{n,2}$	...	$\delta_{n,k}$	R_n
	Total	T_1	T_2	...	T_k	T

Table 3: Agreement between pairs of participants for $k \geq 2$ referents arranged in a tabular form. Note: $\delta_{i,j}$ take values 1 and 0, encoding whether the i-th pair is in agreement for referent r_j or not. T_j represents the sum of column j, R_i the sum of row i, and T the grand total of $\delta_{i,j}$; $n = \frac{1}{2}|P|\,(|P| - 1)$.

The result is the V_{rd} statistic[4] (see Appendix A for the mathematical details of the calculation procedure):

$$\boxed{V_{rd} = (k-1)n \cdot \frac{\displaystyle\sum_{j=1}^{k}\mathcal{AR}^2(r_j) - \frac{1}{k}\left(\sum_{j=1}^{k}\mathcal{AR}(r_j)\right)^2}{\displaystyle\sum_{j=1}^{k}\mathcal{AR}(r_j) - \frac{1}{k}\sum_{t=1}^{k}\sum_{s=1}^{k}\mathcal{CR}(r_t, r_s)}} \tag{9}$$

For the case of comparing two agreement rates only (*i.e.*, $k = 2$), this formula simplifies to[5]:

$$\boxed{V_{rd} = n \cdot \frac{(\mathcal{AR}(r_1) - \mathcal{AR}(r_2))^2}{\mathcal{AR}(r_1) + \mathcal{AR}(r_2) - 2 \cdot \mathcal{CR}(r_1, r_2)}} \tag{10}$$

where $n = \frac{|P|(|P|-1)}{2}$ is the number of pairs of participants.

Cochran showed that the limiting distribution of the Q statistic (under the assumption that the probability of success is the same in all samples) is χ^2 with $k - 1$ degrees of freedom [5] (p. 259). Therefore, the null hypothesis H_0 is rejected at the p level if the statistic is larger than the $1-p$ quantile of χ^2:

$$\text{Reject H}_0 \text{ if } \; V_{rd} > \chi^2_{1-p,k-1} \tag{11}$$

For example, the critical values for 1 degrees of freedom (*i.e.*, two referents) are 3.84 and 6.63 for p=.05 and p=.01, respectively, and 7.81 and 11.34 for 3 degrees of freedom (*i.e.*, four referents). For convenience, Appendix B lists the critical values of the χ^2 distribution at the p=.05, p=.01, and p=.001 levels of significance up to 48 degrees of freedom. In the case in which the null hypothesis is rejected for $k>2$ referents, post-hoc tests can run using equation 10 but, in this case, the Bonferroni correction needs to be applied [36] (p. 261).

Testing for significance against zero

Our statistical test can also be employed to assess whether $\mathcal{AR}(r)$ is significantly greater than 0. To do that, we compare participants' agreement for referent r versus the case of absolute disagreement, which corresponds to a virtual referent $r^\star$, for which all participants would provide different proposals and, therefore, $\mathcal{AR}(r^\star)$=0. We then use the values of $\mathcal{AR}(r)$

[4] Notation V in V_{rd} stands for the *variation* between agreement rates, and the subscript rd denotes a *repeated* measures *design*.

[5] For k=2 conditions, it is customary to use McNemar's test [19]. However, McNemar's test and Cochran's Q for k=2 are equivalent [30] (p. 876).

and $\mathcal{AR}(r^\star)$ under equation 10, which simplifies to:

$$\boxed{V_{rd}^\star = \frac{|P| \cdot (|P|-1)}{2} \cdot \mathcal{AR}(r)} \qquad (12)$$

and decide whether to reject or not the null hypothesis (eq. 11).

☞ EXAMPLE. Let's assume referent r_1 received four distinct proposals from $|P|$=12 participants with frequencies $\{4, 4, 3, 1\}$; referent r_2 received two distinct proposals, $\{10, 2\}$; and referent r_3 received three distinct proposals, $\{5, 5, 2\}$. The agreement rates for the three referents are: $\mathcal{AR}(r_1)$=.227, $\mathcal{AR}(r_2)$=.697, and $\mathcal{AR}(r_3)$=.318. Coagreement rates are: $\mathcal{CR}(r_1, r_2)$=.152, $\mathcal{CR}(r_1, r_3)$=.045, and $\mathcal{CR}(r_2, r_3)$=.197. The V_{rd} statistic (eq. 9) is 28.964, which is significant at the p=.001 level, as indicated by the critical value for the χ^2 distribution with 3−1=2 degrees of freedom (see Appendix B). If we want to further test whether the agreement rates of pairs of referents (r_1, r_2) and (r_2, r_3) are significantly different at p=.05, we compute the statistic for these pairs (either with equation 9 or 10). The values of the V_{rd} statistic are 23.515 and 15.266 respectively, both significant at p=.001, which is below the Bonferroni corrected value of p=.05/2=.025. Furthermore, $\mathcal{AR}(r_1)$=.227 is significantly greater than 0 at p=.001 as $V_{rd}^\star$=14.98, which is above the critical value of 10.83 of the χ^2 distribution with 1 degree of freedom.

