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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

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 \( \chi^2 \) 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 suggesting the most popular proposal for a given referent (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] 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 \( W = .254 \)).

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).
Equation 1 captures this.

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:

\[
AR(r) = \sum_{P_i \subseteq P} \left[ \frac{|P_i|}{|P|} \right]^2
\]

(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 is \( \frac{15}{20} + \frac{5}{20} \), while the total number of pairs that could have been in agreement is \( \frac{20}{2} \). By dividing the two values, we obtain the agreement rate \( AR(r) = \frac{15}{20} = .605 \). 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 \( AR(r) \) 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 steps of simplification of eq. 2, we obtain:

\[
AR(r) = \frac{1}{|P|(|P|-1)} \sum_{P_i \subseteq P} (|P_i|^2 - |P_i|)
\]

\[
= \frac{1}{|P|(|P|-1)} \left( \sum_{P_i \subseteq P} |P_i|^2 - \sum_{P_i \subseteq P} |P_i| \right)
\]

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

\[
AR(r) = \frac{1}{|P|(|P|-1)} \sum_{P_i \subseteq P} |P_i|^2 - \frac{1}{|P|-1}
\]

We continue by placing \( |P_i|^2 \) 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):

\[
AR(r) = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left[ \frac{|P_i|}{|P|} \right]^2 - \frac{1}{|P|-1}
\]

(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 \( AR(r) \) with the two correcting factors. The following properties characterize the differences and relationship between the two definitions:

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

\[
\text{CR}(r_1, r_2) = \frac{\sum_{i=1}^{n} \delta_{i,1} \cdot \delta_{i,2}}{n}, \quad n = \frac{1}{2} |P| (|P| - 1)
\]

(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, 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.

The coagreement rate can be generalized to \( k \geq 2 \) referents:

\[
\text{CR}(r_1, r_2, ..., r_k) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} \delta_{i,j}, \quad n = \frac{1}{2} |P| (|P| - 1)
\]

(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 \( \text{AR} \) calculated from

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

$$Q = k(k - 1) \frac{\sum_{j=1}^{k} (T_j - \frac{1}{k} \sum_{r=1}^{k} T_r)^2}{\sum_{i=1}^{n} R_i (k - R_i)}$$

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

The result is the $V_{rd}$ statistic (see Appendix A for the mathematical details of the calculation procedure):
and \( \mathcal{AR}(r^*) \) under equation 10, which simplifies to:

\[
V_{rd}^* = \left| P \right| \cdot (|P| - 1) \cdot \mathcal{AR}(r)
\]

(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: \( \mathcal{AR}(r_1) = 0.227 \), \( \mathcal{AR}(r_2) = 0.697 \), and \( \mathcal{AR}(r_3) = 0.318 \). Coagreement rates are: \( \mathcal{CR}(r_1, r_2) = 0.152 \), \( \mathcal{CR}(r_1, r_3) = 0.045 \), and \( \mathcal{CR}(r_2, r_3) = 0.197 \). The \( V_{rd} \) statistic (eq. 9) is 28.964, which is significant at the \( p = 0.001 \) level, as indicated by the critical value for the \( \chi^2 \) 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 \( p = 0.05 \), 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 \( p = 0.001 \), which is below the Bonferroni corrected value of \( p = 0.05/2 = 0.025 \). Furthermore, \( \mathcal{AR}(r_1) = 0.227 \) is significantly greater than 0 at \( p = 0.001 \) as \( V_{rd} = 14.98 \), which is above the critical value of 10.83 of the \( \chi^2 \) 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 \( p = 0.05 \), \( 0.01 \), and \( 0.001 \) 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](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 \( (A = 0.409, \mathcal{AR} = 0.336) \). Using our statistical test, we found an overall significant effect of referent type on agreement rate \( (V_{rd}(41, N = 800) = 1466.818, p < 0.001) \). 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 \( \mathcal{AR} = 0.809 \)), but we also detected a significant effect of the direction of alignment (i.e., left, right, bottom, and top) on the resulting agreement \( (V_{rd}(3, N = 80) = 121.737, p < 0.001) \), no significant difference between Align Left and Align Right (both \( p > 0.900 \)), and significantly higher agreement for Align Bottom than for Align Top \( (905 \text{ versus } 632) \). To understand more, we ran coagreement analysis. The coagreement rate between Align Left and Align Right was \( 0.900 \), showing that all the pairs of participants that were in agreement for Align Left were also in agreement for Align Right. The co-agreement between Align Top and Align Bottom was \( 0.632 \), indicating that all the pairs of participants in agreement for Align Top \( (\mathcal{AR} = 0.632) \) were also in agreement for Align Bottom \( (\mathcal{AR} = 0.900) \), 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 \( \mathcal{CR} = 0.632 \), 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 \( (A = 0.446, \mathcal{AR} = 0.417) \). Using the \( V_{rd} \) test statistic, we found an overall significant effect of referent type on agreement rates \( (V_{rd}(39, N = 800) = 3523.962, p < 0.001) \). There were 8 referents that received high (i.e., \( p < 0.900 \) and 1.000) agreement rates, and we found a significant effect of referent type over agreement rate for this subset as well \( (V_{rd}(7, N = 160) = 106.176, p < 0.001) \). There were 10 referents that received agreement rates below \( 0.100 \) \( (V_{rd}(9, N = 200) = 11.033, n.s.) \). 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 vice versa” [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 (CR = .489, 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 (CR(X,Y) = .179, CR(X,Z) = .153, 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 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 (Vrd(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 (Vrd(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 (CR = .436). When analyzing the other dichotomous referents, we found more agreement for Open Menu than for Hide Menu (.118 versus .052, Vrd(1,N=36) = 4.454, p < .05), and nonsignificant differences between the agreement rates for Volume Up and Volume Down (.157 and .157, 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 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 (Vrd(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 AR formula with Wobbrock et al.’s original A measure [37]. We also present the probability distribution function of AR, and discuss the connection between agreement and disagreement.

