Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies

RADU-DANIEL VATAVU, MintViz Lab, MANSiD Center, Ştefan cel Mare University of Suceava, Romania
JACOB O. WOBBROCK, The Information School | DUB Group, University of Washington, USA

We clarify fundamental aspects of end-user elicitation, enabling such studies to be run and analyzed with confidence, correctness, and scientific rigor. To this end, our contributions are multifold. We introduce a formal model of end-user elicitation in HCI and identify three types of agreement analysis: *expert*, *codebook*, and *computer*. We show that agreement is a mathematical *tolerance relation* generating a tolerance space over the set of elicited proposals. We review current measures of agreement and show that all can be computed from an *agreement graph*. In response to recent criticisms, we show that chance agreement represents an issue solely for inter-rater reliability studies and not for end-user elicitation, where it is opposed by *chance disagreement*. We conduct extensive simulations of 16 statistical tests for agreement rates, and report Type I errors and power. Based on our findings, we provide recommendations for practitioners and introduce a five-level hierarchy for elicitation studies.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; User studies;

Additional Key Words and Phrases: End-user elicitation method, gesture elicitation, measures of agreement, consensus, tolerance relations, tolerance spaces, dissimilarity, statistical tests, simulations, inter-rater reliability, Type I error, statistical power

ACM Reference format:
Radu-Daniel Vatavu and Jacob O. Wobbrock. 2022. Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies. *ACM Trans. Comput.-Hum. Interact.* 29, 1, Article 5 (January 2022), 70 pages.
https://doi.org/10.1145/3476101

5

R.-D. Vatavu acknowledges support from a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number PN-III-P4-ID-PCE-2020-0434 (PCE29/2021), within PNCDI III. Original versions of the "user/brain" icons used in Figures 1, 6, and 10 were made by rawpixel.com from https://www.freepik.com ("Brain and mental health icons vector set" pack, https://www.freepik.com/free-vector/brain-mental-health-icons-vector-set_3438021.htm) released under the Freepik license, free for personal and commercial purpose with attribution. The images from Figure 8, top represent snapshots taken from the videos of the memorability gesture dataset [38, 80], freely available to download from https://udigesturesdataset.cs.st-andrews.ac.uk/. The hand pose illustrations used in Figure 3 were provided by GestureWorks® (https://gestureworks.com/), released under a Creative Commons Attribution Sharealike license.

Authors' addresses: R.-D. Vatavu, MintViz Lab, MANSiD Center, Ştefan cel Mare University of Suceava, 13 Universităţii, Suceava 720229, Romania; email: radu.vatavu@usm.ro; J. O. Wobbrock, The Information School | DUB Group, University of Washington, Box 352840, Seattle, WA 98195-2840; email: wobbrock@uw.edu.

1073-0516/2022/01-ART5 $15.00
https://doi.org/10.1145/3476101

1 INTRODUCTION

End-user elicitation studies, in which the results of a system function (e.g., zoom in on a map) are demonstrated, and the user is asked for the *action* (e.g., gesture, voice command) or *symbol* (e.g., icon, button label, command-line term) that would bring about that result, have become exceptionally popular since their advent in 2005 by Wobbrock et al. [140] with the initial phrasing of "guessability studies." Gesture elicitation, in particular, has been the most popular focus of end-user elicitation studies, with Wobbrock et al.'s 2009 paper [141] about uncovering users' preferences for hand gesture input on tabletops being cited nearly 1,300 times, and the original method [140] replicated nearly 300 times; see [57, 69, 120, 124, 127] for overviews of published gesture elicitation studies and their results. A core component of end-user elicitation is the notion of *agreement*, which indicates when actions or symbols proposed by the study participants are, essentially, *the same* or at least *substantially similar* [3]. Agreement is vital both in the grouping of similar actions or symbols and in the calculation of an *agreement score* that quantifies how much agreement participants have exhibited in their proposals. If conceptualized and calculated incorrectly, agreement can mislead designers, system creators, and usability specialists [114] and, ultimately, bring about poor user interface and interactive system designs. Miscalculated agreement could also lead researchers interested in understanding human behavior with interactive technology to draw false conclusions about the proclivities and preferences of end users of interactive systems.

Recently, established measures of agreement in end-user elicitation, such as the agreement score A [140, 141] and agreement rate AR [32, 120, 121], have come under fire for failing to take into account agreement occurring by chance. Statistical inference tests proposed for elicitation data [120, 121] have been criticized as well. An article by Tsandilas [108], especially, has called into question these established agreement calculations by connecting to the literature and practice of inter-rater reliability studies [21, 22, 34, 40, 59], and called out large Type I error rates for analogous statistical tests [120, 121] by modeling bias in gesture elicitation according to the premises and assumptions employed in inter-rater reliability. Other authors have highlighted concerns for end-user elicitation, such as the problem of legacy bias that may cause elicitation studies to get caught in local minima [43, 76], or problems getting reproducible results with the original elicitation method [82, 114]. To add to these, the theoretical landscape of end-user elicitation studies has been rapidly changing due to new formalizations, algorithms and software tools, and variations of the initial method [2–4, 43, 76, 114].

Consequently, *the body of work on end-user elicitation studies can appear confusing and even conflicting* for researchers and practitioners that wish to apply this method to inform the design of their user interfaces, devices, prototypes, and interactive systems, *but also to educators disseminating this method to students.*[1] Against this backdrop, our work attempts to clarify fundamental aspects of end-user elicitation by providing (1) examined perspectives, (2) supporting mathematical theory and results, (3) connections to other fields, and (4) recommendations for conducting *general* end-user elicitation studies in HCI, which subsume popular gesture elicitation studies [140, 141]. To this end, we *alleviate and clarify* **specific concerns** (**SC**) [76, 82, 103, 108], *unify* recent complementary perspectives [3, 114], and *provide a theoretically sound foundation* for end-user elicitation. For example, our close examinations reveal that Tsandilas' [108] concerns about chance agreement in end-user elicitation studies are assuaged by *proportional chance disagreement*, an aspect overlooked in Tsandilas' work, which, we argue, focused on and was overly influenced by the theory and practice of inter-rater reliability studies. This omission prevented Tsandilas from observing the

[1] Such as the "3D User Interfaces" Computer Science course delivered at Colorado State University, Fall 2019, which includes gesture elicitation as part of its syllabus; see https://www.online.colostate.edu/courses/CS/CS567.dot.

subtle, yet key differences between inter-rater reliability and end-user elicitation that, upon close examination, are based on fundamentally different assumptions, as we show. Consequently, end-user elicitation studies need distinct models, methods, measures, and tools for agreement calculation and analysis. To this end, we present both theoretical arguments and practical evidence from experiments conducted on multiple public gesture datasets. In a similar manner, we debunk and clarify other concerns as well. But first, we clearly articulate the SC that researchers and practitioners, and especially newcomers to the method, are likely to find conflicting and potentially a barrier to applying end-user elicitation in their own work. Then, we identify **research questions** (**RQ**) and corresponding practical aspects for conducting end-user elicitation studies, for which we provide corresponding clarifications.

1.1 SC from the Scientific Literature of End-User Elicitation That This Article Clarifies

We outline, in chronological order, seven *Specific Concerns* ([SC_1]–[SC_7]) that were formulated in the literature [76, 82, 103, 108, 114] regarding various practical aspects of conducting user studies with the elicitation method [140, 141] and its variations for calculating and analyzing agreement [32, 120, 121], which we believe are important to clarify for researchers and practitioners wishing to employ this method in their own work:

- [SC_1] Stern et al. [2008] [103]: claim that eliciting proposals by having participants actually performing them, as proposed in Wobbrock et al. [140, 141], may be a less suited approach compared to other ways to elicit end users' preferences for actions, commands, or symbols, such as the "coded gesture entry" method.
- [SC_2] Nebeling et al. [2014] [82]: claim that the end-user elicitation method should be extended toward reproducible and implementable user-defined interaction sets.
- [SC_3] Morris et al. [2014] [76]: claim that legacy bias, i.e., the potential pitfall of users' proposals to be biased by their experience with prior interfaces and technologies, is a limitation of the original end-user elicitation method [140, 141].
- [SC_4] Tsandilas [2018] [108]: claims that the established measures of agreement calculation, A and AR, advocated by Wobbrock et al. [140, 141], Findlater et al. [32], and Vatavu and Wobbrock [120, 121], do not take into account chance agreement.
- [SC_5] Tsandilas [2018] [108]: claims that the guidelines proposed by Vatavu and Wobbrock [120] for interpreting the magnitude of agreement can lead to overoptimistic conclusions about the true level of agreement reached by the participants of end-user elicitation studies.
- [SC_6] Tsandilas [2018] [108]: claims that the V_{rd} and V_b test statistics proposed by Vatavu and Wobbrock [120, 121] yield high Type I error rates.
- [SC_7] Vatavu [2019] [114]: claims that the criteria used to evaluate the similarity of proposals elicited from the participants of end-user elicitation studies can make the magnitude of agreement scores irrelevant, because of the dependency between agreement and the criteria employed. Instead, a holistic approach in which agreement is interpreted as a function of the criteria used to judge the similarity of elicited proposals should be preferred to using specific, possibly subjective criteria.

For some of these SCs, the literature already contains potential improvements on the end-user elicitation method, such as ways to reduce legacy bias [43, 76, 94] and improve the learnability and memorability of user-elicited input [2], measures of agreement that are independent of the criteria used to cluster end-users' proposals according to their similarity to each other [114, 115], or new procedures to perform statistical inference tests for elicitation data [108]. Other concerns,

such as chance agreement potentially occurring in end-user elicitation studies [108], different possible ways to elicit proposals from participants [103], and aspects regarding the reproducibility of end-user elicitation studies [82, 114], are still open, creating a state of uncertainty for practitioners that wish to apply the end-user elicitation method in their own work. Therefore, it is important to clarify such concerns. To this end, *our article provides a close re-examination of agreement calculation and analysis as a core component of end-user elicitation studies, offering numerous contributions, unifications of formulae and current practices, connections to other fields, and many clarifications for researchers and practitioners*. We start by outlining a series of fundamental RQ and corresponding practical aspects for end-user elicitation to which we will refer in the rest of this article.

1.2 Fundamental RQ for End-User Elicitation

We outline four fundamental *Research Questions* ([RQ_1]–[RQ_4]) important for the theoretical foundation and further methodological development of the end-user elicitation method. Together, they subsume nine practical aspects with which researchers, designers, and practitioners are likely to be confronted when running end-user elicitation studies:

[RQ_1] How do end-user elicitation studies compare to inter-rater reliability studies? The following practical aspects are subsumed:

- [$RQ_{1.1}$] *Should the measures of agreement employed in end-user elicitation studies, such as A and AR, be corrected for chance agreement, just like in inter-rater reliability studies? If so, how?*
- [$RQ_{1.2}$] *Is end-user elicitation the same thing as an inter-rater reliability study?*

[RQ_2] What is agreement in end-user elicitation? With the following practical aspects:

- [$RQ_{2.1}$] *How do various measures of agreement relate to each other?*
- [$RQ_{2.2}$] *Which measure(s) of agreement should one use for end-user elicitation studies?*
- [$RQ_{2.3}$] *How to interpret the magnitude of agreement in end-user elicitation studies?*

[RQ_3] Can end-user elicitation be modeled formally? In particular:

- [$RQ_{3.1}$] *Are there viable models for the analysis of elicited proposals?*
- [$RQ_{3.2}$] *Which model should one adopt for the analysis of elicited proposals?*

[RQ_4] What statistical procedures best apply to elicitation data? The following practical aspects are subsumed:

- [$RQ_{4.1}$] *Which statistical test should one use for analyzing agreement data for end-user elicitation studies with between-subjects experimental designs?*
- [$RQ_{4.2}$] *Which statistical test should one use for analyzing agreement data for within-subjects experimental designs?*

1.3 Research Contributions

In this work, we address all of the fundamental RQ outlined above, and clarify all of the specific concerns discussed at the outset. To this end, our article offers many contributions, both theoretical and practical, to end-user elicitation:

(1) We introduce a *formal operational model for end-user elicitation studies* in HCI, which include the popular gesture elicitation studies [140, 141], representing the most comprehensive description of general end-user elicitation provided to date. In this context, we identify three distinct models of agreement analysis: the *expert*, *codebook*, and *computer* models, for which we describe numerical procedures to characterize bias based on the Zipf–Mandelbrot, Bernoulli, and Gaussian distributions and using the Minkowski, squared Euclidean, and Lin dissimilarity functions, thus covering all known variants of end-user elicitation studies.

(2) We show that the notion of agreement employed in elicitation studies can be formalized mathematically as a *tolerance relation* [145] that generates a *tolerance space* [102, 145] over the set of proposals elicited from end users. To this end, the concepts of dissimilarity functions [114], classification [114], and clustering [3] are key to agreement calculation. Furthermore, we show how formalizing agreement calculation with dissimilarity functions and tolerances connects directly to the complete-link method [130], a hierarchical clustering method known for the internal cohesion of its partition and homogeneous clusters.

(3) We review current measures for calculating agreement in end-user elicitation studies, which we evaluate from the perspective of four quality properties. We determine that all those measures can be computed from the *agreement graph*, a concept that we employ to (i) show how *those measures are facets of one single, all-purpose agreement rate*, and (ii) to deliver a key insight about *AR* [120], one of the most popular measures of agreement in elicitation studies, which camouflages as a *measure of central location (the mean)*.

(4) In response to Tsandilas' [108] criticism, we show that chance agreement represents an issue solely for inter-rater reliability studies [21, 22, 34, 40] and not for end-user elicitation, where it is proportionally opposed by "chance disagreement," i.e., for every bit of chance agreement, there is a corresponding amount of chance disagreement to oppose it. We use the concepts of *false positives* and *false negatives* to properly describe and quantify this aspect.

(5) We conduct extensive simulations of 16 statistical inference tests for within- and between-subjects designs for end-user elicitation studies, and report estimations of Type I error rates and statistical power. To this end, we introduce new Monte Carlo procedures that can simulate populations of *exact* agreement rates, an unprecedented level of simulation accuracy in end-user elicitation [108, 121]. We use these simulations to clarify recent concerns [108] regarding the V_{rd} [120] and V_b [121] test statistics.

(6) We provide recommendations for practitioners of end-user elicitation by encouraging numerical representation and acquisition of proposals elicited from end users, and we introduce *a five-level hierarchy for end-user elicitation studies* by considering aspects of recording data computationally, open data, open software, reproducibility, and validation of results. We also distill our theoretical elaborations and empirical findings into readily applicable guidelines for researchers and practitioners regarding what measures of agreement and statistical inference tests to use in their own end-user elicitation studies. The next subsection summarizes these practical guidelines.

1.4 Practical Guidelines for Conducting End-User Elicitation Studies

Our contributions lead to a number of guidelines for the practice of end-user elicitation studies, which are discussed at length in Section 10. A summary is given below:

(1) End-user elicitation studies following the original method [140, 141] are fundamentally different from inter-rater reliability studies: The list of categories is not defined *a priori* in end-user elicitation, the agreement relation is not necessarily transitive, and the measures of agreement *A* [140, 141] and *AR* [32, 120, 121] should not be corrected for chance agreement [108]. These differences are discussed in detail in Section 4, while the distinctive properties of the agreement relation in end-user elicitation are scrutinized in Sections 3 and 5.

(2) The traditional measures to evaluate agreement in end-user elicitation, *A* [140, 141] and *AR* [32, 120, 121], deliver the same ranking order of referents and compute both in the unit interval. Thus, they are interchangeable for analysis purposes, but *AR* conveniently evaluates to 0 when there is no agreement, whereas *A* evaluates to a number greater than zero

dependent upon the number of elicited proposals. The AR_ϵ measure, introduced in this work following [114], encapsulates A and AR under one mathematical formulation; see Sections 6 and 7.

(3) There are three possible models for evaluating the results of end-user elicitation studies: the *expert*, the *codebook*, and the *computer* model, which we discuss in Section 8. Of these, the *codebook* model has been the most used in published elicitation studies. We recommend using the *computer* model whenever possible for reasons of efficiency and replicability of results, but also due to straightforward transfer of study results to actual systems. Our new AR_ϵ measure is compatible with the *computer* model for automated agreement analysis.

(4) We recommend the percentile bootstrap [132, pp. 332–335] statistical test for the analysis of agreement in end-user elicitation studies with between-subjects experimental designs, which seems to control the Type I error rate very well under a variety of testing conditions; details follow in Section 9.1.

(5) We also recommend the percentile bootstrap [132, p. 411] statistical test for end-user elicitation studies using within-subjects experimental designs; details follow in Section 9.2.

To assist practitioners in adopting these guidelines, we provide open-source R code implementing measures of agreement and statistical tests; see details in Section 11.

2 FUNDAMENTALS OF END-USER ELICITATION STUDIES

We review the procedure of conducting end-user elicitation studies [140, 141] to clarify the steps involved, their theoretical support, and key issues for agreement calculation. We use this opportunity to propose an operational model of general end-user elicitation in HCI, representing the most comprehensive and clarifying description to date regarding how end-user elicitation studies work.