TOOLKIT FOR COMPUTING STATISTICAL SIGNIFICANCE TESTS FOR AGREEMENT RATES

To make computation of agreement rates and p values easy, we provide the AGATE tool (AGreement Analysis Toolkit), see Figure 1. The toolkit reads data organized in a matrix format so that each referent occupies one column, and each participant occupies one row. AGATE computes agreement, disagreement, and coagreement rates for selected referents, and reports significant effects of selected referents over agreement rates at p = .05, .01, and .001 levels of significance. The tool was implemented in C# using the .NET 4.5 framework, and is freely available to use and download at **`http://depts.washington.edu/aimgroup/proj/dollar/agate.html`**.

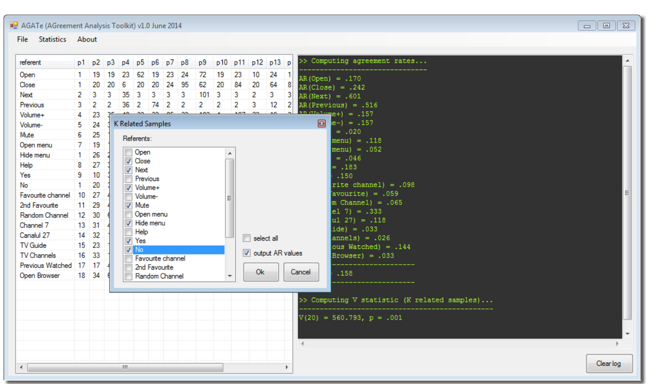

Figure 1: The AGreement Analysis Toolkit (AGATe) computes agreement measures and statistical tests for agreement rates.

CASE STUDIES

In this section, we briefly re-examine previously published data from several elicitation studies [1,24,25,34] from the perspective of our new measures. Our purpose is to show the utility of these measures and statistical test for characterizing user-elicited data in more depth. We do not attempt to be all-encompassing in our analysis, but instead our goal is to show how our measures can be employed on actual data. To this end, we pick a specific finding reported by the authors of each study, on which we then elaborate with our new measures.

Bailly *et al.* [1] (CHI '13)

In this study, 20 participants proposed gestures for 42 referents for the Metamorphé keyboard (A=.409, $\mathcal{AR}$=.336). Using our statistical test, we found an overall significant effect of referent type on agreement rate ($V_{rd(41,N=840)}$=1466.818, p<.001). Moreover, targeted statistical testing revealed more findings about users' agreement. For example, Bailly *et al.* [1] reported that *"highly directional commands (e.g., Align Left) tended to have a high gesture agreement"* (p. 567). Indeed, they did (average $\mathcal{AR}$=.809), but we also detected a significant effect of the direction of alignment (*i.e.*, left, right, bottom, and top) on the resulting agreement ($V_{rd(3,N=80)}$=121.737, p=.001), no significant difference between *Align Left* and *Align Right* (both .900), and significantly higher agreement for *Align Bottom* than for *Align Top* (.805 versus .632). To understand more, we ran coagreement analysis. The coagreement rate between *Align Left* and *Align Right* was .900, showing that all the pairs of participants that were in agreement for *Align Left* were also in agreement for *Align Right*. The coagreement between *Align Top* and *Align Bottom* was .632, indicating that all the pairs of participants in agreement for *Align Top* ($\mathcal{AR}$=.632) were also in agreement for *Align Bottom* ($\mathcal{AR}$=.900), but there were also pairs of participants that agreed on *Align Bottom* and not on *Align Top*. The k-coagreement for all the four referents was $\mathcal{CR}$=.632, showing that all participants in agreement for *Align Top* were also in agreement for the other three referents, but also that only 70% of all pairs that were in agreement for instance for *Align Left* and *Align Right* were also in agreement for *Align Bottom* and *Align Top*. Informed by these findings, the designer can now take a second, informed look at participants' proposals to understand what made the same participants agree on *Align Bottom*, but disagree on *Align Top*, for example.

