Between-Subjects Elicitation Studies: Formalization and Tool Support

Radu-Daniel Vatavu
MintViz Lab | MANSiD Research Center
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

Elicitation studies, where users supply proposals meant to effect system commands, have become a popular method for system designers. But the method to date has assumed a *within-subjects* procedure and statistics. Despite the benefits of examining the relative agreement of independent groups (*e.g.*, men versus women, children versus adults, novices versus experts, etc.), the lack of appropriate tools for *between-subjects* agreement rate analysis have prevented so far such comparative investigations. In this work, we expand the elicitation method to *between-subjects* designs. We introduce a new measure for evaluating coagreement between groups and a new statistical test for agreement rate analysis that reports the exact *p*-value to evaluate the significance of the difference between agreement rates calculated for independent groups. We show the usefulness of our tools by re-examining previously published gesture elicitation data, for which we discuss significant differences in agreement for technical and non-technical participants, men and women, and different acquisition technologies. Our new tools will enable practitioners to properly analyze their user-elicited data resulted from complex experimental designs with multiple independent groups and, consequently, will help them understand agreement data and verify hypotheses about agreement at more sophisticated levels of analysis.

Author Keywords

Guessability study, elicitation study, participatory study, between-subjects design, agreement rate, methodology, statistical test, user-defined gestures, toolkit.

ACM Classification Keywords

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

INTRODUCTION

Participatory design studies are a powerful tool to understand users' perceptions, attitudes, and conceptual models for interacting with new prototypes and applications that are yet in the design stage [4,22]. Consequently, they provide designers with valuable user feedback at the early stages of the design process to inform better prototypes. Elicitation experiments are one instance of participatory design studies that enable user interface designers to collect and cluster users' preferences for specific interactive situations for which little or no design knowledge exists in the community. First proposed by Wobbrock *et al.* in 2005 [50], elicitation studies have since been widely adopted by researchers to compile gesture sets for various interaction scenarios [30,37,42,51] and applications [3,35,36,43,44,48].

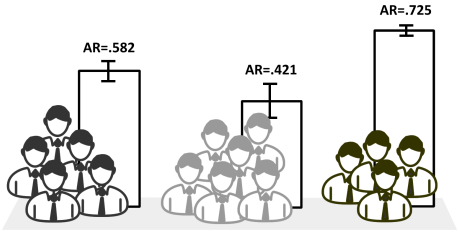

Figure 1: Comparing agreement rates of different groups of participants opens new ways to understand user-elicited data, not possible before this work; *e.g.*, we can now investigate the effect of age group [39], gender [44], technical expertise [44], disability [20], gesture implementer [2], etc., on agreement. Our new tools (measures, statistical test, and software application) enable agreement analysis for elicitation studies with between-subjects designs and complement the tools of Vatavu and Wobbrock [47] for within-subject elicitation experiments.

However, inferring experimental findings of elicitation studies (*i.e.*, agreement rates [50]) to the entire user population cannot be done without statistical reasoning. The lack of tools for analyzing user-elicited data has been a major deficiency of the elicitation methodology [50,51], depriving practitioners from the probabilistic assurance of the significance and validity of their findings when extended to the wider user population. Vatavu and Wobbrock [47] have recently made the first steps toward formalizing agreement analysis in the Human-Computer Interaction community with the V_{rd} statistic that evaluates the effect of referents (*i.e.*, commands to be invoked in some application) on participants' agreement rates. While their test considerably strengthens the designer's confidence in interpreting agreement results (*i.e.*, it enables statements such as *"users' agreement over referent 'zoom-in' was found significantly higher than agreement over referent 'zoom-out',* $p < .05$*"*), the test only applies for data collected from elicitation experiments with *within-subjects* designs in which the same participants contribute to all referents.

CHI'16, May 07–12, 2016, San Jose, CA, USA.

ACM 978-1-4503-3362-7/16/05...$15.00
http://dx.doi.org/10.1145/2858036.2858228

Nevertheless, the opportunity to compare agreement *between* independent groups of participants would open new ways to understand user-elicited data at deeper levels of sophistication; see Figure 1. For example, we may ask whether users with technical backgrounds are more inclined to reach agreement than non-technical users to understand more precisely the practical implications of the legacy bias phenomenon [31,38]. We may also ask whether men tend to reach higher agreement rates than women in terms of the interactive gesture commands they propose, knowing that women generally employ more non-verbal behavior and body movement than men [16]. Also, we may want to study the effect of age group (*i.e.*, child or adult) on the agreement of what is intuitive touch gesture interaction to inform user interface designs for smartphones and tablets adapted to the user's age [45,46]. Or, we may simply want to compare the agreement rates reached by different researchers running independent elicitation studies within the same application domain [43,48,52] as a reliable way to validate previous findings and confirm the reproducibility of agreement results. While these are just a few directions of investigation in the area of gesture set design made possible by our new tools, our formalization of between-subjects elicitation studies is likely to have a much broader impact to virtually any elicitation experiment with a between-subjects design.

The contributions of this work are: (1) we propose a formalization of the guessability methodology [50] for elicitation studies with *between-subjects* experimental designs, for which we introduce a new measure to characterize the degree of agreement shared *between* independent groups of participants; (2) we introduce a statistical test for comparing agreement rates of different groups that reports the *exact p*-value of significance in analogy with the principles behind Fisher's exact test [13]; (3) we conduct a re-examination of several published elicitation datasets that shows the benefits of our methodology to compare agreement rates between independent groups. As an example, this work is the first to examine the effect of participants' technical background on agreement, an important aspect to further understand the legacy bias phenomenon for gesture elicitation studies [31]; also, it is the first work to examine the influence of participants' gender on agreement over elicited gesture proposals; (4) we offer a toolkit to compute agreement measures and run tests of significance for $k \geq 2$ independent groups. It is our hope that these contributions will advance the current knowledge in agreement rate analysis for elicitation studies, allowing researchers and practitioners to understand user-elicited data at unprecedented levels of rigor.

RELATED WORK

We review in this section prior work on measuring agreement between participants or raters in scientific experiments. We also discuss elicitation studies employed in the HCI community, and we examine tools and measures for computing and analyzing agreement observed from users' elicited proposals.

Measures of agreement analysis

The degree of consensus between participants or raters has been evaluated with various measures in the literature of many scientific disciplines [15]. For example, in statistics, inter-rater reliability experiments ask participants (referred to as "raters") to independently produce categorizations of a sample of objects or phenomena into predefined categories or to rank the subjects of a study according to some criterion. For instance, two or multiple raters are asked to categorize the performance achieved by several subjects tested for some task into three categories, *e.g.*, "basic", "intermediate", or "advanced". Or, the raters need to rank subjects by how well they did during the evaluation task. The extent to which raters' categorizations match each other represents the *inter-rater reliability*. The goal of inter-rater reliability testing is to guarantee the interchangeability of raters, *i.e.*, the researcher does not need to worry about which specific categorization they use from the available raters, if they know that categorization is not affected by the rater factor. Inter-rater reliability has been evaluated using the percent agreement [18,32], Scott's π statistic [40], and Cohen's κ (kappa) [8] and Fleiss' κ [14] coefficients.

The percent agreement, popularized by Osgood [32] and Holsti [18] is probably the first (and the most basic) approach to capture the degree of consensus between raters when there is a fixed number of rating categories. The statistic is calculated as the percent of all ratings for which all raters were in agreement. For instance, if 2 raters each classify 10 subjects into one of 3 categories and they all agree 4 times for the first category, 2 times for the second, and 5 times for the third, the percent agreement is $11/(3 \cdot 10)$=36.7%. However, because the problem is to assign subjects to a *fixed number* of predefined categories, agreement by chance may occur when raters are not sure about the right categorization. In that case, even if raters independently make a subjective choice, they could still be in agreement, which is an undesirable outcome since it does not follow from raters mastering the rating process, but rather from chance alone. This issue has caused several chance-corrected agreement coefficients to be proposed in the literature. For example, the Brennan-Prediger coefficient [5] corrects the percent agreement according to the number of nominal categories in the rating scale (q) by considering that the probability of chance agreement is proportional to $1/q$. A popular coefficient to measure agreement of two raters has been Cohen's κ [8], extended by Fleiss' κ [14] to more than two raters. The κ coefficient subtracts the probability of agreement occurring by chance (*i.e.*, the probability of expected agreement, p_e) from the percent agreement and divides the result by the degree of agreement attainable above chance (*i.e.*, $1 - p_e$). Scott's π coefficient uses a similar calculation formula, but computes the probability of chance agreement slightly differently [40]. The weighted κ coefficient reports agreement by weighting different disagreement situations with different weights to properly reflect the seriousness by which those disagreements may affect the final result [9]. When rankings of subjects are required instead of their categorization into predefined classes, Kendall's W coefficient [21] can be used to assess the degree of agreement between multiple raters.