2.1 From Participatory Design and Maximizing Guessability of Symbolic Input to Gesture Elicitation Studies

Participatory Design is the practice where designers and users work together to improve the quality of working life, often through technology design [37, 41, 78, 97]. Applied to HCI [125, 126], participatory design aims at informing and improving the features of interactive systems and user interfaces by involving end-users in the early stages of the design process. In this context, Wobbrock et al. [140] were interested in the "guessability" of symbolic input as *"that quality of symbols which allows a user to access intended referents via those symbols despite a lack of knowledge of those symbols"* (p. 1869). To that end, they proposed a procedure to maximize guessability, in their case of stroke gestures for letters and numbers, by asking end users to propose symbols, based on which agreement was calculated for individual referents. This procedure was then applied by Wobbrock et al. [141] to understand users' preferences for touch gestures on interactive tabletops, the first hand-gesture elicitation study. Since then, gesture elicitation has been growing in popularity, being applied to a variety of contexts of use and applications of gesture input [9, 26, 28, 32, 36, 73, 75, 82, 87, 91, 93, 114, 117, 123], with over 300 studies published to date [124].[2] In these studies, participants propose "gestural signs" in response to "referents" representing actions and system functions, or user behaviors and their effects, respectively [141].

2.2 Toward Formalizing the Steps of a Gesture Elicitation Study

Tsandilas [108] provides the first mathematical formalization of the overall procedure involved in running a gesture elicitation study, highlighting three main steps: (1) *recording* gesture proposals

[2]Papers available in ACM DL, IEEExplore, Scopus, or ScienceDirect.

from participants via notes, videos, logs, or actual digital representations provided by a gesture acquisition device; (2) *classification*, where gesture descriptions are interpreted, either automatically by a computer or manually by the experimenter, into a set of signs; and (3) *agreement analysis* enabled by the resulting sign vocabulary. This formalization is articulate in the distinction that it makes between *gesture descriptions* (i.e., the actual representations of elicited gestures, such as a set of touch points delivered by a smartphone in response to the user's finger drawing a circle on the touchscreen) and *signs* (i.e., the interpretation of the gesture description for the assignment of a label, e.g., "circle" or "swipe left").

This three-step description of a gesture elicitation study is useful, but leaves out key aspects regarding the formation of agreement and implementation of the classification step. The cause of these omissions lies in the fact that some of those missing key aspects were not available when Tsandilas [108] introduced his formalization [2018], as they were revealed only recently by two advances in gesture elicitation methods and tools: (i) the dissimilarity–consensus method for end-user agreement analysis [114] and (ii) new algorithmic approaches to compute agreement based on unsupervised machine learning [3]. The notion of a *dissimilarity measure* [114] and the use of *unsupervised classification* [3] are especially important to clarify and complete the formal description of what elicitation studies are and how they should be conducted for best results. In the following, we provide an improved description of end-user elicitation by introducing an *operational model* with six components that formalize the steps for conducting such studies, from the presentation of referents to the classification of elicited proposals into signs.

2.3 From Gesture Studies to an Operational Model of General End-User Elicitation for Human–Computer Interaction

Tsandilas [108] makes an important distinction between the *description* of a gesture and the *sign* assigned to that gesture as the result of the classification step. This distinction is helpful to understand why two gestures, despite inherent variation in their articulations (e.g., a small and a large circle, or a clockwise and a counter-clockwise circle), may be assigned to the same sign (i.e., the "circle" sign) and, thus, be considered in agreement.

However, there are more aspects at work than first meet the eye in the process leading from referent presentation to the sign assignment for elicited gestures or, in general, for elicited proposals. In the following, we extend Tsandilas' [108] distinction between descriptions and signs into a complete model of the elicitation process. Figure 1 illustrates our operational model for *general* end-user elicitation studies in HCI, highlighting the following steps and components:

(1) Upon the presentation of a *referent*, the participant, through reflection and deliberation, forms a *mental model of the system effect* corresponding to that referent. For example, upon witnessing a map displayed on a tabletop zoom in, the user forms a mental model of zooming, recognizing that the effect on the system of some hypothetical command is to make objects on the map look bigger with greater detail.

(2) The mental model of the system effect is instantiated into a *mental model of the command* that the participant is asked to propose in order to effect the referent just demonstrated.[3] For example, a two-finger model for zooming in and out might be instantiated into either one hand touching the screen (which represents a mental model of single-hand input) or two hands being used at the same time (bimanual coordinated input).

[3]By "command," we mean an *action*, such as a gesture or voice command, or a *symbol*, such as an icon, button label, command-line term, or hyperlink.

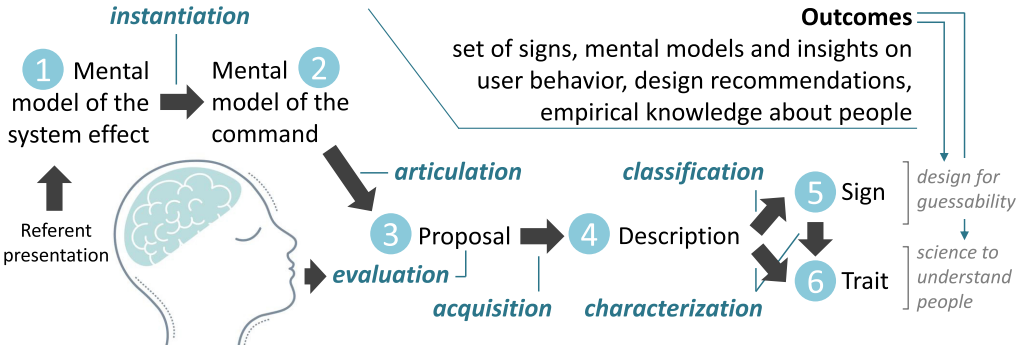

Fig. 1. A model for general end-user elicitation studies in HCI, including gesture elicitation studies [141]: referents presented to participants trigger creation of mental models of the interaction (1) that are instantiated into mental models of the command (2), which is articulated into a specific form of a proposal (3) captured by an actual device or via the experimenter's notes and video recordings (4), and finally grouped and classified into a sign (5) when the goal is the identification of a consensus set of signs and/or compared against other descriptions for characterization purposes (6) when the goal is understanding people.

(3) The command is articulated into an actual *proposal*. Let us assume that the participant chose to zoom by pinching two fingers on one hand. The articulation of the command may take the specific form of using the thumb and the index finger to perform zoom out or even all five of the fingers touching the tabletop. Usually, participants are also asked to evaluate various qualities of their proposals, such as how easy it was to perform a gesture or how well the gesture matched the presented referent [141].

(4) The proposal articulated by the participant is captured by some acquisition device, such as the tabletop in our example, or is logged by the experimenter using notes and video recordings. The result is a *description* of the proposal.

(5) A classifier, including perhaps a human experimenter, distills a set of *signs* based on the similarity and dissimilarity among the descriptions of all elicited proposals. Proposals are grouped, and canonical representations of each group are held up as their *sign*. A sign can be thought of as a class label, where individual proposals are instances of that class.

(6) Following, complementary to, or independent of step (5), when the goal of the end-user elicitation study is not *just* the compilation of a consensus set of signs, a *characterization procedure* is used to compare descriptions of the elicited proposals either with each other, e.g., to quantify the variance, diversity, or consistency of participants' proposals, or against canonical forms in order to understand differences between the elicited proposals and the status quo, such as user-defined *vs.* designer-defined interactions.

This six-step model is helpful to understand that variation in the elicited proposals and, consequently, in the magnitude of agreement between those proposals has various sources, ranging from different mental models of the system effect to instantiations of those models as commands, to different articulations of the proposals, different technologies, devices, and tools to capture the proposals, and finally different criteria to cluster together similar proposals. Steps 5 and 6 lead to the practical outcomes of any end-user elicitation study, including gesture elicitation studies [140, 141]: a set of selected, consensus-driven signs to inform design for guessability [140], the most common application of the method, and empirical results that tell us about how people use, are able to use, or would like to use an interactive system. Next, we provide a mathematical formalization of step 5, classification, which is key for determining agreement and identifying signs, and examine

step 6, characterization, by highlighting other, diverse purposes and goals of end-user elicitation studies. But first, we present an existing formalization from Tsandilas [108] regarding classification of proposals into signs and exemplify current practices employed to classify gestures as signs.

2.3.1 Previous Formalization of the Classification Step. Tsandilas [108] looks at the classification step of a gesture elicitation study from the perspective of *"a function C that takes as input a set of gesture descriptions* $\{g_{ij}\}$ *and produces a set of sign assignments* $\{g_{ij} \rightarrow \sigma_k \mid k = 1 \ldots q\}$*, such that each gesture description* g_{ij} *is assigned a sign* σ_k *that belongs to a sign vocabulary of size q"* (p. 18:4). This formulation is useful as it represents the starting point toward a mathematical formalization of the classification step. However, since no other details are provided in [108], this formulation remains too general and, thus, unclear in terms of how the classification step should be implemented.

For example, Tsandilas' function *C* could be implemented in the context of either *supervised* or *unsupervised classification*, the two main paradigms of pattern recognition [130]. In the former case, proposals are classified based on an existing set of signs, an approach that resembles how inter-rater reliability studies are conducted, where participants (called raters) select categories from a predefined list [21, 34]. This approach to implementing function *C* matches very well the medical doctors example provided by Tsandilas [108, p. 18:9] while, overall, the influence of the practice of inter-rater reliability studies is prominent in Tsandilas' approach to gesture elicitation. Admittedly, the set of signs does not exist prior to the elicitation study, but instead is constructed during the classification step, i.e., if the proposal does not match any of the existing signs, the set is expanded with a new sign. (This aspect is also true in inter-rater reliability studies, because the set of codes that are generated, i.e., the codebook, is usually built inductively from the text of interviews or observations obtained. In the popular grounded theory approach [104], for example, the open coding phase generates codes from the data itself.) This type of supervised approach was implemented before for end-user elicitation, for instance, in a study by Mauney et al. [73], who described their classification procedure of gestures into signs as follows: *"To promote consistency, the moderators created an online gesture glossary that contained pictures and textual descriptions of unique gestures. If a participant made a gesture that was in the glossary, the moderator simply referenced it. If a participant made a unique gesture that was not yet in the glossary, the moderator created a new entry, thereby making that new gesture available to all subsequent moderators to reference"* (p. 4019). According to this procedure, a set of signs is already available to compare against, which makes the classification process *supervised* [130].

The second approach is to look at the function *C* as a clustering procedure, where all the elicited proposals are grouped into classes of similar types, an option that is equally permitted by Tsandilas' formalization of the classification step: *"In most cases, however, sign vocabularies are open-ended, i.e., they are not known or fixed in advance. Instead, they are defined indirectly through an identity or a similarity measure that determines whether any two gestures correspond to the same or two different signs"* [108, p. 18:5]. Although not stated explicitly, this quote suggests a clustering procedure.

There are important differences between the two possible approaches permitted by the function *C* and, unfortunately, Tsandilas [108] does not continue his formalization to clarify them. Although credit is due for starting this formalization process, Tsandilas leaves unresolved a state of uncertainty that existed well before him. For example, when introducing the guessability method, Wobbrock et al. [140] described the classification step as follows: *"symbols are tested for equality and grouped [...] After grouping, the different referents within each group are identified and the number of referring symbols counted,"* remarking that *"it is essential for conflict resolution [...] that captured symbols be testable for equality. Testing equality may be trivial, as in the case of keyword symbols, or more complex, as in the case of (x, y) point traces for unistrokes. For more complex symbols, designers may already have software to interpret them. Human judgment can also determine equality*

among, for example, sketches of icons" (p. 1870). This description is general and essentially leaves the implementation of the classification step to designers. Four years later, in the first hand-gesture elicitation study, Wobbrock et al. [141] again described the classification step at a general level: *"After all 20 participants had provided gestures for each referent for one and two hands, we grouped the gestures within each referent such that each group held identical gestures"* (p. 1087). One of the first gesture elicitation studies that explicitly stated a clustering approach was Ruiz et al.'s [93]: *"For each participant, a transcript of the recorded video was created to extract individual quotes and classify and label each motion gesture designed by the participant. The quotes were then clustered to identify common themes using a bottom-up, inductive analysis approach"* (p. 199). Soon after, the GECKo tool [6], originally introduced to quantify user consistency in stroke-gesture articulation on touchscreens, implemented an automated hierarchical clustering procedure to identify similar gesture articulations and compute agreement (or consistency) rates.

This history shows that clarifications are needed to aid users of the method in the classification step. We argue that both the supervised and unsupervised approaches are useful to understand elicitation data. To this end, we draw inspiration from the 8-stage iterative approach to data analysis from pattern recognition [130, pp. 3–4], from which we adopt the following stages: (1) formulation of the problem, (2) data collection, (3) initial examination of the data to get a feel for the structure, (4) clustering, (5) discrimination, and (6) assessment of results and interpretation. Our operational model from Figure 1 addresses stages 1 and 2, and visualizations of the proposals implements stage 3. Next, we provide the necessary mathematical formalism to implement stages 4–6.

2.3.2 Formalization of the Classification Step. Let $\mathcal{P}$ be the set of all possible proposals for referent r. Let δ be a dissimilarity function defined over the Cartesian product $\mathcal{P} \times \mathcal{P}$ with values in $\mathbb{R}^+$.[4] Higher values of δ indicate proposals that are less similar. For example, δ may be the **Dynamic Time Warping** (**DTW**) function employed in many application domains [54, 79, 106] or the result of the experimenter's judgment on whether two proposals can be assigned the same sign [141]. Note that we do not require any special properties of δ other than non-negativity, i.e., $\delta(x, y) \geq 0$ for all x, y from $\mathcal{P}$. Functions δ that respect symmetry, i.e., $\delta(x, y) = \delta(y, x)$, and the identity of indiscernibles, $\delta(x, x) = 0$ for all $x \in \mathcal{P}$, are referred to as dissimilarity coefficients in the pattern recognition literature [130, p. 419]. Furthermore, dissimilarity coefficients that satisfy the triangle inequality, i.e., $\delta(x, y) \leq \delta(x, z) + \delta(y, z)$ for all x, y, z from $\mathcal{P}$, are called distances or metrics [130, p. 419]. For example, the Euclidean distance is a metric, but the DTW function, popular for gesture classification [106, 143], is not [54]. A special class of metrics, called ultrametrics, satisfy the stronger ultrametric inequality $\delta(x, y) \leq max\,(\delta(x, z), \delta(y, z))$ [130, p. 363]. It is important not to confound dissimilarities with distances because they imply different properties of the computations they perform; for example, some clustering techniques, such as hierarchical methods, employ ultrametric transformations of the dissimilarity values in order to compute the clustering partition [130, p. 363]. This brief overview of dissimilarities and distances/metrics is useful to appreciate our few requirements on the properties of δ's, which makes our formalization of the classification step very general and encompassing of a variety of ways to implement step 5 from Figure 1.

Following the dissimilarity–consensus approach [114], we consider the following rule to decide whether two proposals p_i and p_j, elicited from two participants P_i and P_j, are similar with respect to the dissimilarity function δ:

$$p_i \,\alpha\, p_j \Leftrightarrow \delta(p_i, p_j) \leq \epsilon, \tag{1}$$

[4]Note that our formalism is general and applies to any domain of investigation/elicitation, not just gesture input. However, gesture elicitation [124] has been the most popular application of the end-user elicitation method to date in HCI.

where ϵ is a positive value representing the tolerance at or below which two proposals are considered sufficiently similar to be assigned the same sign, and symbol α denotes the agreement relation. For example, δ can be the DTW function as in [114] and ϵ chosen in physical units, such as all whole-body movement for which the difference in the tracked points on the torso, legs, and arms is cumulatively less than 0.5 meters is considered equivalent. Or, δ can be defined as the number of properties for which two proposals are different, e.g., type of kinematic impulse, dimension, or complexity label for motion gestures [93], and ϵ set to 1. We use $\not\alpha$ to denote the situation when two proposals are not in agreement, i.e., $p_i \not\alpha p_j$. Our notations for elicited proposals follow the original formalization from Wobbrock et al. [140], according to which proposals are denoted by lowercase letters, e.g., p_i represents the ith proposal collected for some referent during the study. Also, we consider that each participant provides just one proposal for any given referent, and we denote participants with uppercase letters, e.g., proposal p_i comes from participant P_i.[5] Thus, the cardinality of the set of elicited proposals $\{p_i\}$ for some referent is equal to the number of participants from the study. Although specific instances of end-user elicitation studies have elicited multiple proposals from the same participant [75, 114], we restrict our discussion to one proposal per participant only, according to the original method introduced by Wobbrock et al. [140, 141]. This premise is particularly convenient since it presents the advantage that all proposals for a given referent are independent of each other, ensuring the independent and identically distributed trials assumption required by the statistical tests that we evaluate in Section 9.

On first look, it may appear that the use of the δ formalism restricts the application of Equation (1) to numerical measures only. However, δ's are defined in our approach with minimal constraints, i.e., all that our formalism requires from δ's is their non-negativity. Therefore, δ's can be dissimilarity coefficients, distances, or even ultrametrics that have explicit mathematical formulations, such as the Euclidean distance between two points in space, but they can also represent any judgment about the dissimilarity of the two proposals being compared, made by a human in a way that is tacit and, perhaps, difficult to explain. Note that, in order to evaluate Equation (1), only two values are needed: how dissimilar the two objects are and how much dissimilarity can be tolerated. The explicit part uses formulas for δ. The implicit part, where a human judges dissimilarity, embeds the formalism. For example, in their elicitation study of ear-based interactions, Chen et al. [18] noted: *"we simply separated gestures that used two or more fingers from those that used only one finger [...] But there was one exception, when the gesture was a metaphor (e.g., using two fingers to perform the scissor), not abstractly used, we would not follow the aforementioned criteria. Loosening the restriction from 'gestures must be identical within each group' to 'gestures must be similar within each group' made this classification better represent the thought underlying the gestures"* (p. 186:10). The δ, operating at an implicit level in Chen et al. [18], is clearly observable.