Piumsomboon *et al.* [24,25] (CHI '13 and INTERACT '13)

In these studies, 20 participants proposed freehand gestures for 40 referents related to interacting with augmented reality (A=.446, $\mathcal{AR}$=.417). Using the V_{rd} test statistic, we found an overall significant effect of referent type on agreement rates ($V_{rd(39,N=800)}$=3523.962, p<.001). There were 8 referents that received high (*i.e.*, .900 and 1.000) agreement rates, and we found a significant effect of referent type over agreement rate for this subset as well ($V_{rd(7,N=160)}$=106.176, p<.001). There were 10 referents that received agreement rates below .100 ($V_{rd(9,N=200)}$=11.033, $n.s.$). Using our additional measures, we can elaborate more on some of the authors' findings. For example, the authors noted that *"we defined similar gestures as gestures that were identical or having consistent directionality although the gesture had been performed with* *different static hand poses. For example, in the Previous and Next tasks, participants used an open hand, an index finger, or two fingers to swipe from left to right or viceversa"* [24] (p. 958). This decision is a reasonable one, but we can now use coagreement rates to find out whether it was the same participants that used hand poses consistently or whether participants also varied their hand poses with the referent type. This investigation is important, because the authors also noted in a follow-up paper that *"variants of a single hand pose were often used across multiple participants, and sometimes even by a single participant"* [25] (p. 296). We found that coagreement equaled the agreement of the two referents ($\mathcal{CR}$=.489, $\mathcal{AR}$=.489 for *Previous* and *Next*) when we considered different hand poses as different gestures, which means that the same participants that were in agreement for *Previous* were also in agreement for *Next* and, even more, they kept their hand pose preference across the two referents. However, we found less consistency for other referents. For example, agreement rates for *Rotate-X-axis*, *Rotate-Y-axis*, and *Rotate-Z-axis* were .247, .263, and .258, while coagreements were less ($\mathcal{CR}(X,Y)$=.179, $\mathcal{CR}(X,Z)$=.153, $\mathcal{CR}(Y,Z)$=.174), showing that not all pairs of participants that agreed on rotating on the X axis necessarily agreed on the other axes as well. In fact, the coagreement rate for all three referents was .126, showing that only 70% of all pairs in agreement for *rotate-X-axis* and *rotate-Y-axis* also agreed on *rotate-Z-axis*. These results can inform further investigation into what made participants change their proposals for these referents.

Vatavu and Zaiti [32] (TVX '14)

In this study, 18 participants proposed freehand gestures to control 21 functions on Smart TVs. The authors found low agreement among participants (A=.200 and $\mathcal{AR}$=.170), explained by the many degrees of freedom of the human hand. Using our tool, we found a significant effect of referent type on agreement rate ($V_{rd(20,N=378)}$=560.793, p<.001). Vatavu and Zaiti [34] reported that *"when encountering referents with opposite effects (e.g., Next and Previous channel, Volume up and Volume down), most participants considered gestures should also be similar."* Our post-hoc tests revealed interesting findings for dichotomous referents. For example, the highest agreement rates were obtained for *Go to Next Channel* and *Go to Previous Channel* (.601 and .516), for which participants proposed hand movements to left and right, but we found a significant difference between the two ($V_{rd(1,N=36)}$=4.568, p<.05). Coagreement analysis showed that not all participants that were in agreement for *Next* were also in agreement for *Previous* ($\mathcal{CR}$=.436). When analyzing the other dichotomous referents, we found more agreement for *Open Menu* than for *Hide Menu* (.118 versus .052, $V_{rd(1,N=36)}$=4.454, p<.05), and nonsignificant differences between the agreement rates for *Volume Up* and *Volume Down* (.157 and .157, $\mathcal{CR}$=.157, showing that all the participants that agreed on *Volume Up* also agreed on *Volume Down*), and *Yes* and *No* (.183 and .150, with low coagreement $\mathcal{CR}$=.046, showing that participants that were in agreement for *Yes* were not also the ones that were in agreement for *No*). Overall, there were eight referents with agreement rates below .100, for which we did not detect significant differences ($V_{rd(7,N=144)}$=7.248, $n.s.$), suggesting the same low level of consensus for these referents.

DISCUSSION

In this section, we compare the agreement rate $\mathcal{AR}$ formula with Wobbrock *et al.*'s original A measure [37]. We also present the probability distribution function of $\mathcal{AR}$, and discuss the connection between agreement and disagreement.