Unfortunately, the above statistics are not appropriate to evaluate agreement for elicitation studies [50,51], during which participants suggest proposals for referents *without* being offered any set of predefined categories. The particularity of an elicitation study is that the researcher wants to understand participants' *unconstrained* preferences over some task, which

ultimately leads to revealing participants' conceptual models for that task [51]. Consequently, the range of proposals is potentially infinite, only limited by participants' power of imagination and creativity. This limitation of existing statistics to reflect agreement between participants for elicitation studies has led Wobbrock *et al.* [50] to introduce the Agreement Rate measure to properly quantify the consensus between participants' proposals. Findlater *et al.* [12] first improved the agreement rate formula and then the measure was updated by Vatavu and Wobbrock [47].

The elicitation methodology and agreement rates

Wobbrock *et al.* [50] proposed a participatory design methodology for maximizing the guessability of symbolic input for user interfaces. To apply the methodology, the practitioner presents participants with commands called "referents," for which "proposals" are asked. For example, the practitioner may want to design a good keyboard shortcut to invoke a new menu command called "Upload to FTP account." He asks 20 participants to think of good combinations of keys for that command, and then groups all elicited data into clusters of equivalent proposals. The number of proposals that are equivalent is used to calculate the agreement rate among participants. For instance, say that 15 participants believe "Ctrl+U" is a good keyboard shortcut for that command, while the remaining 5 are in favor of "Shift+F". The resulted agreement rate is then computed as the sum of square ratios reflecting the support that each proposal has in the group, $A = \left(\frac{15}{20}\right)^2 + \left(\frac{5}{20}\right)^2 = .625$ [50] (p. 1871). Wobbrock *et al.* evaluated the guessability methodology on the EdgeWrite alphabets [50] and, later, applied it for the first gesture elicitation study on tabletops [51].

Findlater *et al.* [12] proposed a variation for Wobbrock *et al.*'s original agreement rate measure [50] that evaluates in $[0..1]$. Recently, Vatavu and Wobbrock [47] introduced a corrected version for the agreement rate formula that has even more desirable properties, such as reported agreement is more consistent under large samples of participants with the same ratios of proposals. The corrected agreement rate for the above example is $\mathcal{AR} = \frac{20}{19}\left(\left(\frac{15}{20}\right)^2 + \left(\frac{5}{20}\right)^2\right) - \frac{1}{19} = .605$, where coefficients $\frac{20}{19}$ and $\frac{1}{19}$ have the role to correct the magnitude of agreement with respect to the sample size of participants according to the actual number of degrees of freedom; see Vatavu and Wobbrock [47] for more examples and discussions.

Researchers have also worked with other measures derived from agreement rates when analyzing data from elicitation studies. For example, Vatavu and Wobbrock [47] discussed disagreement and coagreement measures. They also generated the probability distribution function for agreement rate, which made possible the interpretation in a larger context of the magnitudes of agreement rates reported in the literature. Morris [30] proposed two new measures, max-consensus and the consensus-distinct ratio, to evaluate the agreement between participants when multiple proposals were elicited. Vatavu and Zaiţi [48] used Kendall's W coefficient of concordance [21] together with agreement rates, and reported similar magnitudes. Chong and Gellersen [6] proposed a measure of popularity of preferences function of the number of participants and the number of times the same proposal occurred. Vatavu [44] defined confidence values for referents as the maximum percent of participants in consensus for those referents.

Applications of elicitation studies

Elicitation studies have found valuable application in the HCI community for gesture interface design. For example, Wobbrock *et al.* [51] was the first to employ the elicitation methodology in a study aimed at understanding users' conceptual models of touch gesture interaction on tabletops. From that point on, the community has started to employ elicitation studies for a variety of gesture acquisition technologies and application domains. For example, Ruiz *et al.* [37] examined motion gestures for invoking generic commands on mobile devices; Vatavu [44] and Kuhnel *et al.* [24] investigated users' preferences for controlling home appliances with gestures; Vatavu [43] and Vatavu and Zaiţi [48] examined free-hand Kinect and Leap Motion gestures for Smart TVs; Piumsomboon *et al.* [35,36] explored users' gesture preferences for augmented reality interfaces; Seyed *et al.* [41] and Kurdyukova *et al.* [26] examined gestures for multi-display environments; Anthony *et al.* [2] employed the agreement rates methodology to characterize user consensus in articulating stroke gestures; and Kray *et al.* [23] elicited gestures that span multiple devices.

AN EXACT TEST OF SIGNIFICANCE FOR EVALUATING THE EFFECT OF STUDY GROUP ON AGREEMENT RATE

Preliminaries

Let G_1, G_2, ..., G_k be k independent groups of participants from which proposals were elicited for some referent r. For example, these may be k=2 groups with G_1 composed of male and G_2 of female participants, for which we want to learn whether men reached higher agreement over r than women. Or, they may represent k=3 groups with G_1 composed of children less than 6 years old, G_2 children between 7 and 18 years old, and G_3 composed of adults, for which we want to examine whether age affects consensus over some referent r.

We employ the definition of Vatavu and Wobbrock [47] to evaluate the amount of agreement in each group G_i, $i = 1..k$, expressed as the ratio of the number of pairs of participants from G_i that *are* in agreement over referent r and the maximum number of pairs from G_i that *could have been* in agreement:

$$\mathcal{AR}(r, G_i) = \frac{\sum_{P_j \subseteq G_i} \frac{1}{2}|P_j|\left(|P_j| - 1\right)}{\frac{1}{2}|G_i|\left(|G_i| - 1\right)} \qquad (1)$$

where P_j are subsets of participants from group G_i that are in agreement over r, and $|P_j|$ denotes the cardinality of subset P_j. For example, suppose that in a group of $|G_i|$=10 participants, three distinct proposals emerge for referent r supported by three subgroups of $|P_1|$=5, $|P_2|$=3, and $|P_3|$=2 participants. This elicitation result allows us to write the group size as the partition $10 = 5 + 3 + 2$. The agreement rate of the group is $\mathcal{AR}(r, G_i) = (\frac{5\cdot4}{2} + \frac{3\cdot2}{2} + \frac{2\cdot1}{2}) / \frac{10\cdot9}{2} = \frac{14}{45} = .311$.

Agreement rates can also be expressed using Kronecker's $\delta_{p,q}$ notation [19] (p. 240), where we set $\delta_{p,q}(r)$ to 1 when

participants p and q are in agreement over r and 0 otherwise:

$$\mathcal{AR}(r, G_i) = \frac{a_i}{n_i} = \frac{\sum_{p=1}^{|G_i|} \sum_{q=p+1}^{|G_i|} \delta_{p,q}(r)}{\frac{1}{2}|G_i|\,(|G_i|-1)} \quad (2)$$

where we use a_i to denote the number of pairs in agreement in group G_i and n_i the total number of pairs of that group, $n_i = \frac{1}{2}|G_i|(|G_i|-1)$. Correspondingly, the number of pairs in disagreement is $d_i = n_i - a_i$. Using the a_i and d_i notations, we organize agreement data for referent r as a $2 \times k$ contingency table listing the frequencies of a dichotomous variable AGREEMENT for each group; see Table 1. We denote by n the number of all pairs in all the k groups, $n = \sum_{i=1}^{k} n_i$.

Contingency tables, such as Table 1, are traditionally analyzed for the association between row and column variables using Pearson's Chi-Square test [33], the *G*-test [7], or Fisher's exact test [13]. Pearson's Chi-Square test compares the observed frequencies in the data with the frequencies expected to arise by chance under the null hypothesis and is implemented with the χ^2 statistic as the sum of normalized square differences between observed and expected frequencies. The *G*-test is a log-likelihood ratio significance test for which the statistic takes the form of the Kullback-Leibler divergence [25] for observed and expected frequencies. Both Pearson's Chi-Square test and the *G*-test have asymptotic chi-square distributions (*i.e.*, the larger the sample is, the better their sampling distributions approximate the chi-square distribution under the null hypothesis), and one test may be preferred over the other under various small samples, sparseness, and efficiency assumptions [10]. Actually, it has been shown that Pearson's formulation of the χ^2 statistic is an approximation of the G^2 statistic obtainable by expanding the G^2 log-likelihood formula in a Taylor series, an approximation that works well when the differences between observed and expected frequencies are small [17] (pp. 4-5). However, because they are approximate tests, results may not be accurate when sample sizes are small or when data is unequally distributed among the cells of the contingency table; *e.g.*, Pearson's Chi-Square is not recommended when there are cells in the table with less than 5 observations [13] (p. 96). Nonetheless, low agreement rates are frequently observed in gesture elicitation studies (see Vatavu and Wobbrock [47], Table 5 for an overview of agreement results in gesture elicitation research), which makes Pearson's Chi-Square or the *G*-test inadvisable for analyzing agreement frequencies. On the other hand, Fisher's exact test computes the exact value of the probability of observing the frequencies of a 2×2 contingency table and, consequently, is the preferred test when sample sizes are small. Fisher's exact test also generalizes to $R \times C$ contingency tables [29].