Equation (1) enables us to define the ϵ-Agreement Rate (AR_ϵ) given a set of proposals $\{p_i\}$ for referent r:

$$AR_\epsilon(r) = \frac{\sum_i \sum_{j \neq i} [\delta(p_i, p_j) \leq \epsilon]}{N(N-1)} \times 100\%, \tag{2}$$

where N is the number of participants or proposals, p_i and p_j are the proposals of participants P_i and P_j ($1 \leq i, j \leq N$), and $[\cdot]$ represents Kronecker's function [50, p. 240] that evaluates to 1 when the inner expression is true and to 0 when false.

At this point, it is beneficial to see how Equations (1) and (2) connect to clustering techniques in order to formalize step 5 from Figure 1, where the set of descriptions of the participants' proposals

[5]Note that in Wobbrock et al. [140], P_i was used to denote the ith subset of identical proposals for some referent. Since we refer frequently to participants in this article, and we also consider that just one proposal is elicited from each participant, it is more convenient for our purpose to use notation P_i to denote the participant from which proposal p_i was elicited.

is partitioned so that representative signs emerge, i.e., the consensus set. There are many methods to implement clustering. In their textbook on statistical pattern recognition, Webb [130] discusses hierarchical methods (e.g., the single-link or the complete-link method), mixture models (e.g., maximum likelihood procedures), and sum-of-squares methods (e.g., k-means) as common choices for clustering data. In the following, we focus on hierarchical clustering methods since the other approaches, with some exceptions [42], require the number of clusters to be specified in advance as a general rule, which is less suitable for end-user elicitation studies for which the number of clusters (corresponding to signs) is to be determined. Hierarchical methods construct hierarchical trees, called dendrograms, which represent nested set of partitions, in which individual clusters merge or are divided iteratively, according to the bottom-up or top-down principles. For example, the single-link hierarchical method [130, p. 364] puts two objects into the same cluster if there exists a chain of intermediate signs linking them such that all the intermediate pairwise comparisons are less than a threshold; the complete-link method [130, p. 367] considers the maximum dissimilarity between all the objects from the two clusters; and the general agglomerative algorithm [130, p. 368] produces dendrograms by employing various cluster dissimilarities, e.g., based on centroids or medians. However, no matter how the dissimilarity between two clusters that merge during the construction of the dendrogram is defined, a transformation of the dissimilarities $\delta(p_i, p_j)$ to a new set of values that satisfy the ultrametric inequality is performed [130, p. 363], and the dendrogram is sectioned with a threshold to obtain a specific partition of clusters: the result of the clustering.

In the following, we show the connection between Equations (1) and (2) and complete-link clustering, a method that concentrates on the internal cohesion of the clusters [130, p. 367] producing homogeneous, compact clusters [130, p. 370].[6] We illustrate this connection with an example. Say that a study has elicited five proposals $\mathcal{P} = \{p_1, p_2, p_3, p_4, p_5\}$, for which the dissimilarity values δ are illustrated in Figure 2, left. Note than δ is not a metric since the triangle inequality is not met.[7] If we choose $\epsilon = 25$, then the agreement rate is $4/(5{\cdot}4) = .20$; if we choose $\epsilon = 65$, then the agreement rate is $12/(5{\cdot}4) = .60$. We now compute the dendrogram for this data (Figure 2, right) as follows: at step one, proposals p_1 and p_2 are identified to be the closest in terms of dissimilarity ($\delta(p_1, p_2) = 10$) so they are grouped to form the first cluster; then, proposals p_3 and p_4 form the second cluster since they have the smallest dissimilarity (20); at step three, clusters $\{p_1, p_2\}$ and $\{p_3, p_4\}$ merge with a dissimilarity of 60 (the maximum dissimilarity between their members according to the complete-link method); and, finally, proposal p_5 joins at $\delta = 100$. If we section the dendrogram from Figure 2 at $\epsilon = 25$, we obtain the partition $\{\{p_1, p_2\}, \{p_3, p_4\}, \{p_5\}\}$; if we section at $\epsilon = 65$, we obtain $\{\{p_1, p_2, p_3, p_4\}, \{p_5\}\}$. Thus, the decision that researchers or practitioners need to make with hierarchical clustering (i.e., where to cut the dendrogram) is equivalent to using a tolerance level, as in Equation (1), to evaluate how dissimilar two objects are. More importantly, for each cluster produced with the complete-link method, the dissimilarities between the proposals falling into that cluster is always less than the chosen tolerance, e.g., when $\epsilon = 25$, $\delta(p_1, p_2) \leq 25$ and $\delta(p_3, p_4) \leq 25$; and, when $\epsilon = 65$, $\delta(p_1, p_2) \leq 65$, $\delta(p_1, p_3) \leq 65$, $\delta(p_1, p_4) \leq 65$, $\delta(p_2, p_3) \leq 65$, $\delta(p_2, p_4) \leq 65$, and $\delta(p_3, p_4) \leq 65$; see the dissimilarity matrix from Figure 2, left. In general, in each cluster of the partition delivered by the complete-link method, one that is recommended for the internal cohesion and homogeneous structure of the resulting clusters [13, 130], we will have proposals p_i for which the dissimilarity δ with respect to all the other proposals from the same

[6] Other methods, such as single-link clustering, are subject to the chaining effect, generating long straggly groups, while the centroid and median methods may lead to inversions, making the dendrogram difficult to interpret, as well as to multiple solutions when ties are present in the dissimilarities; see Webb [130, pp. 370–371]. Other authors [13, p. 420] also give preference to the complete-link method (as well as Ward's method that minimizes the total within-cluster variance) compared to single-link, centroid, and average link clustering.

[7] For example, $\delta(p_2, p_4) = 60$, $\delta(p_1, p_2) = 10$, $\delta(p_1, p_4) = 40$.

	p_1	p_2	p_3	p_4	p_5
p_1	0	10	30	40	70
p_2	10	0	50	60	80
p_3	30	50	0	20	90
p_4	40	60	20	0	100
p_5	70	80	90	100	0

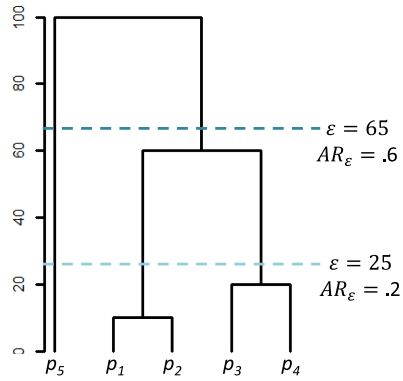

Fig. 2. An example with five proposals illustrating the connection between Equations (1) and (2) and the complete-link clustering method [130]. On the left, the dissimilarity matrix is shown. On the right, the dendrogram computed by the complete-link method.

cluster is less than ϵ, i.e., those proposals will be considered in agreement according to the partition delivered by complete-link clustering, which is exactly what our Equation (1) defines. *Therefore, Equation (1), which specifies agreement as pairwise comparisons between individual proposals, is verified by the result of the complete-link clustering method, while computing agreement based on the clustering partition leads to the same result as when Equation (2) is employed.*

At this point, it is important to anticipate an important result: Equation (2) matches the consensus rate measure of Vatavu [114] and reduces to the *AR* measure of agreement of Vatavu and Wobbrock [120] or, respectively, to the $Agreement_S$ formulation of Findlater et al. [32] of the *A* score [140, 141], once the $[\cdot]$ expressions have been evaluated to 0 or 1. In Sections 6 and 7 of this article, we will come back to these measures of agreement to show how they relate to each other in order to clarify research question $[RQ_2]$.

This formalization of calculating agreement using a dissimilarity function also makes sense from the perspective of real-world application of end-user elicitation studies including gesture recognition, where practitioners wish to implement an actual user interface or interactive system that would recognize the types of proposals (e.g., gestures, voice commands) revealed by the end-user elicitation study in the first place. In that case, a recognizer is needed, which implies the implementation of a dissimilarity function δ, e.g., the Euclidean distance of the $1 recognizer [143] or the point-cloud distance of the $P recognizer [116] in the case of stroke-gesture input, and possibly a threshold ϵ to implement a rejection rule [25]. From this perspective, if the dissimilarity function implemented by the target application cannot ignore the differences between two proposals p_i and p_j, i.e., $\delta(p_i, p_j) > \epsilon$, then it is reasonable to consider that the two proposals should not be declared in agreement during the elicitation stage, i.e., our Equations (1) and (2). For example, after compiling the consensus gesture set, Vogiatzidakis and Koutsabasis [129] found that implementing it in an actual system led to conflicts in gesture recognition since the Kinect-based recognizer could not distinguish between some of the gestures. As a result, the authors refined the consensus gesture set by changing some gestures and simplifying others to match the capabilities of the recognition technology used to implement those gesture commands. This example is revelatory for the need to incorporate a dissimilarity function, used for both recognition and clustering proposals into signs, into the formalism of elicitation studies. Unfortunately, only a few papers (of more than 200 elicitation studies published to date [124]) have followed the results of their studies into an actual implementation of a gesture recognizer or actual system [53, 61, 83, 105, 144]. This problem

has been noted before, such as by Nebeling et al. [82], but unfortunately to no avail: *"most studies thoroughly adopt Wobbrock et al.'s [..] methodology in order to obtain a user-defined interaction set, but without considering implementation issues"* [82, p. 15]. The sad result is that the literature on gesture elicitation has been disconnected from that on gesture recognition. Our formalization of the classification step using dissimilarity functions (Equations (1) and (2)) sets the foundation for addressing research questions [RQ_1]–[RQ_2] listed at the outset of our article. It should also help reconnect the two sides (elicitation study *vs.* implementation of the study results) and encourage a practice of evaluation of gesture recognizers and system implementation following end-user elicitation.

2.3.3 End-User Elicitation as a Scientific Method That Tells Us about People. So far, the scientific literature has primarily reported applications of the end-user elicitation method to compile sets of signs having consensus to inform "design for guessability" [140] for interactive systems, e.g., what users' most common preferences are for gestures to effect specific tasks on touchscreens [141], smartphones [93], interactive television [111], deformable displays [107], smart rings [36], earpieces [18], and so on; see Villarreal et al. [124] for an overview of gesture elicitation studies. However, a less common pursuit, yet nevertheless key result, of an end-user elicitation study is the empirical data that enables new findings and development of knowledge for understanding people, i.e., what we call *traits* in Figure 1, step 6. For instance, the trait of preferring symmetric signs for dichotomous referents [111, 141], or the trait of people who are blind to prefer edge-based gestures as well as gestures that involve tapping on a virtual keyboard for touchscreen input [53].

Several studies have employed the end-user elicitation method to unveil, quantify, and analyze differences in the mental models and/or the proposals elicited from various user groups and even for individual users. For example, Malu et al. [70] conducted a gesture elicitation study to understand accessible smartwatch gestures for people with upper body motor impairments. They noted: *"Unlike the goal of Wobbrock et al.'s original study method [141], we did not compute agreement, as our goal was not to create a highly guessable gesture set but to characterize the range of gestures created and to compare preferences for touchscreen and non-touchscreen gestures"* (p. 488:7). The authors characterized the gestures proposed by the participants with motor impairments in terms of gesture nature and rationale, gesture properties (e.g., number of fingers for gesture articulation), and locations chosen on the smartwatch to articulate those gestures. The key observation here, which highlights step 6 in our model (Figure 1) as independent of the identification of signs, is that Malu et al. were not driven in their scientific investigation by the goal of compiling a consensus set of smartwatch gestures, but by their desire to unveil the preferences for accessible gesture input on smartwatches, and to document and characterize those preferences in various ways. In this context, the end-user elicitation method was employed to conduct science that informs us about people and, in particular, about the traits of gestures that people with motor impairments would like to use.

The end-user elicitation method is therefore a scientific tool to conduct science that tells us about people and is not just a means to arrive at a consensus set of signs [108]. Another practical example is Kane et al. [53], who conducted a gesture elicitation study to analyze and report the differences between touchscreen gestures proposed by blind and sighted participants, which the authors characterized in terms of the number of strokes, on-screen location, nature, ratings of easiness, and other measures, but were not looking for a consensus set of gestures. Also, Vatavu [114] characterized whole-body gestures naturally articulated by small children, between 3 and 6 years old. At that age, children's motor and cognitive skills are still in development and a consensus gesture set is hardly the goal for such users. Instead, elicitation serves the goal of understanding

human movement and gesture production, e.g., the findings from [114] showed that consensus in whole-body gesture articulation increases with age. Other examples of end-user elicitation studies further show that a consensus set of signs is not the only or ultimate goal of applying the elicitation method. For instance, in their study about how user-defined gestures for flatscreens generalize to spherical displays, including both adults and children, Soni et al. [101] noted: *"Our first analysis goal was to understand the characteristics of the participants' gestures for spherical displays, so we analyzed all the gestures for our full set of 16 referents ... Our second goal was to understand how user-defined spherical display gestures differ from those identified for tabletop,"* while a consensus set of gestures was not reported. We refer interested readers to other examples as well: Rädle et al.'s [90] gesture elicitation study to understand preferences for cross-device gestures, where the focus was on the traits of proposed interactions and their labeling as synchronous, spatially aware, or spatially agnostic; Lee et al.'s [62] exploration of hand-to-face gestures, where the authors saw the diversity of elicited proposals as more suitable to their goal than high agreement scores for consensus sets; or Pham et al.'s [86] use of the end-user elicitation method to characterize differences in themes underlying participants' gestures proposed to interact with holograms in Mixed Reality according to the scale of the projected hologram, a scientific approach that was deemed more insightful than coding the observed gestures into signs.

These examples demonstrate an interest in the community toward using end-user elicitation as *a means to understand people* rather than as *a tool to arrive at a consensus set of signs.* This characterization feature of end-user elicitation has not been explicitly acknowledged in the community so far. One reason is probably the strict focus on eliciting gestures [108], with the practical goal of controlling systems, whereas gesture elicitation represents just one particular instance of general end-user elicitation [120, 121]. Also, steps 5 and 6 from our Figure 1, classification and characterization, are independent, but they need not be. For example, the set of signs resulted from classification at step 5 can represent the input for characterization at step 6 instead of characterizing the descriptions themselves.

3 THE NON-TRANSITIVE NATURE OF AGREEMENT

An interesting implication that results from formalizing the classification step using a dissimilarity function relates to the *transitivity of agreement.* Previous work [108, 120, 121] assumed agreement to be transitive and operated with this property, which means that if the proposal p_i elicited from participant P_i is in agreement with the proposal p_j elicited from participant P_j for referent r, and the proposal of P_j is in agreement with proposal p_k of participant P_k, then it must be that participants P_i and P_k also agree over the same referent, i.e., the following implication is always true:

$$(p_i \, \alpha \, p_j) \wedge (p_j \, \alpha \, p_k) \Rightarrow p_i \, \alpha \, p_k. \tag{3}$$

For example, Vatavu and Wobbrock [121] noted that *"transitivity of agreement means that the probability of observations that may turn out to be dependent on previous ones is 1.00"* (p. 3393) and Tsandilas [108] relied on this property to criticize the independence assumptions of the V_{rd} statistic of Vatavu and Wobbrock [120]: *"this solution is problematic because agreement pairs are highly interdependent, which is against the independence assumption of Cochran's Q test [...] if participant P_a agrees both with participant P_b and participant P_c, we can safely deduce that participants P_b and P_c agree with each other. Similarly, if P_a agrees with participant P_b but disagrees with participant P_c, then we can deduce that participants P_b and P_c disagree"* [108, p. 18:25].

However, we are about to show that *the transitivity of agreement is not a valid assumption for end-user elicitation studies.* Let us consider the example of voice commands elicited for controlling a remote display [75] and the Levenshtein distance [63], popular for measuring the difference between text strings [81]. This distance will compute a score of 4 for the commands "turn TV on"

Fig. 3. A simple dissimilarity function for multi-touch gestures, inspired by the Finger-Count [8] technique, is the difference in the number of touches, e.g., $\delta(\mathbf{p_1}, \mathbf{p_2}) = \mathbf{2}$, $\delta(p_1, p_3) = 3$, and $\delta(p_2, p_3) = 1$ for the hand poses illustrated in this figure. However, for a tolerance of 2, transitivity is not met.

and "turn on TV"; 5 for "turn TV on" and "TV on"; and 7 for "turn on TV" and "TV on."[8] If we set $\epsilon = 8$, then all these commands are in agreement according to Equation (1). However, if we choose ϵ anywhere in the real interval $[5, 7)$, transitivity is no longer met. Consider another example, where hand poses are used to select options on a multitouch display using a finger-count menu technique [8], and proposals elicited from participants include a two, four, and five finger tap to effect some referent as illustrated in Figure 3. A simple dissimilarity function could compare two hand poses by the difference in the number of touches simultaneously detected. If the tolerance ϵ is 2, transitivity is not met. To force even this simple dissimilarity function to behave transitively when used to define a relation, one must ignore information, such as the number of fingers in this example, as in the case of Chen et al. [18], who separated gestures that used two or more fingers from those that used only one finger. To verify the invalidity of the transitivity property assumed by prior work [108, 120, 121] for the agreement relation in end-user elicitation studies, we conducted three experiments using 19 publicly available gesture datasets.