The probability distribution function of $\mathcal{AR}$

Figure 2 shows the probability distribution function of $\mathcal{AR}$ that we generated by enumerating all the partitions of the integer $|P|$, which are distinct ways to write $|P|$ as a sum of positive integers, for which the order of the summands is not important [28]. For example, there are 11 distinct ways to write $|P| = 6$ as a sum of positive integers or, equivalently, the ratio $6/6$ as a sum of ratios for which the denominator is 6; see Table 4. We computed the associated agreement rates of these partitions that we binned into 100 equal intervals of $[0..1]$, and counted their frequencies (*e.g.*, the value .200 appears with frequency 2 in Table 4). The result was a discrete version of the probability function of $\mathcal{AR}$.

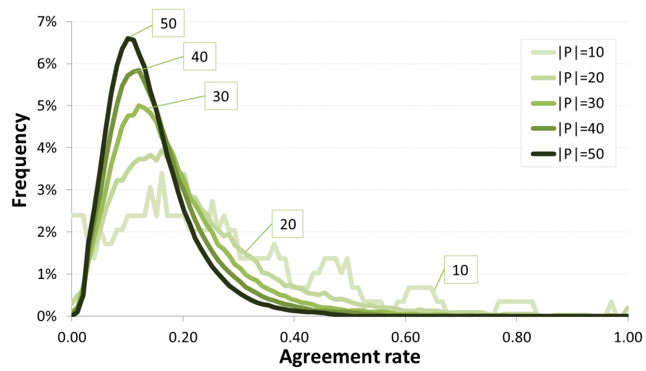

Figure 2: Probability distribution functions of $\mathcal{AR}$ computed[4] for various numbers of participants $|P|$ from 10 to 50.

When we analyze the distribution shown in Figure 2, we find that the cumulative probability of 90% is reached for $\mathcal{AR}\leq$.374, while a cumulative 99% is reached for $\mathcal{AR}\leq$.636 and $|P|$=20 participants. As the number of participants increases, there is a shift in the peak of the probability distribution toward lower values, *e.g.*, 90% cumulative probability is reached for $\mathcal{AR}\leq$.222 and 99% for $\mathcal{AR}\leq$.424 for $|P|$=50 participants. These values may seem low, but remember that we assumed each partition equally probable when we generated the probability distribution function. In the practice of guessability studies, this assumption may not hold for all referents, because some of the referents may trigger the same response from multiple participants simply due to participants' shared experience in a given field, *i.e.*, the legacy bias [21]. However, these probability distributions reflect very well current findings in the literature. For example, the average agreement rate A reported by Wobbrock *et al.* [39] for single-handed tabletop gestures is .320 (the corrected $\mathcal{AR}$ for 20 participants is .284);

[5]To compensate for the low resolution obtained for the probability distributions when bining frequencies into 100 bins at small $|P|$ values (*e.g.*, there are only 42 distinct possibilities to write $|P| = 10$ as a sum of positive integers, and 627 possibilities for $|P| = 20$), all resulted frequencies were smoothed with a central moving average using a window of size 7.

No.	Partition	$\mathcal{AR}(r)$
1	$\frac{6}{6} = \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}$	.000
2	$\frac{6}{6} = \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{2}{6}$	.067
3	$\frac{6}{6} = \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{3}{6}$	.200
4	$\frac{6}{6} = \frac{1}{6}+\frac{1}{6}+\frac{2}{6}+\frac{2}{6}$	.133
5	$\frac{6}{6} = \frac{1}{6}+\frac{1}{6}+\frac{4}{6}$	.400
6	$\frac{6}{6} = \frac{1}{6}+\frac{2}{6}+\frac{3}{6}$	.267
7	$\frac{6}{6} = \frac{1}{6}+\frac{5}{6}$	.667
8	$\frac{6}{6} = \frac{2}{6}+\frac{2}{6}+\frac{2}{6}$	.200
9	$\frac{6}{6} = \frac{2}{6}+\frac{4}{6}$	.467
10	$\frac{6}{6} = \frac{3}{6}+\frac{3}{6}$	.400
11	$\frac{6}{6} = \frac{6}{6}$	1.000

Table 4: All the 11 distinct partitions of the fraction $6/6$ into sums of fractions of positive integers with denominator 6, and their associated agreement rate values.

the average rate of Ruiz *et al.* [27] for motion gestures is .260 (corrected $\mathcal{AR}$ for 20 participants is .221). Table 5 shows more average agreement rate values, all below .450.