Still, none of the above tests can be directly applied to analyze agreement data. The reason is that they all assume *independent* and *identical trials* as a condition to derive the formulas of their test statistics, which is not always the case when measuring agreement at the level of pairs of participants. For example, once we have measured that participant p is in agreement with participant q over some referent, and also that participant p is in agreement with participant u, then it also must be that participants q and u are in agreement for that referent, which makes the (q, u) observation dependent on the first two. On the other hand, we cannot say anything about (q, u) when both p and q and p and u are not in agreement. Because dependence arises not from the study design, but simply in the data itself, it is not *a priori* dependence, but dependence that *may* occur *a posteriori*. Note that independence of observations *is* present from the start because every participant is different from every other and every observation has the chance to be independent, but pairing participants may induce dependence. This transitivity of agreement means that the probability of observations that may turn out to be dependent on previous ones is 1.00. Consequently, the formula of Fisher's exact test [13] (p. 100) may not prove accurate for all cases of agreement, because it assumes the same probability for the event to occur in each trial, *i.e.*, the identical trials assumption. Therefore, a new test would be preferable to prevent possible inaccurate conclusions that traditional tests may produce for agreement data.

AGREEMENT(r)	GROUPS OF PARTICIPANTS				TOTAL
	G_1	G_2	...	G_k	
YES	a_1	a_2	...	a_k	a
NO	d_1	d_2	...	d_k	d
TOTAL	n_1	n_2	...	n_k	n

Table 1: Contingency table showing the number of pairs of participants that are in agreement (a_i) and in disagreement (d_i) over referent r for each independent group G_i, $i = 1..k$.

In the following, we formulate one such significance test to evaluate the effect of independent study groups on agreement. We were inspired by the principle of Fisher's exact test [13] (p. 99-101) that calculates the exact probability to observe a given numerical configuration of the contingency table. In our case, we are concerned with the probability of observing a given configuration of agreement for k independent groups, *i.e.*, the values a_1, a_2, ..., a_k corresponding to column totals n_1, n_2, ..., n_k in Table 1. In other words, we wish to evaluate how likely our observed configuration of agreement is from all possible agreement configurations that could have occurred for the same numbers of pairs n_i in each group. The null and alternative hypotheses for agreement rates over referent r computed for k independent groups are:

H_0: ALL GROUPS HAVE EQUAL AGREEMENT RATES.
H_1: THERE IS A DIFFERENCE AMONG THE AGREEMENT RATES OF THE $k \geq 2$ GROUPS.

The null hypothesis states that the relative proportions of the number of pairs in agreement (a_i) to the total number of pairs in each group (n_i) are independent of the study group, *i.e.*, no association between the rows and the columns of the table.

Counting all possible ways to reach agreement

We show in the following how to count all the possible ways in which a given number of pairs in agreement (a_i) can be reached for a group of size $|G_i|$. We start by noting that while the a_i values are integers that range between 0 and $n_i = \frac{1}{2}|G_i|(|G_i|-1)$ corresponding to the extreme cases of no

PARTITIONS OF PROPOSALS α	$a_i(\alpha)$	$f_{\alpha\|21}$
α_1 : $7=1+1+1+1+1+1+1$	0	1
α_2 : $7=1+1+1+1+1+2$	1	21
α_3 : $7=1+1+1+1+3$	3	35
α_4 : $7=1+1+1+2+2$	2	105
α_5 : $7=1+1+1+4$	6	35
α_6 : $7=1+1+2+3$	4	210
α_7 : $7=1+1+5$	10	21
α_8 : $7=1+2+2+2$	3	105
α_9 : $7=1+2+4$	7	105
α_{10} : $7=1+3+3$	6	70
α_{11} : $7=1+6$	15	7
α_{12} : $7=2+2+3$	5	105
α_{13} : $7=2+5$	11	21
α_{14} : $7=3+4$	9	35
α_{15} : $7=7$	21	1
	TOTAL	877

Table 2: All the possible partitions of proposals for a group of $|G_i|$=7 participants showing the number of pairs in agreement $a_i(\alpha)$ as well as the frequency $f_{\alpha|21}$ of observing each partition α given the $(7 \cdot 6)/2 = 21$ distinct pairs of participants from G_i. NOTE: partition 1+2+4 means that there are three subgroups of participants in agreement: one subgroup of 4, one of 2, and one of 1 participant, with the resulting number of pairs in agreement being $a_i = \frac{4\cdot3}{2} + \frac{2\cdot1}{2} + \frac{1\cdot0}{2} = 7$.

agreement ($a_i = 0$) and absolute agreement ($a_i = n_i$) between n_i pairs of participants, not all the intermediate values between these two extremes are actually attainable. For instance, all the possible values observable for a_i for a group of $|G_i|$=7 participants (n_i=21) are $\{0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 15, 21\}$, and there is no way to obtain, for instance, 8, 13, or 17 pairs in agreement from a set of that size (see Table 2). This result follows from the transitivity of agreement across pairs of participants, *i.e.*, if participant p is in agreement with participant q and also in agreement with u, then it also follows that participants q and u are in agreement. However, while some values cannot be reached at all for some a_i given $|G_i|$, others may occur from multiple partitions. For example, there is more than one way to obtain a value of $a_i = 6$, such as from partitions α_5 and α_{10} listed in Table 2 that correspond to the sets of proposals {♥, ♥, ♥, ♥, ❖, ★, ♣} and {♣, ♣, ♣, ♥, ♥, ♥, ★}. At the same time, there is only one partition (α_{12}) that will evaluate to exactly 5 pairs of participants in agreement, *i.e.*, the set of proposals {❖, ❖, ♥, ♥, ♣, ♣, ♣}. Consequently, some values for a_i will be impossible to obtain for some group size $|G_i|$ and, from the values actually attainable, some will be theoretically more probable to observe than others.

We now consider the number of ways in which participants may permute under the same partition. Each partition α can be observed multiple times depending on which actual participants from G_i are in agreement each time. For instance, partition 7=2+2+3 means that there are three subgroups of participants in agreement: one subgroup of 3 and two subgroups of 2 participants each. However, the actual participants that compose these subgroups are free to vary. The number of ways in which participants may vary for a partition α gives the frequency of occurrence of that partition, which we denote by $f_{\alpha|n_i}$. For instance, the sets of proposals {♣, ♣, ♣, ❖, ❖, ♥, ♥}, {♣, ♣, ❖, ♣, ❖, ♥, ♥}, and {♥, ♣, ♣, ♥, ❖, ❖, ♣} represent different permutations of the same partition 7=2+2+3: in the first permutation, participants 1, 2, and 3 form the subgroup of three (that suggested the proposal ♣); for the second permutation, we have participants 1, 2, and 4 for the same subgroup of three; and, in the third permutation, the subgroup of three is composed of participants 2, 3, and 7.

In the following, we show how to arrive at a formula for calculating frequencies $f_{\alpha|n_i}$ by working with an example. Consider the same partition as before (7=2+2+3) with three subgroups that agreed on three distinct proposals. We can choose the 2 participants of the first subgroup from the set of 7 participants in $\binom{7}{2}$ ways[1], which then leaves us the option to choose the 2 participants of the second subgroup in $\binom{7-2}{2}=\binom{5}{2}$ ways and, lastly, we can choose the 3 participants of the third subgroup in only $\binom{7-2-2}{3}=\binom{3}{3} = 1$ way. Therefore, the total number of ways in which we can form the three subgroups with 7 participants is $\binom{7}{2}\binom{5}{2}\binom{3}{3}$. In general, we write partition $\alpha = \sum_j |P_j|$ (see eq. 1), for which the above product of binomial coefficients for group G_i becomes:

$$\prod_j \binom{|G_i| - \sum_{t<j} |P_t|}{|P_j|} \tag{3}$$

Going back to our example, we note that we have two subgroups of size 2 and it does not matter the order in which these subgroups are considered, *i.e.*, whether the first or the second group comes first in the partition is irrelevant to the structure of the partition. Therefore, we need to divide our result by 2!, which gives a frequency of $\binom{7}{2}\binom{5}{2}\binom{3}{3}/2! = 105$. In general, if we have m subgroups of the same size under partition α, we need to divide the result of eq. 3 by $m!$, because the order in which subgroups of the same size appear in the partition is not relevant. The final formula for $f_{\alpha|n_i}$ is thus:

$$f_{\alpha|n_i} = \frac{1}{\prod_{m\geq 2} m!} \prod_j \binom{|G_i| - \sum_{t<j} |P_t|}{|P_j|} \tag{4}$$

where m represents numbers of subgroups of the same size occurring more than once in partition α.