3.1 Experiment #1: Observing the Non-Transitive Nature of the Agreement Relation

We employed the gesture elicitation dataset of Vatavu [114] consisting of 1,312 gestures acquired with the Microsoft Kinect sensor from 30 participants in response to 15 referents. To our best knowledge, this dataset is the only publicly available elicitation data with gestures collected in a numerical representation, e.g., in this case, as a series of 3-D points representing joints on the human body.[9] These gestures represent body movements produced by small children, between 3 and 6 years old, in response to short verbal commands delivered by a toy teddy bear, such as "fly like a bird." While the goal of this study was not to compile a consensus set of signs (step 5 in our model from Figure 1), the gestures that were elicited were compared for agreement in order to report how agreement between children's gestures increases with age, as children develop their motor and cognitive skills (e.g., step 6 from Figure 1). Furthermore, this dataset is likely to be free of legacy bias [75, 108] due to the young age of the participants.

We computed two measures of a non-transitivity rate (NTR):

(1) NTR_1 is the rate of non-transitive triples out of all the triples of N participants. For example, if 552 triples (p_i, p_j, p_k) do not exhibit transitivity (i.e., $p_i \, \alpha \, p_j$ and $p_j \, \alpha \, p_k$, but $p_i \not\alpha \, p_k$) and $N = 30$, then NTR_1 is $552/\binom{30}{3} = 552/4060 = 13.6\%$.

(2) NTR_2 is the rate of non-transitive triples of all the triples (p_i, p_j, p_k) for which the transitivity premises are satisfied, i.e., there are at least two agreement relations in the triple (p_i, p_j, p_k). Resuming the previous example, assume that the number of triples with at least two agreement relations is 1024. In this case, NTR_2 evaluates to $552/1024 = 53.9\%$.

[8] See an online implementation of the Levenshtein distance at https://planetcalc.com/1721 that produces the numerical results that we use in our example.

[9] Other papers announced public release of the datasets they elicited, such as [82], but only software was finally released; see the source code of KinectBrowser at https://github.com/michaelnebeling/kinectbrowser.

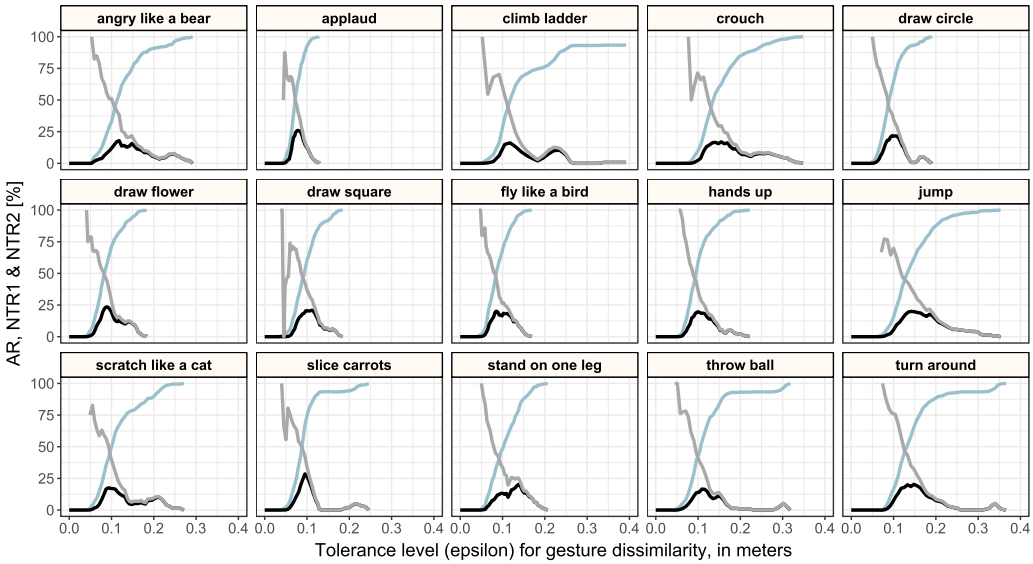

Fig. 4. Agreement rate *AR* (blue color) and non-transitivity rates NTR_1 (black) and NTR_2 (gray) function of the tolerance ϵ for the children gesture elicitation dataset [114].

We use two measures because assessing non-transitivity can be done in various ways. NTR_1 computes values that are normalized with respect to all the possible triples of participants and, thus, it represents an intuitive measure to compare across δ and ϵ's. However, since its denominator considers *all* triples, including those for which the transitivity premises are not met (i.e., less than two agreement relations observed in a triple), the extent of non-transitivity is underestimated, especially for studies with a large number of participants N. To correct for this aspect, NTR_2 uses a denominator that considers only those triples of participants that satisfy the premises for transitivity to occur, but results are now overestimated when there is little transitivity, such as when ϵ is small. To characterize thoroughly the phenomenon of non-transitivity in agreement formation, we report both measures for now, and later we focus on the peak value of NTR_1 as a compromise between NTR_1 and NTR_2.

Figure 4 shows the AR_ϵ growth curves (in blue) as a function of ϵ computed using Equation (2) and the normalized DTW dissimilarity for whole-body gestures from Vatavu [114]. The two NTR measures are shown superimposed: NTR_1 (black curve) increases with ϵ up to some point, after which it decreases to 0%, and NTR_2 (gray curve) decreases from 100% to 0% as the tolerance ϵ increases. These results show that when the criterion used to assess the similarity of gestures is too conservative (i.e., ϵ is small), few proposals will be in agreement according to Equation (1), and the number of non-transitive triples is large compared to transitive ones, resulting in large values for NTR_2. At the same time, NTR_1 will be small. As the tolerance ϵ increases, more proposals will be evaluated in agreement, creating more transitive triples and, thus, decreasing NTR_2 and increasing NTR_1 with respect to all possible triples. When the criterion is too liberal (i.e., ϵ is large), both NTR measures will approach 0%. Since a compromise will be made in practice between too conservative and too liberal criteria, we focus in the rest of this article on the peak values of NTR_1 as an estimation of the non-transitivity rate. For the gesture elicitation dataset of Vatavu [114], peak NTR_1 varied between 16.2% and 28.6% across all the 15 referents, and was normally distributed (Shapiro–Wilk's $W = .910$, $p = .134$) with a mean of 20.5% and **standard deviation** (**SD**) of 3.5%.

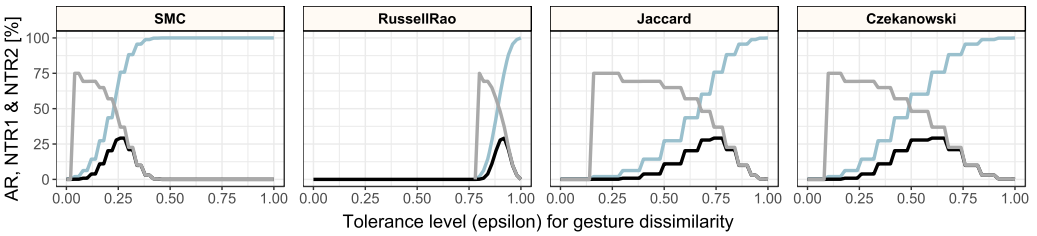

Fig. 5. Agreement rate *AR* (blue color) and non-transitivity rates NTR_1 (black) and NTR_2 (gray) function of the tolerance ϵ for the memorability gesture elicitation dataset [38, 80].

3.2 Experiment #2: Observing the Non-Transitive Nature of the Agreement Relation for Nominal Data

The previous experiment employed gestures represented in a computational form delivered by an actual gesture acquisition device, for which DTW, a popular dissimilarity function for gesture classification, could be easily applied. However, this has been hardly the norm for end-user elicitation studies, where researchers have often employed codebooks to represent the elicited proposals from their analyses [70, 93, 107], proposals that were usually logged as video recordings of their participants. In the following, we show that nominal data, resulting from codebook-based descriptions of elicited proposals, present non-transitivity under a variety of dissimilarity measures. To this end, we employ the memorability dataset[10] consisting of 366 gestures collected from 18 participants [80] that were coded along 12 dimensions [38, p. 28]: localization, number of hands, hand form, additional hand form, hand pose and path, gesture path, relation to action, relation to workspace, and gesture nature. For example, according to the localization dimension, a gesture can be "in air," "on surface," or "mixed"; while hand form has the following four categories: "spread," "flat," "mixed," and "other." Regarding this dataset, Grijincu et al. [38] noted: *"the dataset [is] amenable to classification algorithms that might be able to provide additional insight into the factors that affect memorability. The starting point is the 43-dimensional feature vector for each gesture, where each value is binary and represents each of the possible gesture classifications of the taxonomy"* [38, p. 30], and employed machine learning models, such as **support vector machines** (**SVMs**), to predict the memorability of each gesture from its representation as a set of binary features. In the following, we use the same binary representation for our analysis, where 1 denotes that the gesture has a given property (e.g., its localization is "on surface") while 0 denotes that it does not (e.g., its localization is not "in air"). To evaluate non-transitivity for nominal data coded in a binary form, we implemented four similarity functions, described in Webb [130, p. 424]: the **Simple Matching Coefficient** (**SMC**), Russel & Rao, Jaccard, and Czekanowski. For example, SMC computes the number of properties presented by both gestures (e.g., both gestures were performed "in air" and both gestures were not using the "spread" hand pose), divided by the total number of properties that are being evaluated. Since these measures are actually similarity measures, we convert them into dissimilarities by computing their complement to 1, e.g., $\delta(p_i, p_j) = 1 - SMC(p_i, p_j)$. For each dissimilarity function, we computed 366·365 = 133,590 pairwise comparisons by including all 366 gestures from the memorability dataset in our analysis. Figure 5 shows our results with NTR_1 increasing to a peak and then decreasing and NTR_2 decreasing to zero while ϵ approaches 1 (the maximum value for all the dissimilarity functions employed for this experiment). Peak NTR_1 was 29.2% for this dataset regardless of the dissimilarity measure.

[10] https://udigesturesdataset.cs.st-andrews.ac.uk/.

3.3 Experiment #3: Consolidating Observations about the Non-Transitive Nature of the Agreement Relation

Our previous experiments showed that non-transitivity is present and quite large when evaluating agreement for elicited gestures. This finding is already sufficient to acknowledge fundamental differences between end-user elicitation and inter-rater reliability studies (research question [$RQ_{1.2}$]), where transitivity is implicit in the latter and, consequently, to question the validity of adapting measures of agreement from inter-rater reliability to end-user elicitation [108] (research question [$RQ_{1.1}$]). We will return to clarify this aspect below in the article, as we accumulate more empirical evidence in this regard and build theoretical support. At this point, it is useful to see whether other types of gestures and dissimilarity functions lead to results similar to those reported in our first experiment. Unfortunately, we are not aware of other publicly available elicitation data besides [38, 114]. There are, however, several public datasets that were collected for evaluating the classification accuracy of stroke-gesture recognizers, such as $1 [143], $N [7], HHReco [48], and so on. To understand more about the non-transitivity of agreement, we decided to use this available data, which we reinterpreted from an elicitation perspective, as follows.

Let $\mathcal{D}$ be a dataset where N participants provided samples for a set of distinct gesture types. To simulate elicitation for a given referent, we used one sample from each participant and considered various target AR values from 0% to 100%. For example, assume a target AR of 25%. This level of agreement can be easily obtained with $q = 0.25^{\frac{1}{2}} \cdot N$ of the participants' proposals in agreement.[11] Thus, we randomly picked gesture samples of the same type for the first $q \leq N$ participants, while for the rest we picked different gestures from the dataset. We repeated this process 100 times for each distinct gesture type (referent); e.g., there were $16 \times 100 = 1{,}600$ trials for the $1 dataset [143]. By adopting this procedure, we look at these gesture collection experiments as gesture elicitation studies, where participants are prompted with instructions regarding the shape of the gesture to articulate, e.g., an asterisk sign [7], and in some cases how to perform the gesture, e.g., faster or slower [143], but what is actually elicited is participants' specific ways of articulating gestures. Understanding how different participants (or, the same participant across trials) choose [6] or are able [16] to articulate gestures is key for informing the design of gesture recognizers that would prove tolerant to such differences in articulation compared to the canonical examples from their training sets. In fact, by highlighting "articulation" as the feature that is elicited in gesture collection experiments, previous work [6, 16] has employed the agreement rate measures of the end-user elicitation method to analyze consistency in gesture articulation.

Table 1 reports the peak NTR_1 for 18 gesture datasets, comprising a total of 87,316 gestures of various types[12] collected from 447 participants, and using three dissimilarity measures (Euclidean, DTW, and point-cloud distance) compatible with all of these gesture types [106, 113, 114, 143]. The mean peak NTR_1 across all datasets varied between 18.6% and 29.5% (M = 25.2%, SD = 3.4%), reconfirming our previous results from the first experiment on whole-body gestures.

4 THE PROBLEM OF CHANCE AGREEMENT

Tsandilas [108] argued that agreement measures used in elicitation studies, such as A [140, 141] and AR [32, 120, 121], were defined without considering agreement occurring by chance and, thus, their values are artificially large, as they reflect both intrinsic (i.e., true) and chance agreement.

[11] $AR = \frac{q(q-1)}{N(N-1)} = \frac{0.25^{\frac{1}{2}} N(0.25^{\frac{1}{2}} N-1)}{N(N-1)} \approx 0.25$.

[12] 2-D unistrokes [119, 143], 2-D multistrokes performed with the stylus [7, 48, 135] and the finger [117, 118], 3-D accelerated motion of the hand captured using the Wii Remote controller [17, 44, 65], and whole-body Kinect gestures [35, 113, 114].

Table 1. The Peak Non-Transitivity Rate NTR_1 (Mean and Standard Deviation) Evaluated for Various Dissimilarity Functions, Gesture Datasets, and Gesture Types

Dataset	Gesture types	N	Size	Non-transitivity rate [%]: Mean (SD)		
				Euclidean distance	DTW dissimilarity	Point-cloud distance★
$1-slow [143]	stylus, 2-D unistrokes	10	1,600	27.5 (6.8)	27.2 (7.1)	29.3 (6.5)
$1-medium [143]	stylus, 2-D unistrokes	10	1,600	27.8 (6.9)	27.5 (7.4)	29.6 (6.6)
$1-fast [143]	stylus, 2-D unistrokes	10	1,600	28.2 (7.1)	28.2 (7.3)	31.1 (6.4)
MMG-slow [7]	stylus + finger, 2-D multistrokes	20	3,200	24.8 (6.7)	24.9 (6.8)	24.6 (5.8)
MMG-medium [7]	stylus + finger, 2-D multistrokes	20	3,200	25.4 (6.7)	25.7 (6.7)	24.8 (5.8)
MMG-fast [7]	stylus + finger, 2-D multistrokes	20	3,200	25.6 (6.4)	26.1 (6.4)	24.6 (5.6)
Difficulty-1 [119]	stylus, 2-D unistrokes	14	5,040	24.0 (5.6)	24.3 (6.3)	27.2 (6.2)
Difficulty-2 [119]	stylus, 2-D unistrokes	11	4,400	26.3 (6.9)	25.9 (7.4)	29.2 (7.0)
HHReco [48]†	stylus, 2-D multistrokes	19	7,544	30.0 (5.1)	28.7 (5.3)	27.1 (4.8)
NicIcon [135]‡	stylus, 2-D multistrokes	34	13,819	24.6 (6.8)	25.1 (6.2)	25.6 (4.3)
Low vision [117]	finger, 2-D multistrokes	54	6,562	20.5 (6.3)	19.2 (6.0)	19.7 (5.0)
Motor impairments [118]	finger, 2-D multistrokes	70	9,681	22.3 (6.9)	22.4 (7.1)	19.3 (5.4)
Wiimote [44]	3-D motion, Wii Remote	17	8,500	30.1 (5.3)	29.0 (5.0)	29.2 (4.9)
6DMG [17]	3-D motion, Wii Remote	28	5,600	24.9 (5.6)	25.3 (5.3)	24.7 (5.8)
uWave [65]§	3-D motion, Wii Remote	56	4,480	25.5 (4.0)	25.8 (2.9)	23.4 (3.3)
Children [114]	whole-body gestures, Kinect	15	1,312	19.1 (4.8)	19.7 (4.7)	17.1 (4.2)
MSR Cambridge [35]	whole-body gestures, Kinect	30	5,654	20.1 (4.5)	19.3 (4.4)	18.6 (4.6)
Smart Pockets [113]	whole-body gestures, Kinect	9	324	28.2 (6.2)	28.5 (6.7)	28.3 (6.3)

★The Hausdorff distance is computed for whole-body gestures as in [114], while the $P point-cloud distance [116] is computed for stroke-gestures and 3-D motion.

†The actual number of gestures that we employed from this dataset differs slightly from the number reported in [48] (7,410 gestures) and from the number reported on the HHReco homepage (7,791 gestures) due to conversion issues. The dataset is available from https://ptolemy.berkeley.edu/projects/embedded/research/hhreco/.

‡The actual number of gestures that we employed from this dataset differs from the number reported in [135] (23,641 gestures) and from the one reported on the dataset homepage (26,163 gestures), because only a part of the dataset was available to us. The dataset used to be available at http://www.unipen.org/.