Study	Condition	$\lvert P\rvert$	A	$\mathcal{AR}$
Bailly *et al.* [1]	Metamorphé	20	.406†	.336
Buchanan *et al.* [3]	3-D objects	14	.468†	.430
Liang *et al.* [16]	3-D objects	12	.182†	.108
Obaid *et al.* [23]	robot navigation	35	.230†	.207
Piumsomboon *et al.* [24]	augmented reality	20	.446†	.417
Pyryeskin *et al.* [26]	above the surface	16	.230	.179
Ruiz *et al.* [27]	mobile devices	20	.260†	.221
Seyed *et al.* [29]	multi-displays	17	.160	.108
Valdes *et al.* [31]	horizontal surface	19	.244	.202
Valdes *et al.* [31]	vertical surface	19	.254	.213
Vatavu [32]	TV (Kinect)	12	.415	.362
Vatavu [33]	TV (Wii)	20	.430	.400
Vatavu [33]	TV (Kinect)	20	.330	.295
Vatavu and Zaiti [34]	TV (Leap Motion)	18	.200	.170
Weigel *et al.* [35]	skin as input	22	.250†	.214
Wobbrock *et al.* [37]	EdgeWrite alphabet	20	.349	.315
Wobbrock *et al.* [39]	tabletop (1-hand)	20	.320	.284
Wobbrock *et al.* [39]	tabletop (2-hands)	20	.280	.242

† The authors did not report the average agreement rate, so we calculated (or approximated) it from the individual agreement rates reported in these papers.

Table 5: Average agreement rates A reported in the literature and corrected $\mathcal{AR}$ values. Note how the average $\mathcal{AR}$ values of these studies are less than .450 (compare with Figure 2).

Relationship between agreement and disagreement rates

It is interesting to see how agreement rate $\mathcal{AR}(r)$ compares with the disagreement rate $\mathcal{DR}(r)$ for a given referent r, knowing that the two are complementary with respect to 1, *i.e.*, $\mathcal{AR}(r)+\mathcal{DR}(r) = 1$. This equation tells us that there is more agreement than disagreement for referent r if $\mathcal{AR}(r) > .500$. Figure 2 informs us that the probability of obtaining an agreement rate of this magnitude is less than 1% (under the hypothesis of equal chance partitions, see above) and, consequently, for most referents, participants are more likely to be in disagreement than in agreement (see Table 5). Another way to visualize the relationship between agreement and disagreement is to compute their ratio:

$$\frac{\mathcal{AR}(r)}{\mathcal{DR}(r)} = \frac{\mathcal{AR}(r)}{1-\mathcal{AR}(r)} \tag{13}$$

that takes values between 0 (*i.e.*, no agreement) and ∞ (absolute agreement). Figure 3 shows the values of this ratio. Informed by these results, the average agreement rates reported in the literature [3,16,23,24,26,27,29,31,32,33,39], and inspired by Cohen's guidelines for effect sizes [7], we propose qualitative interpretations for agreement rates, see Table 6.

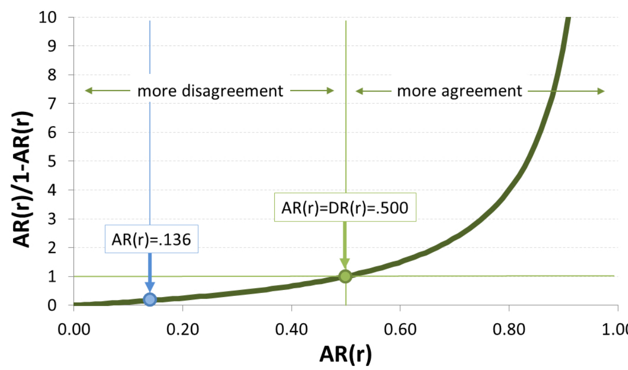

Figure 3: Relationship between agreement and disagreement rates for any referent r. Note the *theoretical* mid-point of .500 for which agreement and disagreement rates are equal, as well as the *expected* value of .136 for $\mathcal{AR}(r)$ (computed as the average of all possible agreement rate values, weighted by their probability of occurrence, according to Figure 2).

$\mathcal{AR}(r)$ interval	Probability †	Interpretation
$\leq .100$	22.9%	*low* agreement
$.100 - .300$	59.1%	*medium* agreement
$.300 - .500$	14.1%	*high* agreement
$> .500$	3.9%	*very high* agreement

† According to the probability distribution functions shown in Figure 2 and $\lvert P\rvert = 20$ participants.

Table 6: Margins for interpreting the magnitude of agreement.