We can now compute the frequency of observing exactly a_i pairs of participants in agreement given n_i total pairs as:

$$f_{a_i|n_i} = \sum_{\alpha \to a_i} f_{\alpha|n_i} \tag{5}$$

where the sum goes over all the partitions α of group size $|G_i|$ that evaluate to exactly a_i pairs in agreement ($\alpha \to a_i$). Considering the data shown in Table 2, the frequency of observing 3 pairs of participants in agreement from a total of 21 pairs is $f_{3|21} = f_{\alpha_3|21} + f_{\alpha_8|21} = 35 + 105 = 140$. The example box next shows more calculation details of how to count frequencies of agreement levels a_i for group sizes $|G_i|$.

[1] $\binom{n}{k}$ is the binomial coefficient of "n choose k", $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

☞ EXAMPLE. Let's assume two independent groups of sizes $|G_1|$=7 and $|G_2|$=8 participants from which proposals were collected for a given referent r. For the first group, the partition of proposals is 7=5+1+1 (*i.e.*, three distinct proposals for r), while for the second group participants agreed in proportions 8=3+3+1+1 (*i.e.*, four distinct proposals were observed for r). The number of pairs in agreement for the two groups are a_1=10 and a_2=6, out of the total numbers of pairs that could have been in agreement n_1=21 and n_2=28, which makes agreement rates $\mathcal{AR}_1(r) = \frac{10}{21} = .476$ and $\mathcal{AR}_2(r) = \frac{6}{28} = .214$. We are interested to learn how frequently a_1=10 and a_2=6 pairs in agreement can be observed for the two groups under random samplings of participants. For the first group, 7=5+1+1 is the only partition that evaluates to 10 (see Table 2), for which we have $\binom{7}{5}$ ways to choose the 5 participants of the first subgroup, $\binom{2}{1}$ ways to choose 1 participant for the second subgroup, and $\binom{1}{1}$=1 way to select the remaining participant for the third subgroup. Because we have two subgroup sizes of one, the frequency of observing 10 pairs in agreement from 21 total pairs is $f_{10|21} = \frac{1}{2!}\binom{7}{5}\binom{2}{1}\binom{1}{1} = 21$. For the second group, there are two partitions that evaluate to a_2=6 pairs in agreement, namely 8=3+3+1+1 and 8=4+1+1+1+1 with frequencies $\frac{1}{2!\cdot 2!}\binom{8}{3}\binom{5}{3}\binom{2}{1}\binom{1}{1} = 280$ and $\frac{1}{4!}\binom{8}{4}\binom{4}{1}\binom{3}{1}\binom{2}{1}\binom{1}{1} = 70$. Therefore, the frequency of observing 6 pairs in agreement from 28 total pairs is $f_{6|28} = 280 + 70 = 350$.

Computing the probability of reaching agreement

Knowing now that the observance of a given number of pairs of participants in agreement a_i is conditioned by the number of all the distinct pairs in that group n_i, we refer to the *conditional probability* of observing a_i given n_i, which we denote $\pi_{a_i|n_i}$ following our previous convention and the notations of Agresti [1] (p. 37) for describing contingency tables. This probability is calculated by dividing the frequency of observing exactly a_i pairs of participants in agreement ($f_{a_i|n_i}$) by the sum of frequencies of observing agreement at any level among the n_i pairs of group G_i. However, what is really of interest to us at this moment is the conditional probability of observing an agreement configuration of exactly $a_1, a_2, ..., a_k$ pairs in agreement given $n_1, n_2, ..., n_k$ distinct pairs of participants (*i.e.*, the column marginal totals of Table 1), which is:

$$\pi_{a_1,a_2,...,a_k|n_1,n_2,...,n_k} = \prod_{i=1}^{k} \frac{f_{a_i|n_i}}{\sum_{\epsilon_1,\epsilon_2,...,\epsilon_k} \prod_{i=1}^{k} f_{\epsilon_i|n_i}} \quad (6)$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_k$ denote all possible configurations of agreement for the k groups. This probability tells us how likely it is to see the particular configuration of agreement that was actually observed during the experiment, given all the possibilities to obtain agreement for the k groups.

Now, similar to Fisher's exact test [13] (pp. 99-101), we want to know what is the probability of observing agreement configurations that are "more extreme" than the configuration resulted from our study, *i.e.*, that deviate more from the null hypothesis according to which the proportions of pairs in agreement are the same across all the k groups. This approach is similar to computing the difference between agreement rates (*e.g.*, $.476 - .214 = .262$; see the previous example) and evaluating the probability of the difference being close to zero. To evaluate extremeness, we use a modified version of the χ^2 statistic comparing observed and expected agreement:

$$V_b(\epsilon_1, ..., \epsilon_k) = \sum_{i=1}^{k} \left(\epsilon_i - \frac{n_i \sum_{j=1}^{k} \epsilon_j}{n} \right)^2 \quad (7)$$

where n is the total number of pairs and $\frac{n_i \sum_{j=1}^{k} \epsilon_j}{n}$ is the expected number of pairs in agreement for group G_i under the null hypothesis, *i.e.*, same proportions for all groups. For example, the expected numbers of pairs in agreement for two groups of size $|G_1|$=7 and $|G_2|$=6 for which a_1=15 and a_2=3 pairs in agreement were observed are $\frac{21\cdot(15+3)}{21+15}$=10.5 and $\frac{15\cdot(15+3)}{21+15}$=7.5. The V_b statistic[2] will be larger for agreement configurations that deviate more from the expected levels for each group. We don't normalize the terms of eq. 7 by the expected agreement (as in the original formula of χ^2), because a zero level of agreement is possible to obtain. Using the V_b statistic, we evaluate one configuration of agreement more extreme than another if its associated V_b statistic is larger.

☞ EXAMPLE. Let's consider two groups of equal size, $|G_1|$=7 and $|G_2|$=7 for which a_1=10 and a_2=6 pairs in agreement were observed ($\mathcal{AR}_1$=.476 and $\mathcal{AR}_2$=.286). We want to calculate the probability of seeing such a configuration of agreement from all possible configurations of n_1=21 and n_2=21 total pairs. Knowing that a_1 and a_2 can take 13 possible distinct values for a group of size 7 (see Table 2), there are 169 ways to reach agreement in the two groups. Therefore, the probability to observe the configuration 10 and 6 is:

$$\pi_{10,6|21,21} = \frac{f_{10|21} f_{6|21}}{\sum_{\epsilon_1}\sum_{\epsilon_2} f_{\epsilon_1|n_1} f_{\epsilon_2|n_2}} = \frac{21 \cdot 105}{769129} = 0.00286$$

which means there is 0.3% chance to observe exactly 10 and 6 pairs in agreement for two groups of size 7 each.

We can now define the conditional probability of observing the configuration of exactly $a_1, a_2, ..., a_k$ pairs in agreement or configurations that are more extreme given $n_1, n_2, ..., n_k$ distinct pairs of participants:

$$\Pi_{a_1,a_2,..,a_k|n_1,n_2,...n_k} = \frac{1}{2}\pi_{a_1,a_2,..,a_k|n_1,n_2,...n_k} + \sum_{\substack{\epsilon_1,...,\epsilon_k \\ V_b(\epsilon_1,\epsilon_2,...,\epsilon_k) > V_b(a_1,...,a_k)}} \pi_{\epsilon_1,\epsilon_2,...,\epsilon_k|n_1,n_2,...,n_k} \quad (8)$$

[2]Notation V in V_b stands for the variation between agreement rates and subscript $_b$ denotes a between-subjects experimental design.

a_1	a_2	V_b	$\pi_{a_1,a_2\|21,21}$	a_1	a_2	V_b	$\pi_{a_1,a_2\|21,21}$
0	15	112.5	.00001	15	0	112.5	.00001
0	21	220.5	.00000	15	1	98.0	.00019
1	15	98.0	.00019	15	2	84.5	.00096
1	21	200.0	.00003	21	0	220.5	.00000
2	15	84.5	.00096	21	1	200.0	.00003
2	21	180.5	.00014	21	2	180.5	.00014
3	21	162.0	.00018	21	3	162.0	.00018
4	21	144.5	.00027	21	4	144.5	.00027
5	21	128.0	.00014	21	5	128.0	.00014
6	21	112.5	.00014	21	6	112.5	.00014
7	21	98.0	.00014	21	7	98.0	.00014

Table 3: Agreement configurations that are more extreme in terms of the V_b statistic than a_1=10 and a_2=6 for two groups of 7 participants each. NOTE: The V_b statistic for 10 and 6 pairs in agreement is 8.0, see the example box in the text.

where the sum goes through all the agreement configurations $\epsilon_1, \epsilon_2, ..., \epsilon_k$ that are more extreme (*i.e.*, less likely) than the observed one, to which we add half the probability for our observed agreement. Equation 8 is known as the mid-*P* method [27] and is recommended over the ordinary *P*-value technique (*i.e.*, without the $\frac{1}{2}$) for small-sample distributions as a *"sensible compromise between having overly conservative inference and using irrelevant randomization to eliminate problems from discreteness"* [1] (p. 20).