§Actually, 8 participants provided gestures during 7 days over a period of about 3 weeks [65]; for this dataset, we considered them as 56 different participants.

The concept of chance agreement comes from inter-rater reliability studies [21, 22, 34, 40, 58, 59], which have been traditionally implemented using nominal data types and raters assigning subjects to predefined categories. For example, physicians may assign patients suffering from spinal pain into the following three categories: "derangement," "dysfunction," and "postural." For such tasks, and especially when the number of categories is small, Cohen [21] argued that a certain amount of agreement is to be expected by chance, i.e., if two raters are unclear about the category to choose, but they still need to make a choice, chance agreement can occur because of their limited options and potential bias for or against specific categories. Cohen evaluated the amount of chance agreement (p_e) by using the joint probabilities of the marginal proportions [21], which he subtracted from the percent agreement (p_a):

$$\kappa = \frac{p_a - p_e}{1 - p_e}, \tag{4}$$

where κ is Cohen's kappa coefficient of agreement [21]. Other authors proposed other ways to estimate p_e, while using the same form of Equation (4), which resulted in various coefficients of agreement, such as Fleiss' κ_F [34]; Brennan and Prediger's β family of coefficients [15], including κ, κ_n, and κ_b; Scott's π [98]; Holley and Guilford's *G-Index* [45]; Krippendorff's α [58]; or Gwet's AC_1 [39]. The percent agreement p_a is computed using the same formula as the *AR* measure [32, 120] employed in elicitation studies by dividing the number of pairs of participants in agreement by the maximum number of pairs that could be in agreement; see Fleiss [34, p. 379]. Although there seems to be general consensus[13] that p_a needs to be corrected for inter-rater reliability studies, how to define and especially how to calculate chance agreement p_e is not a trivial problem and has generated considerable debate [40].

Tsandilas' [108] approach to end-user elicitation is heavily influenced by the practice of inter-rater reliability [21, 34, 40, 59], where raters select options from a predefined list of (usually) nominal categories in order to assign statements made by subjects (e.g., in interviews) or observations about subjects (e.g., during ethnography) to those categories. Tsandilas capitalizes on the concept of "bias" [76] to model chance agreement with certain probability distribution functions. However, what Tsandilas most likely proves with his simulations about distributions that model bias (the two experiments from pages 18:11 and 18:13) is that if one removes from a measure of central tendency, such as *AR*,[14] its expected value, the result will be near zero and, therefore, the new measure is not biased. However, the assumption is that the original measure is biased in the first place and, therefore, removing its expected value based on the probability distribution that models this bias cannot but confirm this assumption. Therefore, bias exists where it is expected to exist. The question, however, is whether this expected level of agreement should be removed from the calculation of agreement measures. In the following, we provide arguments that this should *not* happen for end-user elicitation studies. This aspect becomes clear when we start to examine the differences between end-user elicitation and inter-rater reliability studies and conclude that the two are fundamentally different in their assumptions (research question [$RQ_{1.2}$]) and, thus, are in need of specific models, methods, measures, and tools for the calculation and analysis of agreement. We start highlighting these differences next.

4.1 Inter-Rater Reliability vs. End-User Elicitation Studies

Gwet [40] provides a detailed overview of inter-rater reliability studies and the agreement coefficients employed in those studies to quantify agreement between raters. Among many important aspects addressed in Gwet's critical survey, the difficulty of precisely defining the notions of

[13] Note, however, that some authors [109] consider the need to perform chance correction unconvincing.
[14] As we are about to show in Section 7.2, *AR* can be interpreted as a mean.

"inter-rater reliability" and "agreement" stands out because of the many forms in which inter-rater reliability studies have been conducted in the scientific literature [40, pp. 4–21]. Nonetheless, similarities exist between inter-rater reliability and elicitation, especially regarding the units involved, as remarked by Tsandilas [108]: raters are end-users, subjects are referents, and categories are signs. At a first look, inter-rater reliability and elicitation studies seem equivalent or at least similar. However, there are important differences between them that require calculating agreement and chance agreement in end-user elicitation very carefully.

One important difference is that inter-rater reliability studies operate with a fixed list of categories defined *before* the study. However, the list of categories (or signs) is one outcome of end-user elicitation studies (according to step 6 from our model in Figure 1) and, thus, is known only *after* the end-user elicitation study has completed. As Tsandilas [108] correctly remarks, *"this open-endedness does not affect how Fleiss' κ_F coefficient is computed, because the coefficient requires no prior knowledge or assumption about the number of possible signs q. Equations [..] only depend on the number of observed signs q_ and their frequencies"* (p. 18:12). This observation regarding the calculation of κ_F (and κ, for that matter) is correct, but the quantification of the probability of chance agreement (p_e in Equation (4)) is based on the assumptions that accompany these coefficients, i.e., that a predefined list of categories exists, and that *"the categories of the nominal scale are independent, mutually exclusive, and exhaustive"*; see Cohen [21, p. 38] regarding κ. These assumptions can be verified only when the categories are known, which is only *after* an elicitation study. From the perspective of the study participants, when participants make proposals (steps 2 and 3 in Figure 1), they have no idea about the final list of categories or signs; therefore, they are not picking categories, but rather making proposals without knowing the final categories that will emerge. From the perspective of the experimenter, when classification starts (step 6 in Figure 1), categories or signs are again unknown and emerge as the classification progresses. In this context, it is not possible to determine whether the categories are independent, mutually exclusive, or exhaustive [21] before the study, neither by the participants nor the experimenters. Therefore, any verification of these assumptions to justify adoption of κ or related coefficients that correct for chance agreement when analyzing data in end-user elicitation studies can take place only *a posteriori* and, consequently, the probability of chance agreement (p_e) is based on categories that emerge instead of categories that are known *a priori*, *which makes the notion of chance agreement dependent on categories that do not exist when measurements about agreement are collected.* We believe this perspective makes sense, but we do not wish for it to devolve into the same never-ending debate as whether and how to estimate p_e from Equation (4) in inter-rater reliability studies [40]. Instead, we want to make practitioners aware that, in order to be affected by chance agreement, the conditions for chance agreement to occur, i.e., the categories and assumptions about those categories [21], must be fulfilled *a priori* to the study. In most elicitation studies, they are not.

Another important difference between inter-rater reliability and end-user elicitation studies is how agreement is defined. Having a fixed set of predefined categories on a nominal scale [21, 34] makes the transitivity of agreement a consequence of agreement formation in inter-rater reliability studies, whereas we showed that transitivity should not be assumed for end-user elicitation. For example, Cohen's κ and Fleiss' κ_F (also employed by Tsandilas [108] to analyze elicitation data by correcting for chance agreement) assume that categories are defined on a nominal scale. When the scale is ordinal, Cohen's weighted κ [22] as well as other measures was proposed to account for situations of partial agreement when selected categories are close on the ordinal scale. However, when the scale is interval or ratio, the recommendation is not to use these indices (unless the ratings are predetermined *before* the experiment; see Gwet [40, p. 24]), but rather to report intraclass correlations. In fact, Gwet [40] notes that for interval or ratio data, *"the very notion of agreement must be revised. Given the large number of different values a score may take, the likelihood of two*

raters assigning the exact same score to a subject is slim" (p. 17) and *"with agreement no longer referring to an exact match, the notions of chance agreement and percent agreement evaporate"* (p. 186).

We illustrate this insight with an example adapted from Fleiss [34, p. 379]. Say that three psychiatrists, P_1 to P_3, diagnose patients into one of five categories: "depression," "personality disorder," "schizophrenia," "neurosis," and "other." If P_1 and P_2 both pick "personality disorder" for a given patient, then we say that they are in agreement. If P_3 picks the same category, then all pairs of psychiatrists are in agreement; if P_3 assigns the patient to another category, then P_3 disagrees with both P_1 and P_2. *Whenever the list of categories is fixed and their scale of measurement is nominal, the transitivity of agreement is a mathematical consequence.* Put otherwise, while Cohen's [21] and Fleiss' [34] approaches to agreement calculation operate based on the equivalence relation "is equal to" between categories, end-user elicitation studies implement the tolerance relation "is similar to" or "is approximately equal to," as summarized by Equation (1). The result is a fundamental difference in how chance agreement can be regarded.

What would happen if one still used κ or κ_F when the agreement relation was non-transitive, such as when P_1 and P_2 agree, P_1 and P_3 agree, but P_2 does not agree with P_3? In that case, Cohen's [21] assumptions of independent and mutually exclusive categories are note met because of the conflict generated by P_2 and P_3. Or, in an attempt to salvage those two assumptions, one could consider that the category on which P_2 and P_3 do not agree was not available when P_1 and P_2 made their choice, but that scenario would break Cohen's third assumption: exhaustive categories [21]. In either case, κ (and κ_F) and non-transitive agreement relations are incompatible.

4.2 Clarifying Chance Agreement in End-User Elicitation Studies

Our mathematical formalism enables us to express the probability of two participants P_i and P_j being in agreement about their proposals for referent r, as follows:

$$P(P_i \, \alpha \, P_j \mid r) = \frac{\sum_{p_1 \in \mathcal{P}_i} \sum_{p_2 \in \mathcal{P}_j} [\delta(p_1, p_2) \leq \epsilon]}{|\mathcal{P}_i| \cdot |\mathcal{P}_j|}, \tag{5}$$

where $\mathcal{P}_i$ and $\mathcal{P}_j$ represent the pools of all possible proposals available to participants P_i and P_j according to their mental models of the system effect, and their developing mental models of possible commands to issue to bring about that effect; see Figure 1. In theory, sets $\mathcal{P}_i$ and $\mathcal{P}_j$ are infinite,[15] but for practical purposes, we can consider them finite. Moreover, because of inherent differences between participants, it is reasonable to assume that $\mathcal{P}_i$ and $\mathcal{P}_j$ are not identical, i.e., participant P_i might have access to different mental models leading to different types of commands or proposals not necessarily accessible to P_j and vice versa. In support of this argument, we refer readers to cultural differences highlighted by previous gesture elicitation studies [28, 73].

Equation (5) can be used both when referent r is specified, but also when r is not known in order to express the probability of two participants being in agreement when they are ignorant of the referent or metaphorically blindfolded, a situation considered by Tsandilas [108]. In the latter case, sets $\mathcal{P}_i$ and $\mathcal{P}_j$ should be larger than in the former. Various forms of bias, such as legacy bias [76] or performance bias [94], may affect both the size and structure of sets $\mathcal{P}_i$ and $\mathcal{P}_j$. Equation (5) shows that the probability of any two participants being in agreement depends on the models to which they have access, but also on the dissimilarity function δ and tolerance ϵ. Considering the classification perspective that we adopted in this article for end-user elicitation, there are four possible outcomes for proposals elicited from two participants when presented with referent r:

(1) *True agreement.* Both participants develop the same mental model of the system effect and of the command, and they present proposals for which the descriptions A and A' are evaluated

[15] Because any of their elements can be composed indefinitely many times, thus generating new elements [108].

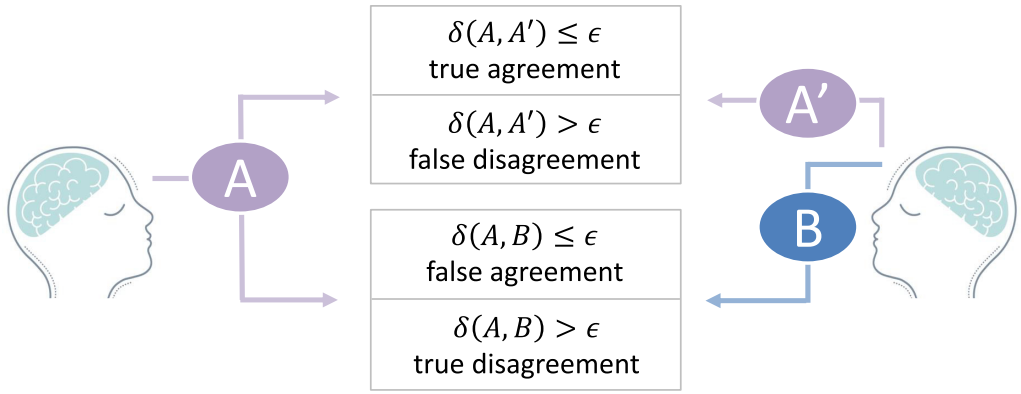

Fig. 6. Proposal **A** elicited from the participant on the left generates four possible outcomes when compared to proposals **A′** and **B** of the participant on the right.

as similar by the dissimilarity function, i.e., $\delta(A, A') \leq \epsilon$. In this case, there is *true agreement* between the two proposals; see Figure 6, top.

(2) *True disagreement.* The two participants develop different mental models of the system effect or of the command, which results in different articulations A and B, and the difference is detected correctly. In terms of the δ and ϵ formalism, we can write $\delta(A, B) > \epsilon$. The outcome is *true disagreement*, where the two participants and, correspondingly, their proposals for the referent genuinely disagree; see Figure 6, bottom.

(3) *False agreement.* The two participants develop different mental models, but the dissimilarity function fails to discriminate between the corresponding descriptions, i.e., $\delta(A, B) \leq \epsilon$, although A and B are different. This is a case of *false agreement* or a Type I error, i.e., the agreement hypothesis between participants is confirmed, whereas it should have been rejected.

(4) *False disagreement.* The two participants develop the same models, but δ fails to confirm the similarity of their corresponding descriptions, i.e., $\delta(A, A') > \epsilon$, although A and A' are actually sufficiently similar. In this case, we have *false disagreement* or a Type II error, i.e., the agreement hypothesis is rejected, where it should have been accepted.

False agreement includes chance agreement in inter-rater reliability studies [21, 34, 40]: two ratings result in agreement although agreement is not justified since, in the absence of suitable mental models for the task (i.e., an informed rating), random or biased models are applied instead (i.e., a rating caused by chance, or a rating that is biased by some factor). However, in traditional inter-rater reliability studies, false disagreement is not threatening for the main purpose of those studies, which is the assessment of "reliability": whenever two raters pick their options from a list of categories, and those options turn out to be different, the two raters are disagreeing with each other. However, the cause of disagreement, such as both picking categories randomly and it just happened that the categories were different, is less important because there is no "ground truth" of what they should have picked–rather, it is about what they did pick and the fact that they did not agree and, hence, the reliability of them agreeing was not affected. In Cohen's [21] words, *"there is no criterion for the 'correctness' of judgments, and the judges are a priori deemed equally competent to make judgments"* (p. 38). Thus, false disagreement is ruled out directly in inter-rater reliability by the definition of agreement used in such studies in direct relation to quantifying the reliability of agreement; see Gwet [40, p. 15]: *"With a nominal scale, two raters agree when their respective ratings assigned to a subject are identical, and are in disagreement otherwise."* Since disagreement

by chance is not interesting for inter-rater reliability studies that focus on surfacing agreement, which is corrected by subtracting agreement occurring by chance so that reliability is reflected by the actual, intrinsic agreement between raters, it was not considered by Cohen [21] nor Fleiss [34] when defining their coefficients κ and κ_F, respectively. However, when formalizing the classification step of end-user elicitation studies through the prism of a dissimilarity function δ and a tolerance threshold ϵ (Equations (1) and (2)), implemented either explicitly by a computer or implicitly by a human observer, both false agreement and false disagreement outcomes are relevant. *While the former artificially increases the observed agreement rate, just like chance does for inter-rater reliability studies, the second artificially decreases it.* As a consequence, we believe that agreement rate measures used in end-user elicitation data analysis should not force corrections in one direction or the other, unlike what happens in inter-rater reliability for κ and related coefficients. The κ coefficient reports agreement compared to a baseline represented by agreement occurring by chance or random allocation and, for some applications, this baseline can prove to be distracting. For example, in other fields, researchers have advocated for abandoning κ coefficients that focus on corrected agreement and recommended reporting only disagreement instead [52]. While the community is still on the fence about such issues, we believe that not correcting measures of agreement in end-user elicitation reflects best the interplay of agreement and disagreement occurring in the data, regardless of its nature and source.

A similar situation exists in pattern recognition, where *sensitivity* measures the actual positives that are correctly identified, and *specificity* does the same for negatives [30]. A perfect classifier would be 100% sensitive and specific, leaving no room for false positives or false negatives. In practice, a tradeoff is sought between specificity and sensitivity, usually represented in the form of a **Receiver Operating Characteristic** (**ROC**) graph [30], a useful tool for visualizing and comparing classifiers' performance. Just like in hypothesis testing with Type I and II errors, classifier design compromises in terms of minimizing false positive and false negative rates. Note that by classifier design, we mean a wide array of options, from supervised learning to unsupervised procedures, including those employing codebooks for grouping descriptions into signs during end-user elicitation analysis. For example, the researcher iteratively assigns new proposals to clusters depending on how similar those proposals are to the proposals already included in those clusters or with the representative "sign" of each cluster, based on the categories of the codebook. Even for such cases, the possible outcomes of the classification process remain the same as above.