CONCLUSION

We introduced in this paper new measures, a statistical test, and a companion toolkit to assist researchers and practitioners with agreement rate analysis of user-elicited data collected during guessability studies. We showed the benefits of our measures and toolkit by re-examining some published data in the literature. Further work will address new useful aspects for reporting agreement rates, such as confidence intervals, and new ways to distill agreement and coagreement into a single measure to facilitate analysis of users' consensus. We hope the contributions of this work will provide researchers and practitioners with a solid foundation for analyzing and interpreting agreement rate data and, consequently, will lead to improved user interface designs informed by more careful and in-depth examination of user-elicited data.

ACKNOWLEDGMENTS

The authors would like to thank Gilles Bailly and Thammathip Piumsomboon as well as their co-authors from [1,24,25] for kindly providing access to their gesture elicitation data. This research was conducted under the project Mobile@Old, ref. 315/2014 (PN-II-PT-PCCA-2013-4-2241), financed by MEN-UEFISCDI, Romania.

REFERENCES

1. Bailly, G., Pietrzak, T., Deber, J., and Wigdor, D. J. Métamorphe: Augmenting hotkey usage with actuated keys. In *Proc. of CHI '13*, ACM (2013), 563–572.
2. Bergvall-Kåreborn, B., and Ståhlbrost, A. Participatory design: One step back or two steps forward? In *Proc. of PDC '08*, Indiana University (2008), 102–111.
3. Buchanan, S., Floyd, B., Holderness, W., and LaViola, J. J. Towards user-defined multi-touch gestures for 3D objects. In *Proc. of ITS '13*, ACM (2013), 231–240.
4. Chong, M. K., and Gellersen, H. W. How groups of users associate wireless devices. In *Proc. of CHI '13*, ACM (2013), 1559–1568.
5. Cochran, W. G. The comparison of percentages in matched samples. *Biometrika 37*, 3/4 (1950), 256–266.
6. Cohen, J. A coefficient of agreement for nominal scales. 37–46.
7. Cohen, J. A power primer. *Psychological Bulletin 112*, 1 (1992), 155–159.
8. Findlater, L., Lee, B., and Wobbrock, J. Beyond QWERTY: Augmenting touch screen keyboards with multi-touch gestures for non-alphanumeric input. In *Proc. of CHI '12*, ACM (2012), 2679–2682.
9. James, R. C. *The Mathematics Dictionary, 5th Ed.* Chapman & Hall, New York, 1992.
10. Kaptein, M., and Robertson, J. Rethinking statistical analysis methods for CHI. In *Proc. of CHI '12*, ACM (2012), 1105–1114.
11. Kaptein, M. C., Nass, C., and Markopoulos, P. Powerful and consistent analysis of likert-type ratingscales. In *Proc. of CHI '10*, ACM (2010), 2391–2394.
12. Kendall, M. G., and Babington Smith, B. The problem of m rankings. *Annals of Math. Stats. 10*, 3 (1939), 275–287.
13. Kensing, F., and Blomberg, J. Participatory design: Issues and concerns. *Computer Supported Cooperative Work 7*, 3-4 (Jan. 1998), 167–185.
14. Kray, C., Nesbitt, D., Dawson, J., and Rohs, M. User-defined gestures for connecting mobile phones, public displays, and tabletops. In *Proc. of MobileHCI '10*, ACM (2010), 239–248.
15. Kurdyukova, E., Redlin, M., and André, E. Studying user-defined ipad gestures for interaction in multi-display environment. In *Proc. of UI '12*, ACM (2012), 93–96.
16. Liang, H.-N., Williams, C., Semegen, M., Stuerzlinger, W., and Irani, P. User-defined surface+motion gestures for 3D manipulation of objects at a distance through a mobile device. In *Proc. of APCHI '12*, ACM (2012), 299–308.
17. Martens, J.-B. Interactive statistics with illmo. *ACM Trans. Interact. Intell. Syst. 4*, 1 (Apr. 2014), 4:1–4:28.
18. Mauney, D., Howarth, J., Wirtanen, A., and Capra, M. Cultural similarities and differences in user-defined gestures for touchscreen user interfaces. In *Proc. of CHI EA '10*, ACM (2010), 4015–4020.
19. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika 12*, 2 (June 1947), 153–157.
20. Morris, M. R. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proc. of ITS '12*, ACM (2012), 95–104.
21. Morris, M. R., Danielescu, A., Drucker, S., Fisher, D., Lee, B., schraefel, m. c., and Wobbrock, J. O. Reducing legacy bias in gesture elicitation studies. *Interactions 21*, 3 (May 2014), 40–45.
22. Morris, M. R., Wobbrock, J. O., and Wilson, A. D. Understanding users' preferences for surface gestures. In *Proc. of GI '10*, Canadian Inf. Proc.Soc. (2010), 261–268.
23. Obaid, M., Häring, M., Kistler, F., Bühling, R., and André, E. User-defined body gestures for navigational control of a humanoid robot. In *Proc. of ICSR'12*, Springer-Verlag (2012), 367–377.
24. Piumsomboon, T., Clark, A., Billinghurst, M., and Cockburn, A. User-defined gestures for augmented reality. In *Proc. of CHI EA '13*, ACM (2013), 955–960.
25. Piumsomboon, T., Clark, A., Billinghurst, M., and Cockburn, A. User-defined gestures for augmented reality. In *Proc. of INTERACT '13*, Springer (2013), 282–299.
26. Pyryeskin, D., Hancock, M., and Hoey, J. Comparing elicited gestures to designer-created gestures for selection above a multitouch surface. In *Proc. of ITS '12*, ACM (2012), 1–10.
27. Ruiz, J., Li, Y., and Lank, E. User-defined motion gestures for mobile interaction. In *Proc. of CHI '11*, ACM (2011), 197–206.
28. Sedgewick, R. Permutation generation methods. *ACM Computing Surveys 9*, 2 (June 1977), 137–164.
29. Seyed, T., Burns, C., Costa Sousa, M., Maurer, F., and Tang, A. Eliciting usable gestures for multi-display environments. In *Proc. of ITS '12*, ACM (2012), 41–50.
30. Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press, 2004.
31. Valdes, C., Eastman, D., Grote, C., Thatte, S., Shaer, O., Mazalek, A., Ullmer, B., and Konkel, M. K. Exploring the design space of gestural interaction with active tokens through user-defined gestures. In *Proc. of CHI '14*, ACM (2014), 4107–4116.
32. Vatavu, R.-D. User-defined gestures for free-hand TV control. In *Proc. of EuroiTV '12*, ACM (2012), 45–48.
33. Vatavu, R.-D. A comparative study of user-defined handheld vs. freehand gestures for home entertainment environments. *Journal of Ambient Intelligence and Smart Environments 5*, 2 (2013), 187–211.
34. Vatavu, R.-D., and Zaiti, I.-A. Leap gestures for TV: Insights from an elicitation study. In *Proc. of TVX '14*, ACM (2014), 131–138.
35. Weigel, M., Mehta, V., and Steimle, J. More than touch: Understanding how people use skin as an input surface for mobile computing. In *Proc. of CHI '14*, ACM (2014), 179–188.