☞ EXAMPLE. Let's compute the V_b statistic for each of the 169 possible ways to obtain agreement for the two groups of the previous example. The expected number of pairs in agreement corresponding to a=10+6=16, n_1=21, and n_2=21 is $\frac{16 \cdot 21}{42}$=8 for each group. Therefore, the V_b statistic for 10 and 6 pairs in agreement, respectively, is $(10-8)^2+(6-8)^2$=8. Table 3 lists the values of the V_b statistics and probabilities for all configurations of agreement that are more extreme than the observed one. The probability $\Pi_{10,6|21,21}$ to observe our agreement data or more extreme proportions is then .0039, which results from adding up all the probabilities of more extreme configurations and half of the probability of the configuration of agreement observed in the study.

A test of significance for between-subjects designs

The value of the cumulative probability $\Pi_{a_1,a_2,..,a_k|n_1,n_2,...n_k}$ tells us how extreme our observed agreement data is when considering all the possible agreement outcomes that could have occurred. Similar to the principle of Fisher's exact test [13], we reject the null hypothesis H_0 at the p level (*e.g.*, p=.05) if this probability is smaller than p:

$$\text{Reject } H_0 \text{ if } \Pi_{a_1,a_2,..,a_k|n_1,n_2,...n_k} < p \qquad (9)$$

which says that if the null hypothesis were true, agreement observations of this type would be highly exceptional. For the previous example, $\Pi_{10,6|21,21} = .0039$ and we will reject the null hypothesis that agreement rates .476 and .285 for the two groups of size 7 are nonsignificantly different. The V_b statistic is $V_{b(2,N=42)}$=8.000 and can be interpreted as the size of the effect, in the case in which a significant difference was detected. The number of degrees of freedom is k=2 because both a_1 and a_2 are free to vary, while we only constrain the numbers of pairs n_1 and n_2 to be the same across different samplings of agreement.

SOFTWARE SIMULATION

To evaluate the accuracy of the new V_b test statistic, we ran a simulation procedure by repeatedly generating populations (of size 100) of various controlled $\mathcal{AR}$ rates (from .100 to .900), from which we repeatedly drew two samples of equal size ($|P| = 20$) and we computed the number of Type I errors for $p = .05$, .01, and .001; see Table 4. After 9×10^5 simulation runs, we found that the average number of Type I errors was close to the p value for $\mathcal{AR} < .200$ and at least one order of magnitude below the p value for $\mathcal{AR} > .200$; it becomes nearly zero for $\mathcal{AR} > 300$ and $p < .01$. These results show that the predictions of V_b are accurate in general and very accurate for agreement rates of practical significance.

p	Agreement rate (average value, simulated over 9×10^5 runs)								
	.100	.200	.300	.400	.500	.600	.700	.800	.900
.050	0.190	0.041	0.004	0.000	0.000	0.003	0.001	0.002	0.000
.010	0.103	0.017	0.002	0.000	0.000	0.000	0.000	0.000	0.000
.001	0.050	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000

Table 4: Average number of Type I errors produced by the V_b test statistic compared to p values. Note how the number of errors is one order of magnitude below p for $\mathcal{AR} > .200$.

COAGREEMENT BETWEEN INDEPENDENT GROUPS

Each individual agreement rate (eq. 1) captures how much consensus there is within its group but, considered alone, cannot describe the consensus *between* groups. For example, consider three independent groups of participants, for which the following proposals were elicited in response to some referent: G_1={♣, ❖, ♣, ♥, ♣, ♣}, G_2={❤, ❤, ♠, ❤, ❤, ♦}, and G_3={❖, ♣, ❖, ♣, ♣, ❖}. The agreement rates for every group are numerically identical, *i.e.*, $\mathcal{AR}_1$=$\mathcal{AR}_2$=$\mathcal{AR}_3$=.400, but looking at participants' actual proposals, there is clearly more agreement between participants from groups G_1 and G_3, whom all suggested proposals ♣ and ❖, compared to the proposals ❤ and ♠ preferred by the participants of group G_2. Consequently, the use of the agreement rate formula alone is not always reflective of between-group agreement behavior: although agreement may be reached numerically, the actual preferences of the independent groups may vary. To capture such behavior accurately, we introduce the *between-group coagreement rate*:

$$\mathcal{CR}_b(G_1, G_2, ...G_k) = \frac{\sum_{i=1}^{k} \sum_{j=i+1}^{k} \sum_{p=1}^{|G_i|} \sum_{q=1}^{|G_j|} \delta_{p,q}}{\sum_{i=1}^{k} \sum_{j=i+1}^{k} |G_i| \cdot |G_j|} \qquad (10)$$

where $\delta_{p,q}$ is Kronecker's notation from eq. 2 that evaluates to either 1 or 0 depending whether participants p and q are in agreement or not, and the sum goes for all pairs of participants selected from all pairs of groups G_i and G_j,

$1 \leq i < j \leq k$. For example, the between-group coagreement rate for the $k = 3$ groups from the example above is $\mathcal{CR}_b(G_1, G_2, G_3) = \frac{0+15+0}{6 \cdot 6+6 \cdot 6+6 \cdot 6} = .139$, which is low for the reasons discussed above, but the coagreement between groups G_1 and G_3 is much higher, *i.e.*, $\mathcal{CR}_b(G_1, G_3) = \frac{15}{6 \cdot 6} = .417$. The between-group coagreement rate can be used by the practitioner in conjunction with individual agreement rates and the V_b statistic to further investigate causes of differences in agreement between groups; the Case Studies section of this paper illustrates the joint use of agreement and coagreement.

SOFTWARE TOOL: AGATE 2.0

To make computation of between-subject agreement rates and p values easy, we provide the AGATE 2.0 tool (AGreement Analysis Toolkit). The tool is implemented in C# .NET 4.5, and is freely available to download at `http://depts.washington.edu/aimgroup/proj/dollar/agate.html`.

Because the computation of the exact p value requires generation of all possible configurations of agreement for the k groups, the complexity of the algorithm implementing the test is exponential with power k. Knowing that a_i can take the maximum value n_i, we can estimate the upper margin of this complexity to be $O(\prod_{i=1}^{k} n_i) = O(\prod_{i=1}^{k} |G_i|^2)$. For the case in which all the groups have the same number of participants $|G|$, the complexity of running our test is $O(|G|^{2k})$. Actual time measurements showed that the test completes immediately for k=2 groups of 20 participants each (which is the maximum size employed for gesture elicitation studies so far), it takes 6 seconds for k=3 groups of 20 participants each (*i.e.*, 60 participants in total), and 10 minutes for k=4 (80 total participants). Time measurements were performed on a 2.4 GHz Intel Core 2 Quad CPU running Windows 7 on 32 bits and 2 GB RAM. These results show that the time performance of our toolkit is reasonable for practical scenarios, despite the exponential complexity required by the algorithm in the general case. We estimate that most between-subjects designs will consist of k=2 groups (*e.g.*, men versus women, novices versus experts, etc.), because more complex experimental designs generally require too many participants, *e.g.*, 80+ participants for k=4 groups. Nevertheless, for designs of that size, we expect that a Monte Carlo approach [11] will reduce the running time, an optimization that we leave for future work.

CASE STUDIES

In this section, we briefly re-examine data from several published elicitation studies [3,35,36,44,48] from the perspective of between-subjects analysis. We do not attempt to be comprehensive in our analysis, but instead we want to reveal the capability of our measures to unveil new discoveries for user-elicited data, not attainable prior to this work. We hope that researchers and practitioners will benefit from our case studies to inform their own data analysis for independent groups.