At this point, we have accumulated sufficient empirical evidence and theoretical support to start clarifying the RQ outlined for end-user elicitation at the outset of this article. The discussion from this section, specifically, enables us to provide an answer for our first research question, [$RQ_{1.1}$]:

> **Research Question [$RQ_{1.1}$]:** Should the measures of agreement employed in end-user elicitation studies, such as *A* and *AR*, be corrected for chance agreement, just like in inter-rater reliability studies? If so, how?
>
> **Clarification:** Chance agreement should be considered and corrected *only* if the end-user elicitation study was conducted with a fixed set of nominal categories from which participants picked their proposals, i.e., end-user elicitation was implemented in the form of an inter-rater reliability study, such as in Stern et al. [103]. Otherwise, when elicitation studies are conducted following the original method [140, 141], agreement by chance is always opposed by disagreement by chance when agreement calculation is formalized with a dissimilarity function δ and tolerance threshold ϵ. Consequently, the magnitude of the intrinsic agreement always lies in the tension of false positives and false negatives, and the measures of agreement *A* [140, 141] and *AR* [32, 120, 121] traditionally employed in end-user elicitation should not be corrected for chance agreement.

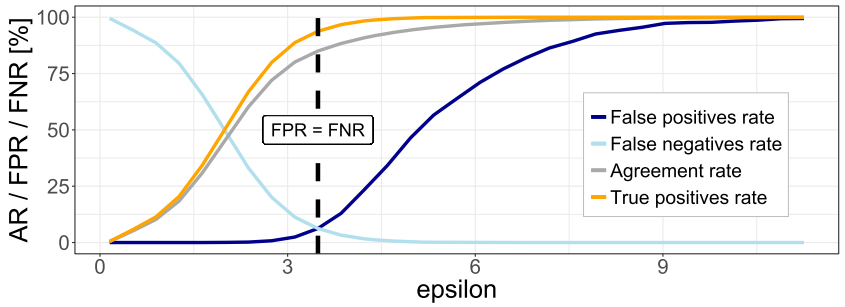

Fig. 7. Illustration of false positive and false negative rates as a function of the tolerance ϵ for the point-cloud dissimilarity function [116] and the MMG stroke-gestures dataset [7].

Next, we will be addressing the rest of the research questions fundamental to end-user elicitation studies. To this end, we continue our elaboration of theoretical support and presentation of corresponding empirical evidence.

4.3 Experiment #4: Observing False Positives and False Negatives when Calculating Agreement in End-User Elicitation Studies

Although the concepts of false positives and negatives should be well known to readers from the field of pattern recognition [130], inferential statistics [132], or the practice of HCI [137], we believe it is useful to illustrate them for the specific case of agreement defined as in Equation (2). To this end, we employ one of our previous gesture datasets: 3,200 samples of 16 distinct multistroke gestures performed by 20 users [7], and the point-cloud dissimilarity function [116]. Figure 7 highlights the point where the false positive rate (increasing with ϵ) and false negative rate (decreasing with ϵ) curves meet. If ϵ is below this point (i.e., the similarity criteria are conservative), the agreement rate is underestimated because there are more false negatives than false positives. If ϵ is above this point (i.e., the similarity criteria are more liberal), the agreement rate is overestimated because there are more false positives than false negatives. The magnitude of the agreement rate lies in the tension between false positives and false negatives, but there is a point where the two cancel each other out, leaving just the true, intrinsic agreement. No methods exist in the literature of end-user elicitation to locate the ideal ϵ that corresponds to the intrinsic level of agreement, simply because the community was not aware of this issue before. However, we believe that methods inspired by ROC analysis from pattern recognition [30] may be useful for this purpose; more details about potential future developments in this direction are available in the Discussion section.

We continue our exposition with specific illustrations from the memorability gesture dataset [38, 80], for which the elicited gestures were coded using a taxonomy composed of 12 dimensions, e.g., localization, number of hands, and so on. Since the authors did not compile a consensus set of signs (their goal was instead to understand memorability of user-defined *vs.* designer-defined gestures, i.e., they implemented step 6 instead of step 5 from our model in Figure 1), we do not have access to the signs for the referents examined in that work as ground truth data to be able to automate the computation of the false **disagreement rate** (**DR**) as in our previous example. However, the coded proposals, according to the 12-level taxonomy, are available to compute δ's. Therefore, we discuss a few examples that we identified by looking at the videos from this dataset and then computed δ's from the code-based descriptions of those videos available in the dataset. Figure 8 shows

the six gestures proposed by six participants to effect the "Homepage" action for a web browser. It is straightforward to identify four themes: draw the symbol of a house (proposals 1 and 2), use a specific hand pose (proposal 3), draw a circle (4 and 5), and draw a corner (proposal 6). However, executions of these themes are different, e.g., the house symbol proposed by the first participant is segmented, whereas the one drawn by the second participant is continuous; the circles drawn by participants 4 and 5 have opposite directions, while participant 5 also places their hand on the tabletop after completing the circle. The bottom of Figure 8 shows the dissimilarity matrices computed using four dissimilarity functions (SMC, Russell & Rao, Jaccard, and Czekanowski; see Section 3) and the corresponding dendrograms generated by the complete-link clustering method. Although the house symbols illustrate the same mental model, their codebook descriptions differ on 4 of the 12 dimensions: the path direction (straight *vs.* flexible), path flow (segmented *vs.* continuous), path shape (n/a[16] *vs.* closed), and relation to action (arbitrary *vs.* iconographic). Consequently, the dissimilarity between these proposals is not zero, but instead 0.15 according to SMC, 0.85 (Russell & Rao), 0.50 (Jaccard), and 0.33 (Czekanowski). Although this is fine (their executions are different, as we showed above), all the dissimilarity matrices also show that proposal 1 is more similar to proposal 6 and, indeed, when looking at their codebook-based descriptions, proposals 1 and 6 differ along fewer dimensions.

Now look at proposals 4 and 5: the two circles. According to the codebook, they differ along four dimensions: gesture nature (metaphorical *vs.* abstract), relation to action (iconic *vs.* arbitrary), hand orientation (n/a[17] *vs.* horizontal), and hand form (other form *vs.* spread) and, therefore, the dissimilarity between them is larger than zero. However, just like in the previous case, proposal 2 is more similar to proposal 4 based on the descriptions from the codebook, according to all the dissimilarity measures. The consequence is that proposals 1 and 6 and, respectively, 2 and 4 will form their own clusters first (see the dendrograms from Figure 8) and the only way to have them as part of the same cluster (i.e., to denote that they represent the same sign) is to increase the tolerance ϵ, which also then includes proposals 2 and 4 in that cluster. At least, this is what the data from the codebook tells us for a variety of dissimilarity measures. If, at this point, the researcher decides to form a cluster with 1 and 2 and another with 4 and 5, they would go against their own codebook, whereas classification based on other criteria, independent of the codebook, would make the codebook pointless in the first place.

Having established this example, we see how false agreement and false disagreement interplay: while clustering together proposals 1 and 6 is an example of false agreement, not clustering 1 and 2 is false disagreement. We acknowledge that the 12-level taxonomy employed by Grijincu et al. [38] was not necessarily developed with the goal to emerge signs from proposals, and other, more relevant dimensions could be employed for that purpose and work much better for the "Homepage" gestures. However, the codebook determines the dissimilarity matrix and, as the pattern recognition literature has been showing, false agreement and false disagreement are outcomes of any classification process. Since such codebook-based approaches have been used by researchers to characterize participants' proposals in end-user elicitation studies, when clustering is based on a codebook, situations of both false agreement and disagreement will emerge.

[16]In the memorability dataset, the value "n/a" is present for the path shape dimension to characterize the house symbol gesture produced by the first participant from Figure 8. Another possible coding could have been "open" to contrast the closed articulation of the house symbol produced by the second participant.

[17]In the memorability dataset, the value "n/a" is present for the hand orientation dimension to characterize the gesture produced by the fourth participant from Figure 8. Another possible coding could have been "index finger pointed" to contrast the flat horizontal hand employed by the fifth participant to mark the ending of their gesture.

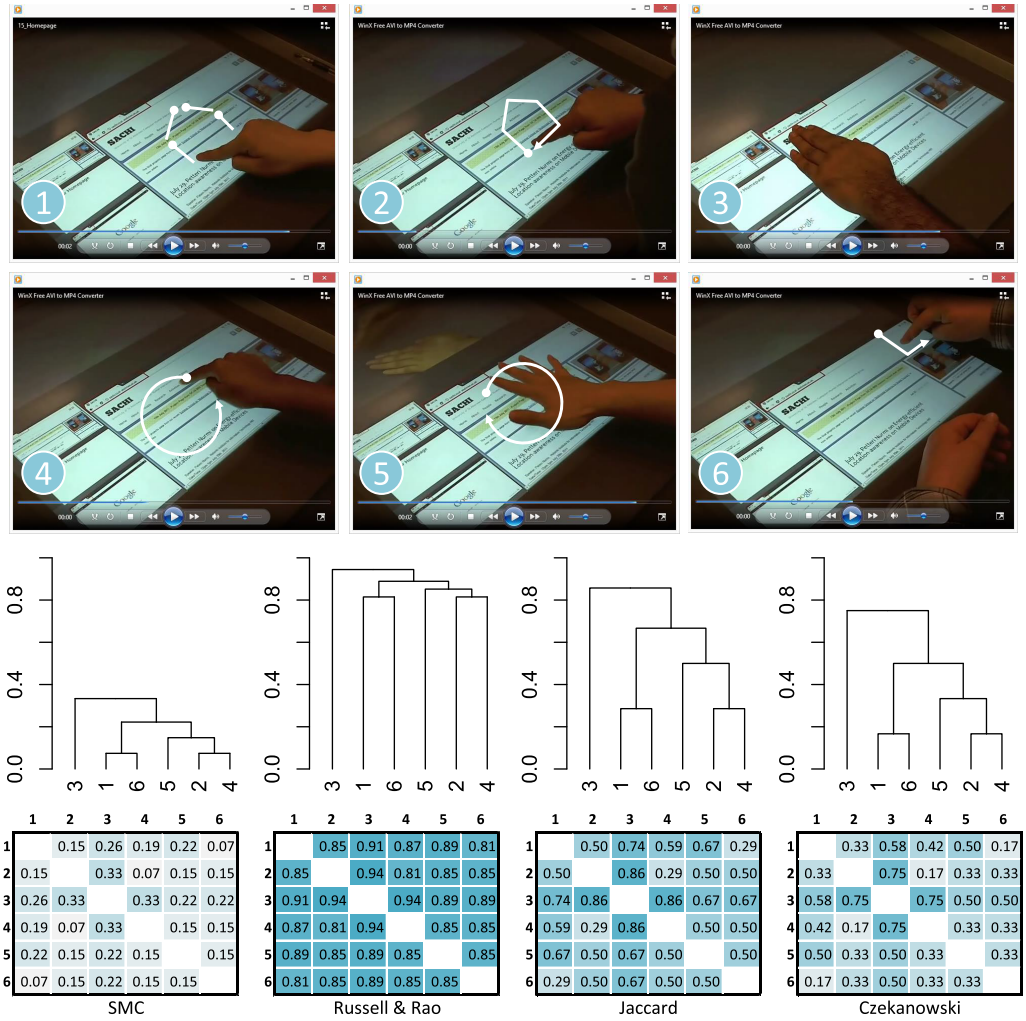

Fig. 8. Top: examples of gestures elicited from six participants in the memorability study [38, 80] to effect the "Homepage" action for a web browser. From top to bottom and left to right: (1) and (2) draw a house symbol, (3) lay flat hand on the tabletop, (4) draw a circle, (5) draw a circle and place the hand on the tabletop, and (6) small drawing in the top-right corner of the web browser. Our annotations, in white, show the geometrical shape of each gesture. Middle and bottom: dendrograms computed with the complete-link clustering method and four distinct dissimilarity functions.

These results strengthen our previous clarification regarding research question [$RQ_{1.1}$] about not correcting measures *A* and *AR* for chance agreement in end-user elicitation studies. Moreover, at this point, our analysis of chance agreement (and disagreement, respectively) should be put in a larger context to gain perspective. According to Gwet's [40] critical survey on inter-rater reliability studies, *"the idea of adjusting the percent agreement p_a for chance agreement is often controversial, and the definition of what constitutes chance agreement is part of the problem"* (p. 32). As we have shown, end-user elicitation studies are fundamentally different in their

assumptions from inter-rater reliability studies. Moreover, our previous analysis unveiled two important aspects: the non-transitive nature of the agreement relation and the existence of chance disagreement (i.e., false negatives). The first aspect is not present in inter-rater reliability, where transitivity is a consequence of raters picking their options from a codebook of fixed and usually nominal categories. The second aspect is not considered when applying inter-rater reliability coefficients to end-user elicitation [108], since those coefficients focus on chance agreement alone. However, both false positives and false negatives represent the basis for computing the sensitivity and specificity measures of performance of recognizers, which offer an overall view on the recognizers' discrimination capabilities. From this perspective, Uebersax [109] argues that *"in measuring accuracy of a diagnostic test, we don't correct sensitivity or specificity for the effects of chance; why do so in measuring rater agreement?"* as well as *"by considering both sensitivity and specificity together, there is no obvious, compelling need to correct for possible effects of chance."*

At this point, we have sufficient empirical and theoretical support to clarify research question [$RQ_{1.2}$], as follows:

> **Research Question [$RQ_{1.2}$]:** Is end-user elicitation the same thing as an inter-rater reliability study?
>
> **Clarification:** Despite apparent similarities, inter-rater reliability studies and end-user elicitation are fundamentally different in their goals, methods, and measures with respect to calculating agreement. Unlike inter-rater reliability studies, however, the list of categories or signs is not defined *a priori* in end-user elicitation, the agreement relation is not necessarily transitive, and the measures of agreement *A* [140, 141] and *AR* [32, 120, 121] incorporate both chance agreement and chance disagreement, where for every bit of chance agreement, there is a corresponding amount of chance disagreement to oppose it.

In the next section, we focus on the properties of the agreement relation in end-user elicitation studies, for which we provide further theoretical support.

5 TOLERANCE RELATIONS AND SPACES

We showed so far that (i) there are many causes for the differences observable in the proposals elicited from end users as revealed by our operational model (Figure 1), (ii) agreement can be reached in ways that are not always intuitive, because of the limited previous understanding of agreement formation in elicitation studies, and that incorporate the effects of both false positives and false negatives, and (iii) that the agreement relation is not necessarily transitive. These findings enabled us to clarify research questions [$RQ_{1.1}$] and [$RQ_{1.2}$] listed at the outset of our article.

In the following, we show that agreement can be further formalized in terms of tolerance relations and tolerance spaces [102, 145], and that the process of agreement formation is just one part of a bigger picture regarding our scientific understanding of human perception and action [85].

5.1 From Human Perception to a Mathematical Theory of Tolerance Relations

Our starting point is the concept of "tolerance," already hinted at in Equation (1) with our variable ϵ, which we connect in this section to the mathematical theory of "tolerance spaces" [102, 145].

At the end of the 19th century, Henri Poincaré [88, 89] formalized "sets of sensations" to characterize the physical spectrum of human perception, and concluded that similar perceptions can be described using a mathematical space where the concept of tolerance plays a key role. Poincaré made a distinction between *l'espace géométrique* and *l'espace représentatif* [88] or between *le continu mathématique* and *le continu physique* [89], where the latter, in both cases,

indexes human representations and sensations and is limited in resolution compared to the descriptive power of the former. However, the projection of the physical onto the mathematical continuum led to logical contradictions for Poincaré [89]: *"It has, for instance, been observed that a weight A of 10 grammes and a weight B of 11 grammes produced identical sensations, that the weight B could no longer be distinguished from a weight C of 12 grammes, but that the weight A was readily distinguished from the weight C. Thus the rough results of the experiments may be expressed by the following relations: A = B, B = C, A < C, which may be regarded as the formula of the physical continuum"* (pp. 27–28). By characterizing this fact as an *"intolerable disagreement with the law of contradiction, and [a] necessity of banishing this disagreement"* [89], Poincaré conceptualized the existence of regions in the mathematical continuum mapped to sets of sensations that aggregate perceptually indistinguishable objects. Even before Poincaré, Weber [131] had already noticed the imprecision of human senses to detect differences in the intensities of stimuli that are smaller than specific thresholds, referred to today as "difference limens" or "just noticeable differences." Weber formalized his observations in his eponymous law stating that the difference threshold divided by the original intensity of the stimulation is constant, i.e., the Weber constant.

However, Poincaré's insight represented the foundation for tolerance theory [102], and it was Zeeman [145] who properly defined the notion of a "tolerance relation" on a given set as a binary relation on the Cartesian product of that set that is *reflexive* and *symmetric*. Following this definition, a "tolerance space" is the set supplied with a tolerance relation. Moreover, a "perceptual tolerance space" applies tolerance spaces to the study of resemblances between perceived objects and sensations. In their overview of tolerance spaces applied to human perception, Peters and Wasilewski [85] defined *"[the] tolerance on a set [as a] mathematical structure that formalizes the idea of resemblance, i.e., the idea of being the same within some tolerance. Put another way, objects are considered near each other up to a small, allowable error"* (p. 211). Overall, the tolerance theory formalizes the idea of resemblance between objects up to some error [102], which is ϵ in our Equation (1) for proposals elicited from participants in end-user elicitation studies.