df	p=.05	p=.01	p=.001	df	p=.05	p=.01	p=.001	df	p=.05	p=.01	p=.001	df	p=.05	p=.01	p=.001
1	3.84	6.64	10.83	13	22.36	27.69	34.53	25	37.65	44.31	52.62	37	52.19	59.89	69.35
2	5.99	9.21	13.82	14	23.69	29.14	36.12	26	38.89	45.64	54.05	38	53.38	61.16	70.70
3	7.82	11.35	16.27	15	25.00	30.58	37.70	27	40.11	46.96	55.48	39	54.57	62.43	72.06
4	9.49	13.28	18.47	16	26.30	32.00	39.25	28	41.34	48.28	56.89	40	55.76	63.69	73.40
5	11.07	15.09	20.52	17	27.59	33.41	40.79	29	42.56	49.59	58.30	41	56.94	64.95	74.75
6	12.59	16.81	22.46	18	28.87	34.81	42.31	30	43.77	50.89	59.70	42	58.12	66.21	76.08
7	14.07	18.48	24.32	19	30.14	36.19	43.82	31	44.99	52.19	61.10	43	59.30	67.46	77.42
8	15.51	20.09	26.13	20	31.41	37.57	45.32	32	46.19	53.49	62.49	44	60.48	68.71	78.75
9	16.92	21.67	27.88	21	32.67	38.93	46.80	33	47.40	54.78	63.87	45	61.66	69.96	80.08
10	18.31	23.21	29.59	22	33.92	40.29	48.27	34	48.60	56.06	65.25	46	62.83	71.20	81.40
11	19.68	24.73	31.26	23	35.17	41.64	49.73	35	49.80	57.34	66.62	47	64.00	72.44	82.72
12	21.03	26.22	32.91	24	36.42	42.98	51.18	36	51.00	58.62	67.99	48	65.17	73.68	84.04

Table 7: Critical values of the χ^2 distribution for $p = .05$, .01, and .001 levels of significance up to 48 degrees of freedom.