The effect of users' technical background on agreement

Previous elicitation studies have shown that users' prior experience with technology (*e.g.*, with Windows-like graphical user interfaces) affects their proposals for gesture commands [36,43,44,48,51], a phenomenon known as the "legacy bias" [31]. However, the lack of tools for agreement analysis has hindered rigorous examination of the significance and effect size of this phenomenon. In the following, we present the first analysis on the effect of users' technical expertise on agreement rate enabled by our statistics and using the dataset of the gesture elicitation study of Vatavu [44]. In that study, 20 participants (out of which 7 were non-technical) were asked to propose free-hand gestures to control 22 functions of a multi-screen TV system. Figure 2 shows the agreement rates[3] computed for each referent and each group of participants.

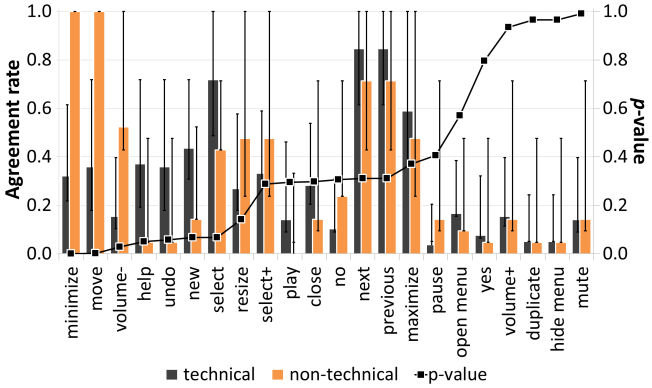

Figure 2: Agreement rates computed for technical and non-technical participants using the dataset of Vatavu [44]. NOTE: Referents are shown on the horizontal axis in ascending order of the exact p value of the V_b test; error bars show 95% CIs.

Using the V_b test, we found several differences in the agreement of the technical and non-technical groups. For example, there was more agreement for non-technical participants for *minimize* (1.000 versus .321, $V_{b(2,N=20)}$=252.784, p=.001), *move* (1.000 versus .359, $V_{b(2,N=20)}$=224.977, p=.001), and *volume down* (.524 versus .154, $V_{b(2,N=20)}$=74.938, p=.028), while technical participants achieved higher consensus for *help* (.372 versus .048, $V_{b(2,N=20)}$=57.537, p=.050) and *undo* (.359 versus .048, $V_{b(2,N=20)}$=53.076, p=.057); see Figure 2. To understand more about these differences, we computed between-group coagreement rates for each referent. For example, coagreement for *minimize* was $\mathcal{CR}_b$=.538, showing that only half (53.8%) of all pairs of participants across the two groups were in agreement about how to minimize a TV window, *i.e.*, by drawing hands together [44] (p. 202). The reason why the other half disagreed was that while all non-technical participants minimized content by drawing their hands together (corresponding to shrinking something in the real-world), the technical group proposed more variations, such as one and two-hand gestures, employed interaction metaphors, such as double click, and worked with an imaginary space outside the display [44] (p. 206). All these proposals elicited from the technical group indicate a clear influence of the legacy bias that was significant (p=.001) for *minimize* and exhibited the largest effect size (V_b=252.784) among all the 22 referents. The same effect was observed for *maximize*

[3]There is a difference in the magnitude of the agreement rates that we report and those from Vatavu [44] (pp. 196-197), because Vatavu used the original definition of agreement rate introduced by Wobbrock *et al.* [50], while here we employ its corrected version [47].

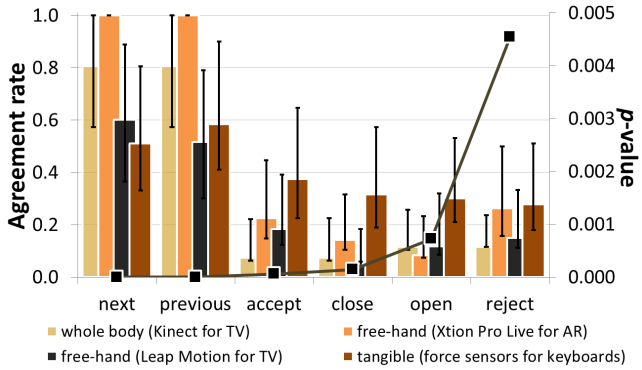

Figure 3: Agreement rates computed for 6 referents common to 4 independent studies with 78 participants [3,36,44,48]. NOTE: Referents are shown in ascending order of the exact p value of the V_b test; error bars show 95% CIs.

as well, but this time from another perspective: although the agreement rates of the two groups were similar (.590 versus .476) and the difference not significant ($V_{b(2,N=20)}$=7.060, *n.s.*), the coagreement rate showed different gesture preferences for the two groups ($\mathcal{CR}_b$=.549).

Similar findings emerged for other referents as well. For example, most of the technical participants drew the question mark symbol to invoke *help* ($\mathcal{AR}$=.372), which is a common icon in graphical user interfaces, while non-technical participants came up with significantly more different suggestions ($\mathcal{AR}$=.048), which were grounded in the non-technical real-world, *e.g.*, raise shoulders or raise arms in a shrug. This finding was also reflected by low a coagreement value for *help*, $\mathcal{CR}_b$=.198. On the other hand, referents with directional mappings received high coagreement, *i.e.*, $\mathcal{CR}_b$=.791 for *go to previous channel* and *go to next channel*, for which agreement rates were also similar (.846 versus .714, *n.s.*). We also found that for abstract referents, such as *open menu*, *hide menu*, *duplicate channel*, and *mute*, there were no significant differences between the agreement of the two groups. Coagreement was between .022 and .187, which shows the need for specific gesture designs for those referents, regardless of users' technical backgrounds.

These examples of between-group analysis illustrate how the original discussion of Vatavu [44], which was limited to only reporting qualitative differences between groups, could have been consolidated with numerical tools that evaluate significance and coagreement between groups quantitatively. We hope that our brief illustration of how to apply the between-group methodology will inspire researchers and practitioners to investigate the "legacy bias" phenomenon in more depth.

The effect of gesture application domain on agreement

Previous elicitation studies have employed various acquisition technologies to capture users' gestures, such as free-hand [35,36,48], accelerated motion [37], whole body [43,44], and touch gestures [51] for various application domains, such as appliance control [24,43,44,48], mobile device interaction [37], augmented keyboards [3], stroke-gesture alphabets [50], augmented reality [35,36], and generic touch interaction on tabletops [51]. Inadvertently, these studies have examined similar referents, such as *previous*, *next*, *accept*, *open*, *close*, etc. due to the application-independent nature of these commands. While analyzing these studies, we were able to find similar results reflective of generic user behavior, such as users falling back on already acquired interaction models or users assigning similar gestures for dichotomous tasks [36,37,43,44,51], but also many observations particularly related to the gesture technology or application context under evaluation, which suggests a potential effect of application domain on agreement. However, no work so far has investigated users' agreement over referents across application domains, one reason being the lack of tools to properly quantify and evaluate the significance of differences between independent groups of participants. In the following, we present the first analysis of such an effect by using our new V_b test statistic and 4 datasets with a total number of 78 participants: (1) the whole-body gesture dataset of Vatavu [44] for TV control (Microsoft Kinect, 20 participants, 22 referents), (2) the free-hand gesture dataset of Piumsomboon *et al.* [36] for augmented reality (Asus Xtion Pro Live, 20 participants, 40 referents), (3) the free-hand dataset of Vatavu and Zaiţi [48] for TV control (Leap Motion, 18 participants, 21 referents), and (4) the gesture dataset of Bailly *et al.* [3] for gesture-enhanced keyboards (Métamorphe keyboard, 20 participants, 42 referents). Figure 3 shows agreement rates of 6 referents that we found common to all these elicitation datasets together with the exact p-values of the V_b test.