5.2 Agreement in End-User Elicitation is Indeed a Tolerance Relation

Our previous empirical results from Section 3 showed that the agreement relation α is not transitive [108, 120, 121]. Using tolerance theory, we can now enunciate that the agreement relation *is* a tolerance with reflexive and symmetric properties only, as follows:

(1) Reflexivity of agreement: $p_i \, \alpha \, p_i \; \forall p_i \in \mathcal{P}$

(2) Symmetry of agreement: $p_i \, \alpha \, p_j \Leftrightarrow p_j \, \alpha \, p_i \; \forall p_i, p_j \in \mathcal{P}$

Together with the set $\mathcal{P}$, the agreement relation α generates a tolerance space $< \mathcal{P}, \alpha >$. Similar to Poincaré's [89] mapping between the physical and mathematical continuum, $< \mathcal{P}, \alpha >$ contains sets of descriptions (i.e., signs) that are indistinguishable given the dissimilarity function δ and tolerance level ϵ (see Equation (1)), although their descriptions may be different.

Our previous use of the dissimilarity function δ did not impose any constraints on its properties. Our only assumption was that δ could be defined in a reasonable way to compute a small value when two proposals are similar and a larger value when they are less similar. A more proper way to define δ is to require that it holds the properties of a *metric*,[18] although this might be too constraining for some dissimilarity functions, such as DTW [54, 79, 106]. According to Sossinsky [102], any metric space determines tolerance relations with respect to some positive threshold ϵ. Thus, if $< \mathcal{P}, \delta >$ is a metric space and ϵ a real, positive value, then the relation α defined by

[18] δ is a metric if it satisfies the identity, symmetry, and the triangle inequality properties [99].

our Equation (1) represents a tolerance over $\mathcal{P}$ and generates the tolerance space $< \mathcal{P}, \alpha >$. This mathematical result provides theoretical support for the definitions used by the dissimilarity-consensus approach to agreement analysis in the end-user elicitation work by Vatavu [114], which is reflected in Equation (1). Moreover, Sossinsky [102] presents several examples of tolerance spaces, including one about information structured in graphs: if X is the set of vertices of a (nonoriented) graph, the relation "x and y are vertices of the same edge" is a tolerance, which connects the tolerance theory to a recent method from Ali et al. [3] for computing agreement based on clustering nodes to form subgraphs within larger fully connected graph structures.

> **Clarification (continuation):** In the previous section, we clarified research questions [$RQ_{1.1}$] and [$RQ_{1.2}$] in that measures of agreement A [140, 141] and AR [32, 120, 121] do not need to be corrected for chance agreement ([$RQ_{1.1}$]) and that inter-rater reliability and end-user elicitation studies make fundamentally different assumptions ([$RQ_{1.2}$]). The insight of this section regarding the agreement relation being a tolerance relation strengthens further the differences between the two types of studies: where agreement in inter-rater reliability is an *equivalence relation*, agreement in end-user elicitation takes the form of a *tolerance relation*.

The discussion so far has showed that end-user elicitation is different from inter-rater reliability studies and, consequently, needs specific measures to calculate and report agreement. Below, we continue our discussion with an overview of the measures of agreement that have been widely employed in end-user elicitation studies, highlighting their specific advantages and shortcomings in the light of our finding from this section that agreement is a tolerance relation.

6 AN OVERVIEW OF AGREEMENT CALCULATION IN END-USER ELICITATION

Agreement has been evaluated in various ways in the practice of end-user elicitation. Probably the simplest measure of quantifying agreement has been to count the frequency of proposals elicited from participants [51, 72, 103, 110, 112]. Other, more elaborate measures include the Jaccard similarity coefficient [73], Kendall's coefficient of concordance [123], and measures tailored to the specifics of the topic under examination [19, 75, 114, 128]. The latter category includes, for example, the *popularity* of user-defined techniques for associating smart devices [19] or measures devised to characterize agreement when multiple proposals are elicited from each participant, such as the *max-consensus* and *consensus-distinct ratio* [75] and *aggregator functions* [114]. Recently, Tsandilas [108] advocated the use of *coefficients of agreement* traditionally employed in inter-rater reliability studies, such as Cohen's κ [21], Fleiss' κ_F [34], or Krippendorff's α [59], to report the magnitude of agreement in end-user elicitation studies. The literature on inter-rater reliability encompasses a wide variety of such coefficients; see Gwet [40] for a critical survey on when and how to apply them. However, since inter-rater reliability studies operate on fundamentally different assumptions than end-user elicitation, directly adopting coefficients of agreement from inter-rater reliability to elicitation studies is debatable; see our in-depth discussion of this aspect in Section 4.

6.1 Quality Properties for Measures of Agreement

In the following, we review measures of agreement widely employed in end-user elicitation studies [120, 140, 141], and we offer a way to generalize them into one single, "all-purpose" measure. But first, we establish criteria to assess the *quality properties of any measure of agreement*:

(1) *Lower bound (LB)*: The measure admits a fixed lower bound that corresponds to perfect disagreement (e.g., 0) that is independent of the number of end users from which proposals are elicited and also independent of the actual proposals.
(2) *Upper bound (UB)*: The measure admits a fixed upper bound that corresponds to perfect agreement (e.g., 1 or 100%), independent of the number of end users and their proposals.
(3) *Interpretation (IN)*: The measure admits a simple and intuitive interpretation of its values.
(4) *Non-Transitivity (NT):* The measure can be employed to compute agreement between the elicited proposals when the agreement relation turns out to be non-transitive.

Ideal measures of agreement have fixed lower and upper bounds, intuitive interpretations, and are applicable to both transitive and non-transitive data. Next, we briefly overview the measures of agreement most used in end-user elicitation studies, in chronological order, and discuss their compliance with our quality properties.

6.2 2005: The Agreement Score *A*

The most widely used measure of agreement in end-user elicitation studies is probably the *Agreement score A* proposed by Wobbrock et al. [140] in 2005 as a practical way to evaluate numerically the guessability of symbolic input; see the first row of Table 2 for its formula and quality properties. The *A* measure was also employed in the first hand-gesture elicitation study published by Wobbrock et al. [141] in 2009. One strength of *A* is that it is remarkably easy to understand and calculate, even by hand. For example, if from a group of 20 elicited proposals, four subgroups of similar proposals of sizes 8, 6, 4, and 2 emerge such that all the proposals of each subgroup are in agreement, then *A* is the sum $(\frac{8}{20})^2 + (\frac{6}{20})^2 + (\frac{4}{20})^2 + (\frac{2}{20})^2 = .300$. *A* is upper bounded by 1, which is the case of perfect agreement when all the participants' proposals are the same or substantially similar [140]. The components of the sum represent the squared ratios of the size of each subgroup to the total size of the group. These ratios can be interpreted as the probabilities p_i that a new end user, not part of the study, would "guess" the same proposal as the participants forming the ith subgroup, e.g., p_1 is $\frac{8}{20} = 40\%$ and p_4 is $\frac{2}{20} = 10\%$ in our example. (It is easy to verify that $\sum p_i = 1$.) Thus, it is more likely that a new user will think of a command similar to what the eight participants of the first subgroup proposed than to the proposal of the subgroup of two. Although Wobbrock et al. [140, 141] did not present this interpretation of *A*, we believe it represents a useful clarification for practitioners.

Wobbrock et al. [140, 141] did not provide a theoretical basis for their *A* formula, but its straightforwardness and wide popular adoption offer a clue about its instinctual origins. In the support of this argument, we quote Good, who affirmed in a comment[19] to a 1982 paper [84] on the topic of evaluating diversity that *"if $p_1, p_2, \ldots, p_t$ are the probabilities of t mutually exclusive and exhaustive events, any statistician of this century who wanted a measure of homogeneity would have taken about two seconds to suggest $\sum_{i=1}^{t} p_i^2$, which I shall call ρ"* (p. 561). Furthermore, Ellerman [29] argues that the quantity $\sum p_i^2$ represents the probability of getting non-distinct values in two independent samplings of the random variable for which the probability distribution is $\{p_i\}$ (p. 129) and, thus, the sum is a measure of *homogeneity* or *concentration*. In his work, Ellerman was interested in logical entropy,[20] defined as $h(p) = 1 - \sum_{i=1}^{t} p_i^2$, which is the complementary of *A*. Ellerman [29] also presents a history of $h(p)$ and *A* being employed in various scientific fields with applications in cryptography, biostatistics, economics, and so on; see, for example, Gini [1992] or Friedman [1922]

[19]The comment was published together with Patil and Taillie's article [84] on page 561.
[20]Ellerman's article was brought to our attention by Tsandilas' work [108].

Table 2. An Overview of Agreement Calculation Formulae Frequently Employed in End-User Elicitation Studies

	Measure	Refcerence(s)	Original formula†	Agreement graph equivalent formula	Quality Properties‡ LB	UB	IN	NT	Numerical example§
Measures of agreement for single elicitation (one proposal per participant, according to the original implementation of end-user elicitation studies [140, 141])									
1	Agreement score (*A*)	Wobbrock et al. [2005] [140]; Wobbrock et al. [2009] [141]	$\sum_{P_i \subseteq P} \left(\frac{\lvert P_i \rvert}{\lvert P \rvert}\right)^2$	$\frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} a_{i,j}+1\right)$	–	✓	✓	–	$\left(\frac{8}{20}\right)^2+\left(\frac{6}{20}\right)^2+\left(\frac{4}{20}\right)^2+\left(\frac{2}{20}\right)^2 = .300$
2	Agreement rate (*AR*)	Stern et al. [2008] [103]; Findlater et al. [2012] [32]; Vatavu & Wobbrock [2015] [120]	$\frac{\sum_{P_i \subseteq P} \frac{1}{2}\lvert P_i \rvert (\lvert P_i \rvert - 1)}{\frac{1}{2}\lvert P \rvert (\lvert P \rvert - 1)}$	$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}}{N(N-1)}$	✓	✓	✓	–	$\frac{\frac{8\cdot(8-1)}{2}+\frac{6\cdot(6-1)}{2}+\frac{4\cdot(4-1)}{2}+\frac{2\cdot(2-1)}{2}}{\frac{20\cdot(20-1)}{2}} = .263$
3	Agreement rate (*AR*), alternative formulation	Vatavu & Wobbrock [2016] [121]	$\frac{\sum_{p=1}^{\lvert G_i \rvert}\sum_{q=p+1}^{\lvert G_i \rvert}\delta_{p,q}}{\frac{1}{2}\lvert G_i \rvert (\lvert G_i \rvert - 1)}$	$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}}{N(N-1)}$	✓	✓	✓	✓	$\frac{100}{20\cdot 19} = .263$
4	Consensus (*C*)	Vatavu [2019] [114]	$\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}[\delta(g_i,g_j)\le\tau]}{\frac{1}{2}N(N-1)}\cdot 100\%$	$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}[a_{i,j}\le\tau]}{N(N-1)}\cdot 100\%$	✓	✓	✓	✓	Depends on δ; see [114] for examples.
5	Growth rate (*r*)	Vatavu [2019] [114]	The growth rate of the growth curve of *C* as a function of τ using a logistic model	–	–	–	–	✓	Not calculated by hand; needs a computer; see [114] for examples.
Measures of agreement for repeated elicitation (multiple proposals per participant, e.g., production [77] or elicitation using multiple input modalities [75])¶									
1	Max-consensus	Morris [2012] [75]	Percent of participants suggesting the most popular proposed interaction	Cardinality of the largest connected component¶ divided by *N*	✓	✓	✓	–	32% for "open browser" [75]
2	Consensus-distinct ratio	Morris [2012] [75]	Percent of the distinct interactions that achieved a given consensus threshold	Number of connected components¶ with cardinality greater than the consensus threshold divided by the total number of connected components	✓	✓	✓	–	.300 for "select URL" [75]
3	Consensus (*C*)	Vatavu [2019]	$\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}[\zeta(\delta(g_{i,t},g_{j,u})\forall t,u)\le\tau]}{\frac{1}{2}N(N-1)}\cdot 100\%$	$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}[a_{i,j}\le\tau]}{N(N-1)}\cdot 100\%$¶	✓	✓	✓	✓	Depends on δ; see [114] for examples.
4	Growth rate (*r*)	Vatavu [2019] [114]	The growth rate of the growth curve of *C* as a function of τ using a logistic model	–	–	–	–	✓	Not calculated by hand; needs a computer; see [114] for examples.

†We used the same notations as the original authors: $\lvert P \rvert$ denotes the cardinality of the set of proposals in [140]; P_i [140] and G_i [120] denote the ith subgroup of proposals that are in agreement; g_i and g_j represent two proposals elicited from two participants for some referent [114]; p and q denote participants in [121]; $\delta_{p,q}$ is Kronecker's symbol that evaluates to either 1 or 0, depending whether participants p and q are in agreement or not [121]; Δ is a dissimilarity function and τ denotes the tolerance level in [114]; ζ is an aggregating function, such as min, max, or mean [114].

‡LB - admits fixed lower bound; UB - admits fixed upper bound; IN - simple interpretation; NT - formula works for non-transitive agreement relations.

§The numerical example is provided for a set of $N = 20$ participants or, equivalently, $\lvert P \rvert = 20$ proposals forming four subgroups of sizes 8, 6, 4, and 2 participants/proposals in agreement.

¶In this work, we address single elicitation only, following the original implementation of Wobbrock et al. [140, 141]. For repeated elicitation, further work is needed to formalize the agreement graph.

referenced in [29]. From this perspective, A can be viewed as the *degree of order* in the set of proposals elicited from the participants of the study as opposed to the disorder-like interpretation commonly accepted for entropy.

6.3 2015: The Agreement Rate *AR*

An inconvenience of the A measure [140, 141] is that it never reaches 0, even when all participants are in disagreement with each other. The minimum value attainable by A is $\frac{1}{N}$, which depends on the number of proposals put forward by participants. This aspect is inconvenient, because a study with N= 20 participants in full disagreement leads to A = .050, while the same study with N = 40 participants also in full disagreement will yield A = .025. Wobbrock et al. [140] challenged this discrepancy with the fact that *"the lower bound [of A] is non-zero because even when all proposals disagree, each one trivially agrees with itself"* (p. 1871). By using our interpretation of the agreement relation as a tolerance (see Section 5), we can now formally assert that Wobbrock et al. [140] incorporated the reflexivity property directly into their definition of the A measure. Unfortunately, this approach led to a measure with an unstable zero-agreement level.

However, there is a simple way to make agreement fall exactly in the closed interval $[0, 1]$, shown by Vatavu and Wobbrock [120] and referred to as the *Agreement Rate* (*AR*); see the 2nd row of Table 2. *AR* is the ratio between the number of pairs of participants in agreement and the total number of pairs that could be in agreement. Considering our previous example with the partition $20 = 8 + 6 + 4 + 2$, the first subgroup contains $\frac{8 \cdot (8-1)}{2}$ pairs of proposals in agreement, the second one $\frac{6 \cdot (6-1)}{2}$ pairs, and so on, which makes $AR = \frac{28+15+6+1}{190} = .263$. Another advantage of *AR* is that it admits a simple interpretation as a percentage, i.e., 26.3% pairs of proposals are in agreement. This straightforward interpretation alleviates problems caused by reporting A incorrectly, e.g., May et al. [74] reported A scores as percentages, where they should not be interpreted as such.[21]

Just like for the A measure [140, 141], the straightforwardness of defining agreement in the manner advocated by Vatavu and Wobbrock [120] was intuited before 2015. Findlater et al. [32] used the *AR* formula in a 2012 elicitation study for touchscreen keyboards (but without making the explicit mathematical connection between *AR* and *A*), and Stern et al. [103] employed it, among other measures, in a 2008 paper regarding the intuitiveness of hand gestures (but the denominator was not fixed to $N \cdot (N-1)/2$, but rather varied according to the specific proposals elicited from the participants; see Stern et al. [103, p. 100] for an example). In fact, the *AR* formula seems to have been employed in many other fields.[22] In his comment[23] to Patil and Taillie [84], Good referred to the *AR* formula as an unbiased estimate of A (p. 561).

AR has inspired the invention of other measures to characterize agreement, such as the DR [120, 121] and the ***Coagreement Rate*** (**CR**) between referents for within-subjects designs [120] and for independent groups of participants [121]. *AR* has also inspired adaptations, such as a measure of the consistency of users interacting on multiple devices in contrast to multiple users employing a single device [128]. Both A and *AR* were implemented by GECKo [6], a software tool that reports users' consistency of stroke-gesture articulation on touchscreens.

[21] Note that although the initial definition of the A measure employed percentages [140], subsequent uses of A removed them [3, 141].
[22] A conscientious observation from Tsandilas [108].
[23] The comment was published together with Patil and Taillie's article [84] on page 561.

6.4 2016: An Alternative Definition for *AR*

Both the original formulations of *A* and *AR* [32, 120, 140, 141] have assumed the transitivity property for the agreement relation, which we debunked in Section 3. For example, when *A* computes $(\frac{8}{20})^2$, it is implied that all the participants forming the subgroup of size 8 are in agreement with each other [140]. And when *AR* computes $\frac{8\cdot(8-1)}{2}$, the same assumption is implied [32, 120]. However, our analysis above showed that the agreement relation is not transitive and it should be viewed as a tolerance relation [102, 145] that generates a tolerance space over the set of distinct proposals elicited from participants. In other words, the agreement relation is reflexive and symmetric, but not transitive. This finding *limits the application of the A and AR formulae just to those situations where transitivity can be verified* due to some specificity of the study, analysis process, or application domain; e.g., when ϵ is always zero and the similarity test is actually a test for equality, as it was for the design of EdgeWrite stroke-gestures [140, 142], which encoded letters as sequences of corners represented as integers. Inter-rater reliability studies with nominal categories are another example.