The probability distribution function of AR
Figure 2 shows the probability distribution function of 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 AR.

![Figure 2: Probability distribution functions of AR computed for various numbers of participants |P| from 10 to 50.](image)

When we analyze the distribution shown in Figure 2, we find that the cumulative probability of 90% is reached for AR ≤ .374, while a cumulative 99% is reached for AR ≤ .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 AR ≤ .222 and 99% for AR ≤ .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 AR for 20 participants is .284);

---

5To compensate for the low resolution obtained for the probability distributions when binning 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.
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.

Table 5 shows more agreement rate values, all below $.450$.

Table 5: Average agreement rates $\hat{AR}$ reported in the literature and corrected $AR$ for 20 participants is $.221$. Table 5 shows qualitative interpretations for agreement rates, see Table 6.

### Relationship between agreement and disagreement rates

It is interesting to see how agreement rate $AR(r)$ compares with the disagreement rate $DR(r)$ for a given referent $r$, knowing that the two are complementary with respect to 1, i.e., $AR(r) + DR(r) = 1$. This equation tells us that there is more agreement than disagreement for referent $r$ if $AR(r) > .500$. 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:

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

that takes values between 0 (i.e., no agreement) and $\infty$ (absolute agreement). Figure 3 shows the values of this ratio. Informed by these results, the 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 $AR(r)$ (computed as the average of all possible agreement rate values, weighted by their probability of occurrence, according to Figure 2).](image-url)

<table>
<thead>
<tr>
<th>$AR(r)$ INTERVAL</th>
<th>PROBABILITY</th>
<th>INTERPRETATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq .100$</td>
<td>22.9%</td>
<td>low agreement</td>
</tr>
<tr>
<td>$.100 - .300$</td>
<td>59.1%</td>
<td>medium agreement</td>
</tr>
<tr>
<td>$.300 - .500$</td>
<td>14.1%</td>
<td>high agreement</td>
</tr>
<tr>
<td>$&gt;.500$</td>
<td>3.9%</td>
<td>very high agreement</td>
</tr>
</tbody>
</table>

† According to the probability distribution functions shown in Figure 2 and $|P| = 20$ 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 Thammatip Piumsomboon as well as their co-authors from [1,2,4,25] 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
The sum at the denominator of eq. 14 can be written as:

\[ \sum_{i=1}^{n} R_i (k - R_i) = k \sum_{i=1}^{n} R_i - \sum_{i=1}^{n} (R_i)^2 = kT - \sum_{i=1}^{n} (R_i)^2 \]

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:

\[ \sum_{j=1}^{k} \left( T_j - \frac{T}{k} \right)^2 = \frac{1}{k} \sum_{j=1}^{k} T_j \left( T_j - \frac{T}{k} \right)^2 = \frac{1}{k} \sum_{j=1}^{k} (T_j)^2 - \frac{T^2}{k} \]

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

\[ = n^2 \sum_{j=1}^{k} \text{AR}^2(r_j) - \frac{n^2}{k} \left( \sum_{j=1}^{k} \text{AR}(r_j) \right)^2 \]

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

\[ \sum_{i=1}^{k} \delta_{i,t} \cdot \delta_{i,s} = kT - \sum_{i=1}^{n} \left( \frac{k}{i} \cdot \delta_{i,t} \cdot \delta_{i,s} \right) = kT - \sum_{i=1}^{n} \delta_{i,t} \cdot \delta_{i,s} \]

\[ = kT - \sum_{i=1}^{n} \left( \sum_{s=1}^{k} \delta_{i,t} \cdot \delta_{i,s} \right) \]

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

\[ kT - \sum_{i=1}^{n} \left( \sum_{j=1}^{k} \text{AR}(r_j) \right)^2 \]

and, after simplification by \( k \cdot n \):

\[ \sum_{j=1}^{k} \text{AR}(r_j) - \frac{1}{k} \sum_{j=1}^{k} \text{AR}(r_j) \]

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

**APPENDIX B: CRITICAL VALUES OF THE CHI-SQUARE DISTRIBUTION**

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