Using the V_b test statistic, we found significant effects of the application domain on the agreement rates of all groups and all referents ($p \leq 0.0045$); see Figure 3. Follow-up post-hoc tests (Bonferroni corrected at $p = .05/6 = .0083$) revealed more precise differences. For instance, we found significantly more agreement for key gestures performed on the Métamorphe keyboard to invoke *accept* ($\mathcal{AR}$=.374) than for whole-body gestures ($\mathcal{AR}$=.074), free-hand gestures ($\mathcal{AR}$=.183) performed to control a Smart TV, and free-hand gestures elicited for augmented reality ($\mathcal{AR}$=.226), with exact p-values under .0033. This result shows that a tangible constraint, such as a key on a keyboard, affects significantly the number of potentially discoverable gestures, which results in higher agreement. There was no significant difference between the agreement rates of free-hand gestures for augmented reality and TV (.226 versus .183, $V_{b(2,N=38)}$=35.693, p=.279), which are acquisition scenarios that leverage similar gesture types; but the whole-body scenario led to significantly smaller agreement than hand gestures (p=.003 and p=.033). Both *next* and *previous* referents received maximum agreement (1.000) for free-hand gestures in the augmented reality context, significantly larger than all other groups (p<.001). However, we found no significant difference between free-hand gestures captured with Leap Motion and key gestures on the Métamorphe keyboard (.601 and .511, $V_{b(2,N=38)} = 118.392$, $n.s.$ for *next* and .516 and .584, $V_{b(2,N=38)} = 66.175$, $n.s.$ for *previous*). This non-significant result suggests that similar agreement levels may be reached across application domains for directional referents when gestures are elicited at the same scale of the body, *i.e.*, finger scale in this case. Note that results presented in this section do not come from a single experiment in which participants would

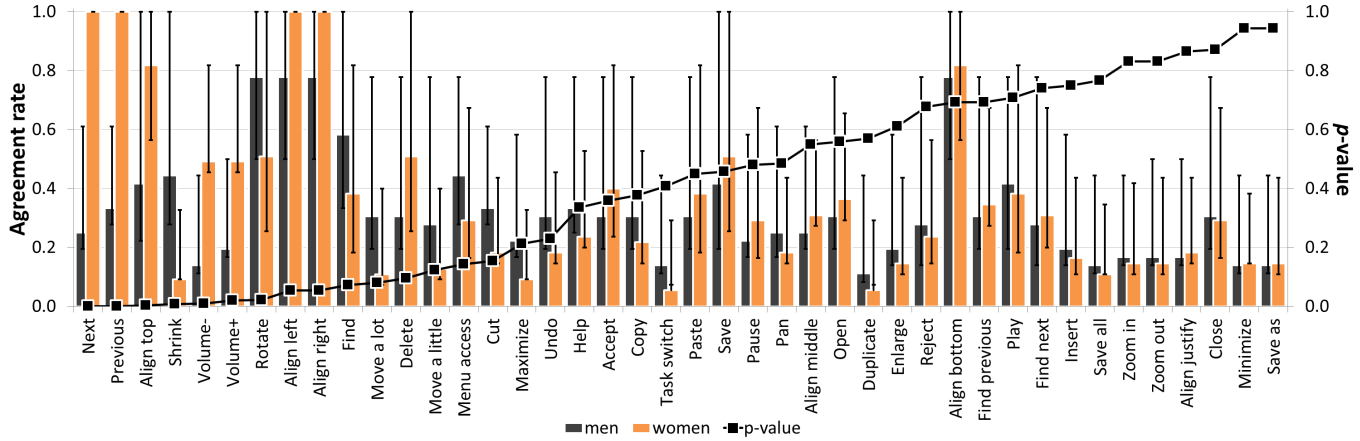

Figure 4: Agreement rates computed for male and female participants using the dataset of Bailly *et al.* [3]. NOTE: Referents are shown on the horizontal axis in ascending order of the exact p value of the V_b test; error bars show 95% CIs.

have been randomly allocated to the various devices, so further investigations in this direction are necessary to confirm our findings. However, we believe that such investigations could reveal even more aspects to understand potential *invariance of gesture commands* across application domains.

The effect of gender on agreement

We know from the literature of behavioral science that women generally express more non-verbal behavior than men in the form of smiles, laughs, and head and body movements [16]. We also know that there are cognitive differences between men and women, and that women generally exhibit more verbal fluency than men [34], while men perform better during visuo-spatial tasks [49]. Given that gestures and language are one system of mutual relationships expressing thought and that gestures have inherent visuo-spatial representations [28], it may be that women and men reach different levels of agreement in gesture elicitation studies. In the following, we present the first analysis of the effect of gender on agreement rates. We use the dataset of the elicitation study of Bailly *et al.* [3] in which 20 participants (9 males and 11 females) proposed key gestures for 42 referents for the Métamorphe keyboard.

Figure 4 shows the agreement rates reached by men and women for each referent. Overall, men and women reached similar levels of agreement (.322 and .353), but per-referent analysis showed significant differences ($p<.05$) for 7 referents. For instance, women reached significantly higher agreement than men for *next* (1.000 versus .250, $p=.00004$), *previous* (1.000 versus .333, $p=.00014$), *align top* (.818 versus .417, $p=.0038$), *increase volume* (.491 versus .194, $p=.0196$), and *decrease volume* (.491 versus .139, $p=.0085$). On the other hand, men were more in consensus for *shrink* (.444 versus .091. $p=.0077$) and *rotate* (.778 versus .509, $p=.0214$). Differences in agreement rates were marginally significant ($p=.0544$) for *align left* and *align right* (1.000 versus .778). These results show that women and men reach consensus over gestures in different ways that depend on the nature of the referent and that may be driven by different cognitive processes and capabilities to employ analogies of how things work. For instance, the perfect agreement for *next* and *previous* reached by women indicates very stable and predictable metaphors for referents of such semiotic nature. On the other hand, men were more in consensus for ergotic tasks, *e.g.*, *rotate* and *shrink*, that relate to the idea of work and mechanical modeling of the world. Furthermore, coagreement rates varied between a minimum of .111 (for *save as* and *task switch*) and a maximum of .889 (for *align left* and *align right*), which further highlights differences for some referents and agreement for others. Informed by these findings, the designer can now take a second, informed look at participants' proposals to understand what made them agree for some referents and disagree for others. While we only point to such differences to demonstrate the usefulness of our measures, we see benefits of further explorations of the types of gestures preferred by each gender and their corresponding levels of agreement that will lead to valuable understanding of *gender-related gesture preferences*, a topic under-examined so far by the gesture community.

CONCLUSION AND FUTURE WORK

We introduced in this paper new measures to evaluate differences in agreement between independent groups of participants and we showed their applicability to reveal new findings for various experimental conditions, findings unattainable prior to this work. Future work will consider a statistical test to evaluate whether the coagreement between independent groups is different from 0, which will enable the use of coagreement rate alone during analysis of significance. A new test for analyzing agreement data from elicitation experiments with mixed designs will prove useful for complex studies. However, the joint use of the V_b test for agreement rates and examinations of coagreement values can help designers today to examine differences between their groups of participants and use that knowledge to inform their current designs. It is our hope that this work will consolidate the practice of designing elicitation studies and analyzing agreement results, which will in turn inform improved user interface designs.

ACKNOWLEDGMENTS

The authors would like to thank Gilles Bailly and Thammathip Piumsomboon as well as their co-authors from [3,35,36] for kindly providing access to their gesture elicitation data. This work was supported from the project PN-II-RU-TE-2014-4-1187 financed by UEFISCDI, Romania.