36. Weisstein, E. W. *CRC Concise Encyclopedia of Mathematics, 2nd Ed.* Chapman & Hall, New York, 2003.
37. Wobbrock, J. O., Aung, H. H., Rothrock, B., and Myers, B. A. Maximizing the guessability of symbolic input. In *Proc. of CHI EA '05*, ACM (2005), 1869–1872.
38. Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. of CHI '11*, ACM (2011), 143–146.
39. Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-defined gestures for surface computing. In *Proc. of CHI '09*, ACM (2009), 1083–1092.

APPENDIX A: TEST STATISTIC FOR REPEATED MEASURES

We start with Cochran's Q test formula (see Table 3 for the notations we employ and their relation to agreement rates):

$$k(k-1)\cdot\frac{\displaystyle\sum_{j=1}^{k}\left(T_j-\frac{T}{k}\right)^2}{\displaystyle\sum_{i=1}^{n}R_i\left(k-R_i\right)} \tag{14}$$

We adapt this formula to our problem by expressing it in terms of agreement and coagreement rates.

The sum at the numerator of eq. 14 can be written as:

$$\sum_{j=1}^{k}\left(T_j\right)^2-2\frac{T}{k}\underbrace{\sum_{j=1}^{k}T_j}_{T}+\sum_{j=1}^{k}\left(\frac{T}{k}\right)^2=\sum_{j=1}^{k}\left(T_j\right)^2-\frac{T^2}{k}$$

and, knowing that $T_j = n\cdot\mathcal{AR}(r_j)$ and $T=\sum_{j=1}^{k}T_j$:

$$= \quad n^2\sum_{j=1}^{k}\mathcal{AR}^2(r_j)-\frac{n^2}{k}\left(\sum_{j=1}^{k}\mathcal{AR}(r_j)\right)^2$$

The sum at the denominator of eq. 14 can be written as:

$$\begin{aligned}
\sum_{i=1}^{n}R_i\left(k-R_i\right) &= k\underbrace{\sum_{i=1}^{n}R_i}_{T}-\sum_{i=1}^{n}\left(R_i\right)^2=kT-\sum_{i=1}^{n}\left(R_i\right)^2\\
&= kT-\sum_{i=1}^{n}\left(\sum_{t=1}^{k}\delta_{i,t}\right)^2\\
&= kT-\sum_{i=1}^{n}\left(\sum_{t=1}^{k}\sum_{s=1}^{k}\delta_{i,t}\cdot\delta_{i,s}\right)\\
&= kT-\sum_{t=1}^{k}\sum_{s=1}^{k}\underbrace{\left(\sum_{i=1}^{n}\delta_{i,t}\cdot\delta_{i,s}\right)}_{n\cdot\mathcal{CR}(r_t,r_s)}\\
&= kn\sum_{j=1}^{k}\mathcal{AR}(r_j)-n\sum_{t=1}^{k}\sum_{s=1}^{k}\mathcal{CR}(r_t,r_s)
\end{aligned}$$

The agreement rate test statistic can then be described solely in terms of agreement and coagreement rates between the k referents, as follows:

$$k(k-1)\cdot\frac{\displaystyle n^2\sum_{j=1}^{k}\mathcal{AR}^2(r_j)-\frac{n^2}{k}\left(\sum_{j=1}^{k}\mathcal{AR}(r_j)\right)^2}{\displaystyle kn\sum_{j=1}^{k}\mathcal{AR}(r_j)-n\sum_{t=1}^{k}\sum_{s=1}^{k}\mathcal{CR}(r_t,r_s)}$$

and, after simplification by $k\cdot n$:

$$(k-1)n\cdot\frac{\displaystyle\sum_{j=1}^{k}\mathcal{AR}^2(r_j)-\frac{1}{k}\left(\sum_{j=1}^{k}\mathcal{AR}(r_j)\right)^2}{\displaystyle\sum_{j=1}^{k}\mathcal{AR}(r_j)-\frac{1}{k}\sum_{t=1}^{k}\sum_{s=1}^{k}\mathcal{CR}(r_t,r_s)}$$

where $n=\frac{1}{2}|P|\left(|P|-1\right)$.

APPENDIX B: CRITICAL VALUES OF THE CHI-SQUARE DISTRIBUTION

For convenience, Table 7 lists the critical values of the χ^2 distribution for p=.05, .01, and .001 significance levels for 1 to 48 degrees of freedom.