REFERENCES

1. Alan Agresti. 2002. *Categorical Data Analysis, 2nd Ed.* USA: John Wiley & Sons.
2. Lisa Anthony, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2013. Understanding the Consistency of Users' Pen and Finger Stroke Gesture Articulation. In *Proc. of Graphics Interface 2013 (GI '13)*. Canadian Inf. Proc. Soc., Toronto, Ont., Canada, 87–94. http://dl.acm.org/citation.cfm?id=2532129.2532145
3. Gilles Bailly, Thomas Pietrzak, Jonathan Deber, and Daniel J. Wigdor. 2013. Métamorphe: Augmenting Hotkey Usage with Actuated Keys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 563–572. http://dx.doi.org/10.1145/2470654.2470734
4. Birgitta Bergvall-Kåreborn and Anna Ståhlbrost. 2008. Participatory Design: One Step Back or Two Steps Forward?. In *Proc. of the 10th Anniversary Conference on Participatory Design 2008 (PDC '08)*. 102–111. http://dl.acm.org/citation.cfm?id=1795234.1795249
5. Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement* 41, 3 (1981), 687–699. http://dx.doi.org/10.1177/001316448104100307
6. Ming Ki Chong and Hans W. Gellersen. 2013. How Groups of Users Associate Wireless Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1559–1568. http://dx.doi.org/10.1145/2470654.2466207
7. William G. Cochran. 1952. The χ^2 Test of Goodness of Fit. *Annals of Mathematical Statistics* 23, 3 (1952), 315–345. http://dx.doi.org/10.1214/aoms/1177729380
8. J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20 (April 1960), 37–46. http://dx.doi.org/10.1177/001316446002000104
9. Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (October 1968), 213–220. http://dx.doi.org/10.1037/h0026256
10. Noel Cressie and Timothy R.C. Read. 1989. Pearson's χ^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review. *International Statistical Review* 57, 1 (1989), 19–43. http://www.jstor.org/stable/1403582
11. Sonjoy Das, James C. Spall, and Roger Ghanem. 2010. Efficient Monte Carlo computation of Fisher information matrix using prior information. *Computational Statistics & Data Analysis* 54, 2 (2010), 272–289. http://dx.doi.org/10.1016/j.csda.2009.09.018
12. Leah Findlater, Ben Lee, and Jacob Wobbrock. 2012. Beyond QWERTY: Augmenting Touch Screen Keyboards with Multi-touch Gestures for Non-alphanumeric Input. In *Proc. of CHI '12*. ACM, 2679–2682. http://dx.doi.org/10.1145/2207676.2208660
13. Ronald A. Fisher. 1954. *Statistical Methods for Research Workers*. London: Oliver and Boyd.
14. Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (November 1971), 378–382. http://dx.doi.org/10.1037/h0031619
15. Kilem L. Gwet. 2012. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Advanced Analytics, LLC.
16. Judith A. Hall. 1990. *Nonverbal Sex Differences: Communication Accuracy and Expressive Style*. Johns Hopkins University Press.
17. Jesse Hoey. 2012. The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way χ^2 Test. *arXiv:1206.4881* (2012). http://arxiv.org/abs/1206.4881
18. Ole R. Holsti. 1969. *Content analysis for the social sciences and humanities*. Addison-Wesley.
19. Robert C. James. 1992. *The Mathematics Dictionary, 5th Ed.* Chapman & Hall, New York.
20. Shaun K. Kane, Jacob O. Wobbrock, and Richard E. Ladner. 2011. Usable Gestures for Blind People: Understanding Preference and Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 413–422. http://dx.doi.org/10.1145/1978942.1979001
21. Maurice G. Kendall and B. Babington Smith. 1939. The Problem of m Rankings. *Annals of Math. Stats.* 10, 3 (1939), 275–287. http://www.jstor.org/stable/2235668
22. Finn Kensing and Jeanette Blomberg. 1998. Participatory Design: Issues and Concerns. *Computer Supported Cooperative Work* 7, 3-4 (Jan. 1998), 167–185. DOI: http://dx.doi.org/10.1023/A:1008689307411
23. Christian Kray, Daniel Nesbitt, John Dawson, and Michael Rohs. 2010. User-defined Gestures for Connecting Mobile Phones, Public Displays, and Tabletops. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '10)*. ACM, New York, NY, USA, 239–248. http://dx.doi.org/10.1145/1851600.1851640
24. Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. 2011. I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies* 69, 11 (2011), 693–704. http://dx.doi.org/10.1016/j.ijhcs.2011.04.005
25. S. Kullback and R.A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* 22, 1 (1951), 79–86. http://www.jstor.org/stable/2236703
26. Ekaterina Kurdyukova, Matthias Redlin, and Elisabeth André. 2012. Studying User-defined iPad Gestures for

Interaction in Multi-display Environment. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, USA, 93–96. http://dx.doi.org/10.1145/2166966.2166984

27. H.O. Lancaster. 1961. Significance Tests in Discrete Distributions. *J. Amer. Statist. Assoc.* 56, 294 (1961), 223–234. http://www.jstor.org/stable/2282247
28. David McNeill. 1992. *Hand and Mind: What Gesture Reveals about Thought*. University Chicago Press.
29. Cyrus R. Mehta and Nitin R. Patel. 1983. A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables. *J. Amer. Statist. Assoc.* 78, 382 (1983), 427–434. http://www.jstor.org/stable/2288652
30. Meredith Ringel Morris. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces (ITS '12)*. ACM, New York, NY, USA, 95–104. http://dx.doi.org/10.1145/2396636.2396651
31. Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, m c schraefel, and Jacob O. Wobbrock. 2014. Reducing Legacy Bias in Gesture Elicitation Studies. *Interactions* 21, 3 (May 2014), 40–45. http://dx.doi.org/10.1145/2591689
32. Charles E. Osgood. 1959. Representational model and relevant research methods. In *Trends in Content Analysis*, I. Pool (Ed.). Illinois Press, Urbana, IL.
33. Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* 50, 5 (1900), 157–175. http://dx.doi.org/10.1080/14786440009463897
34. M. Pilar Matud, C. Rodrguez, and J. Grande. 2007. Gender differences in creative thinking. *Personality and Individual Differences* 43, 5 (2007), 1137–1147. http://dx.doi.org/10.1016/j.paid.2007.03.006
35. Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013a. User-Defined Gestures for Augmented Reality. In *Human-Computer Interaction INTERACT 2013 (Lecture Notes in Computer Science)*, Paula Kotz, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.), Vol. 8118. Springer Berlin Heidelberg, 282–299. http://dx.doi.org/10.1007/978-3-642-40480-1_18
36. Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013b. User-defined Gestures for Augmented Reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 955–960. http://dx.doi.org/10.1145/2468356.2468527
37. Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined Motion Gestures for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 197–206. http://dx.doi.org/10.1145/1978942.1978971
38. Jaime Ruiz and Daniel Vogel. 2015. Using Soft-Constraints to Reduce Legacy and Performance Bias in Gesture Elicitation Studies. In *Proceedings of CHI'15, the 33rd ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3347–3350. http://dx.doi.org/10.1145/2702123.2702583
39. Karen Rust, Meethu Malu, Lisa Anthony, and Leah Findlater. 2014. Understanding Childdefined Gestures and Children's Mental Models for Touchscreen Tabletop Interaction. In *Proceedings of the 2014 Conference on Interaction Design and Children (IDC '14)*. ACM, New York, NY, USA, 201–204. http://dx.doi.org/10.1145/2593968.2610452
40. William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 19, 3 (1955), 321–325. http://dx.doi.org/10.1086/266577
41. Teddy Seyed, Chris Burns, Mario Costa Sousa, Frank Maurer, and Anthony Tang. 2012. Eliciting Usable Gestures for Multi-display Environments. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces (ITS '12)*. ACM, New York, NY, USA, 41–50. http://dx.doi.org/10.1145/2396636.2396643
42. Consuelo Valdes, Diana Eastman, Casey Grote, Shantanu Thatte, Orit Shaer, Ali Mazalek, Brygg Ullmer, and Miriam K. Konkel. 2014. Exploring the Design Space of Gestural Interaction with Active Tokens Through User-defined Gestures. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 4107–4116. http://dx.doi.org/10.1145/2556288.2557373
43. Radu-Daniel Vatavu. 2012. User-defined Gestures for Free-hand TV Control. In *Proceedings of the 10th European Conference on Interactive TV and Video (EuroiTV '12)*. ACM, New York, NY, USA, 45–48. http://dx.doi.org/10.1145/2325616.2325626
44. Radu-Daniel Vatavu. 2013. 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. http://dx.doi.org/10.3233/AIS-130200
45. Radu-Daniel Vatavu, Lisa Anthony, and Quincy Brown. 2015a. Child or Adult? Inferring Smartphone Users' Age Group from Touch Measurements Alone. In *Proceedings of the 15th IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT '15)*. 1–9. http://dx.doi.org/10.1007/978-3-319-22723-8_1
46. Radu-Daniel Vatavu, Gabriel Cramariuc, and Doina Maria Schipor. 2015b. Touch Interaction for Children Aged 3 to 6 Years: Experimental Findings and Relationship to Motor Skills. *International Journal of*

Human-Computer Studies 74 (2015), 54–76. http://dx.doi.org/10.1016/j.ijhcs.2014.10.007

47. Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of CHI'15, the 33rd ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1325–1334. http://dx.doi.org/10.1145/2702123.2702223
48. Radu-Daniel Vatavu and Ionut-Alexandru Zaiti. 2014. Leap Gestures for TV: Insights from an Elicitation Study. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '14)*. ACM, New York, NY, USA, 131–138. http://dx.doi.org/10.1145/2602299.2602316
49. Tomaso Vecchi and Luisa Girelli. 1998. Gender differences in visuo-spatial processing: The importance of distinguishing between passive storage and active manipulation. *Acta Psychologica* 99, 1 (1998), 1–16. http://dx.doi.org/10.1016/S0001-6918(97)00052-8
50. Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1869–1872. http://dx.doi.org/10.1145/1056808.1057043
51. Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1083–1092. http://dx.doi.org/10.1145/1518701.1518866
52. Huiyue Wu and Jianmin Wang. 2012. User-Defined Body Gestures for TV-based Applications. In *Proceedings of the 2012 Fourth International Conference on Digital Home (ICDH '12)*. IEEE Computer Society, Washington, DC, USA, 415–420. http://dx.doi.org/10.1109/ICDH.2012.23