However, *AR* allows for a different expression that computes the same numerical results as the original formula [120], but without assuming transitivity. Vatavu and Wobbrock [121] introduced this alternative formula in 2016 in the context of formalizing between-subjects designs for end-user elicitation studies; see the 3rd row of Table 2. Instead of computing the pairs of proposals in agreement from a subgroup by employing multiplication (e.g., $\frac{8\cdot(8-1)}{2}$), the alternative formula *counts* how many pairs of proposals are in agreement. This alternative formulation makes it possible to use *AR* even when the agreement relation is not transitive.

6.5 2019: The Consensus Rate *C*

Vatavu [114] introduced the concept of a *dissimilarity function* to compute agreement and defined the *Consensus Rate* (*C*). The reasoning behind *C* is that a dissimilarity function δ can be applied to all pairs of proposals and the result compared to a tolerance value τ (which we call ϵ in this article). If the dissimilarity value is less than the tolerance, then one pair of proposals in agreement gets counted. Just like *AR*, *C* is obtained by dividing the total number of pairs in agreement to the total number of pairs of proposals that could be in agreement; see the 4th row of Table 2. One advantage of *C* is that it enables great flexibility (in terms of δ and τ), while keeping all the properties of *AR*: fixed lower and upper bounds 0 and 1, simple interpretation as a percentage, and independence of the assumption of transitivity of agreement. Moreover, Vatavu [114] was interested in calculating agreement beyond any fixed set of criteria (represented by τ) and defined *C* as a function of τ, which he modeled using growth curves and logistic functions. The growth rate *r* of the logistic model was used as a measure of agreement to analyze whole-body movements [114] and touchscreen stroke-gestures [115]; see the 5th row of Table 2. Because *C* and *r* use dissimilarity functions, they cannot be easily calculated by hand, but software was made available.[24]

6.6 Measures of Agreement for Repeated Elicitation

Some implementations of end-user elicitation studies, especially those that employ production as a practical way to reduce the influence of legacy bias [77], elicit multiple proposals per participant, i.e., interaction synonyms, in some cases using different input modalities [75]. We refer to these cases as repeated elicitation to differentiate them from the original method introduced by

[24] http://www.eed.usv.ro/~vatavu/projects/DissimilarityConsensus.

Wobbrock et al. [140, 141], where one proposal was elicited per participant. Morris' [75] "web on the wall" elicitation study and Vatavu's [114] examination of whole-body gestures produced by young children are two relevant examples of repeated end-user elicitation.

To correctly analyze and report agreement when multiple proposals are elicited per participant, new measures have been introduced. For example, Morris [75] defined the *max-consensus* metric as the percent of participants suggesting the most popular proposed interaction for a given referent and the *consensus-distinct ratio* as the percent of the distinct interactions proposed for a given referent that achieved a given consensus threshold among participants. These two metrics offer practitioners flexibility in informing the design of possible interactions in their prototypes, i.e., *"if the goal is to design a system with a single, highly guessable command per referent, then max-consensus may be more important, whereas if the goal is to understand diversity of opinion surrounding a referent, or conceptual complexity of a referent ... consensus-distinct ratio may be more helpful"* [75, p. 98]. Vatavu [114] introduced aggregate dissimilarity functions ζ that applied a dissimilarity function δ for all pairs of proposals elicited from the same participant to compute an aggregate statistic, such as the minimum, maximum, or average of the dissimilarity values computed by δ. The numerical result computed by ζ was employed in the formula of the consensus rate C [114].

Note that in our approach to formalizing the classification step presented in Section 2.3.2, we considered one proposal elicited per participant, following the original end-user elicitation method [140, 141]. This approach is convenient since all proposals for a given referent are independent of each other, ensuring the independent and identically distributed trials assumption required by the statistical tests that we evaluate in Section 9. In the following, we continue with this premise (one proposal per participant), and leave aggregate measures of agreement and the corresponding form of the agreement graph for detailed examination in future work. Nevertheless, we include the measures from Morris [75] and Vatavu [114] in Table 2 (the repeated elicitation section) to offer readers the complete perspective on measures of agreement calculation.

6.7 Summary

So far, we discussed in this section the most influential measures of agreement employed in end-user elicitation studies, addressing their limitations and improvements over the years, up to their most recent formulations involving dissimilarity functions. In this context with several measures of agreement available, it may be difficult for practitioners to determine which one to use in their work. For example, as we have shown, Tsandilas [108] criticized the fact that *"the A and AR indices do not take into account that agreement between participants can occur by chance"* (p. 18:2) and suggested the use of coefficients of agreement from inter-rater reliability to correct for chance agreement. Du et al. [27] favored *AR* to *A* since when *"compared to the widely used agreement score introduced by [Wobbrock et al., 2005], the agreement rates are more accurate measures of agreement."* Felberbaum and Lanir [31] acknowledged that while *"there are several ways to calculate agreement scores [..], we followed the original method presented in [Wobbrock et al., 2009]."* Magrofuoco and Vanderdonckt [69] reported both *A* and *AR* in their GELICIT platform for conducting gesture elicitation studies. Ali et al. [3] used the original *A* formula [140] on the basis that *"subsequent variations to this formula have been published but all are similar"* (p. 178). Moreover, the various notations from Table 2 can be confusing to newcomers to end-user elicitation. In this context, clarifications are needed regarding what measure of agreement to use. To address this aspect, we discuss next "agreement graphs," a useful concept that we employ to show that *A, AR, and C are multiple facets of one single, all-purpose measure of agreement.*

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 34 & 36 & 51 & 85 & 70 & 76 \\ 34 & 0 & 93 & 65 & 59 & 25 & 69 \\ 36 & 93 & 0 & 12 & 33 & 65 & 13 \\ 51 & 65 & 12 & 0 & 75 & 26 & 60 \\ 85 & 59 & 33 & 75 & 0 & 22 & 42 \\ 70 & 25 & 65 & 26 & 22 & 0 & 42 \\ 76 & 69 & 13 & 60 & 42 & 42 & 0 \end{bmatrix}$$

Fig. 9. An agreement graph can be represented as an adjacency matrix with 0 and 1's (left) or as a cost matrix with real values (right). *Note:* the adjacency matrix from the left was obtained by thresholding the cost matrix shown on the right with the tolerance $\epsilon = 45$.

7 THE AGREEMENT GRAPH

The agreement graph is a simple, yet useful concept to visualize agreement between elicited proposals, where vertexes correspond to proposals and edges implement the agreement relation, i.e., an edge connects vertexes i and j of the agreement graph if the proposals p_i and p_j elicited from participants P_i and P_j are in agreement over some referent.

Let $\mathbb{A} = \{a_{i,j}\}$ be a square matrix of size $N \cdot N$, where N is the number of participants and $a_{i,j}$ quantifies the agreement between proposals p_i and p_j elicited from participants P_i and P_j. When $a_{i,j}$ values are restricted to 0 (disagreement) and 1 (agreement), $\mathbb{A}$ takes the form of an adjacency matrix; see Figure 9, left for an example. However, $a_{i,j}$ can also take any value from $\mathbb{R}$, in which case the graph is weighted and the $\mathbb{A}$ matrix is referred to as a cost matrix; see Figure 9, right. In that case, the values $a_{i,j}$ represent the dissimilarity between the proposals of participants P_i and P_j. A cost matrix can be easily transformed into an adjacency matrix by applying a tolerance ϵ to its weights, i.e., if $a_{i,j} \leq \epsilon$, the result is 1 and 0 otherwise.

It is worth noting that all the measures of agreement from Table 2 (except the growth rate r) can be expressed in terms of the agreement matrix $\mathbb{A}$; see the corresponding definitions from Table 2. For example, the number of edges of the agreement graph corresponds to the number of pairs of proposals in agreement and, by dividing that number to the total number of edges in a complete graph, we get the *AR* measure [32, 120]. This ratio is referred to in graph theory under the name "graph density" [95, p. 29].

7.1 An All-Purpose *AR* Measure: The ϵ-Agreement Rate (AR_ϵ)

We stress that our definition of AR_ϵ, introduced in Equation (2) from Section 2 to operationalize the five-step model for general end-user elicitation studies in HCI, represents a generalization of all previous measures of agreement outlined in Table 2, as follows. The connection to C [114] is direct. (The difference in notations, i.e., τ from [114] becomes ϵ, comes from the new connection between end-user elicitation studies and the mathematical theory of tolerance spaces [102], where the notation "ϵ" has been traditionally used.) Once the $[\cdot]$ expressions from Equation (2) have been evaluated to either 1 or 0, AR_ϵ reduces to the non-transitive expression of *AR* [121], which computes the same numerical values as the original *AR* formula [32, 120]. Finally, *AR* and *A* are linearly related:

$$AR = \frac{N}{N-1} \cdot A - \frac{1}{N-1}. \tag{6}$$

For example, for our previous example, $AR = \frac{20}{19} \cdot .300 - \frac{1}{19} = .263$. Based on these insights, we can now clarify research questions [$RQ_{2.1}$] and [$RQ_{2.2}$]:

Research Question [RQ$_{2.1}$]: How do various measures of agreement relate to each other?

Clarification: Measures of agreement A [140, 141] and AR [32, 120] are linearly related. Overall, AR is less optimistic compared to A (i.e., the inequality $AR \leq A$ is always true), but it has a stable zero-agreement level and admits an expression [121] that can be used to compute the degree of agreement even when the agreement relation is non-transitive. The consensus rate C and growth rate r [114] belong to a different class of measures of agreement: while A and AR need the similarity criteria to be available and agreed upon beforehand, C and r are criteria-independent. Finally, AR_ϵ encapsulates all these previous measures under one single mathematical formulation.

Research Question [RQ$_{2.2}$]: Which measure(s) of agreement should one use for end-user elicitation studies?

Clarification: Both A [140, 141] and AR [32, 120] are simple to understand and straightforward to calculate by hand, and can be employed when the similarity criteria to judge whether two proposals are identical, equivalent, or similar can be unambiguously defined and applied. AR has the advantage of a stable zero-agreement level that does not depend on the number of proposals put forth for each referent in the study. When the similarity criteria are not straightforward or are loosely defined and, thus, subjective to misinterpretation, C, r, and growth curve modeling [114] should be employed. When the elicitation study takes the form of an inter-rater reliability study (i.e., participants choose from predefined list of proposals or categories), the coefficients of agreement κ [21], κ_F [34], or others [40] are recommended along with confidence intervals [108]. C and AR_ϵ were specifically designed to compute agreement for proposals stored in a computational representation, which is one of our recommendations for future end-user elicitation studies, when possible.

7.2 The Agreement Rate as a Mean

We continue our discussion about the measures of agreement frequently employed in end-user elicitation, and prove in the following that the AR measure [32, 120, 121], an extension of A [140, 141] and an instance of AR_ϵ (Equation (2)), represents a measure of location, i.e., a mean. We start our proof by defining the agreement score a_i of the proposal elicited from participant P_i as the number of other proposals with whom P_i's proposal is in agreement. For example, if participant P_3 from a set of $N = 20$ is in agreement with P_1, P_7, P_{12}, and P_{19} over some referent, then the agreement score of P_3 is $a_3 = 4$. For convenience, we normalize a_i in the $[0, 1]$ interval by dividing it by $N-1$, which is the maximum number of participants with whom P_i could potentially be in agreement. We use $\hat{a}_i$ to denote this normalized score. For our example, $\hat{a}_3 = 4/(20 - 1) = .211$. Consider now the set of values $\hat{a}_i$ for all participants P_i, $i = 1..N$:

$$\left\{\hat{a}_1 = \frac{a_1}{N-1}, \hat{a}_2 = \frac{a_2}{N-1}, \ldots, \hat{a}_N = \frac{a_N}{N-1}\right\}.$$

A practical problem of interest is to characterize the level of agreement for the group of N participants from their individual agreement scores, for which the immediate option is to take the average of values $\hat{a}_i$. This approach leads to the formula of AR exactly:

$$\frac{\sum_{i=1}^{N} \hat{a}_i}{N} = \frac{\sum_{i=1}^{N} \frac{a_i}{N-1}}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \delta_{i,j}}{N \cdot (N-1)} = AR. \tag{7}$$

To the best of our knowledge, this is the first time that this derivation of AR has been presented from the perspective of elicitation studies, despite its multiple origins [32, 103, 114, 120]. (Instead,

AR has been interpreted as the percentage of pairs of participants in agreement out of all pairs of participants [120, 121].) In reality, however, *AR* is an arithmetic mean, as the leftmost part of Equation (7) clearly shows. This insight has implications for agreement analysis, which we discuss in Section 8. But first, we provide clarifications regarding the qualitative interpretation of the agreement rate.

7.3 How to Interpret the Magnitudes of Agreement Rates?

Vatavu and Wobbrock [120] proposed recommendations for the qualitative interpretation of the magnitude of the *AR* measure, as follows: values less or equal to .10 can be interpreted as "low agreement"; values between .10 and .30 as "medium agreement"; between .30 and .50 as "high agreement"; and values larger than .50 as "very high agreement" [120, p. 1332]. Observing an agreement rate above the .10 margin means that at least 10% of all the pairs of participants' proposals are in agreement or, according to our ϵ formalism, at least 10% of the pairwise comparisons between proposals are less than the tolerance ϵ. Observing an agreement rate above .50 means that more than 50% of the pairwise comparisons between participants' proposals are less than the tolerance. These recommendations were inspired by Cohen's [23] guidelines for interpreting effect sizes for the most common statistical inference tests used in psychological research—see Table 1 from [23, p. 157]—but also by an analysis of the magnitudes of agreement rates reported by published end-user elicitation studies, for which the average *AR* value was .261 and by a probabilistic analysis of the *AR* distribution; see [120, p. 1332]. To simplify their analysis, Vatavu and Wobbrock [120] assumed partitions of agreement equally probable, while acknowledging that their assumption may not hold for all referents because of legacy bias. In a follow-up work, Vatavu and Wobbrock [121] provided a formalism on which to compute the frequency of occurrence of each partition, i.e., there are 35 ways in which 7 participants can form two groups of sizes 3 and 4 in agreement and 105 different ways in which the same 7 participants can form three groups of 4, 2, and 1 participants in agreement.

Tsandilas [108] criticized the use of these guidelines in the practice of end-user elicitation by claiming that they can lead to overoptimistic conclusions about the true level of agreement reached by the participants of an end-user elicitation study. Also, Tsandilas criticized the probabilistic framework that was used to arrive at these guidelines because that framework did not incorporate bias (an assumption acknowledged in [120]; see above) and because of the assumption of equally probable partitions (later addressed in [121]). Unfortunately, Tsandilas did not provide an alternative solution to interpret the magnitude of agreement rates. This unresolved state of recommendations from Vatavu and Wobbrock [120] and criticisms from Tsandilas [108] warrants clarification for practitioners that wish to interpret agreement rates.

To clarify this issue, we must understand what influences the magnitude of the agreement rate measure. So far in this article, we showed that the criteria used to evaluate the similarity between proposals elicited from participants (implemented in Equation (1) by the variable ϵ) have a direct influence on the magnitude of the agreement rate (Equation (2)). Figure 4 illustrated this dependency between ϵ and AR_ϵ for the 15 referents of the public gesture elicitation dataset of Vatavu [114]: as ϵ increased (i.e., criteria are more permissive), AR_ϵ increased as well until it reached 100%. Vatavu [114] notes that practitioners of end-user elicitation compromise somewhere between the extremes of 0% and 100%, *"favoring some criteria and dismissing others, but it is evident ... how the choice of the clustering criteria affects the magnitude of reported consensus"* (p. 224:3). An observation from Tsandilas [108] complements this perspective: *"comparing agreement rates across different studies can be misleading because chance agreement can be high for some studies and low for others"* (p. 18:22), although it brings into consideration only chance agreement and not chance

disagreement; see Section 4. Tsandilas [108] also argues that *"agreement and disagreement do not generally occur with the same probability. For example, full agreement [...] is very unlikely to occur when participants randomly choose from a very large set of possible signs. Full disagreement [...] is far more likely to occur in this case."* (p. 18:21). This observation is contained by the more general result characterizing the dependency between ϵ and AR_ϵ, e.g., if the criteria are strict and conservative, full agreement has less probability to occur, but when criteria are more liberal, full disagreement is less likely. So far, the facts are that (i) different end-user elicitation studies use different criteria to evaluate the similarity of proposals elicited from the study participants, (ii) the criteria influence the magnitude of the agreement rate and, furthermore, (iii) the criteria themselves are often not stated *a priori*, but are intuited emergently and inductively from the proposals that arise. In this context, Vatavu [114] proposed a holistic approach in which agreement is interpreted as a function of the criteria (ϵ) used to judge the similarity of elicited proposals. While this option is recommendable for advanced models of end-user elicitation and agreement analysis, as we elaborate on in Section 8 and recommend in our guidelines from Section 10, what should practitioners do regarding the opposing views of Vatavu and Wobbrock [120] and Tsandilas [108]? To understand more about the use of guidelines to interpret agreement rates, we turn to Cohen [23].

Cohen [23] was motivated to suggest qualitative scales, which he characterized as *subjectively defined* (p. 156). In the history of statistics, several recommendations to interpret the magnitude of test statistics have been proposed and employed for turning quantitative numerical results into qualitative interpretations. For instance, Landis and Koch [60] recommended using the threshold .20 to denote "slight" agreement for Cohen's κ, the interval [.21, .40] to denote "fair" agreement, [.41, .60] for "moderate" agreement, and so on. However, the overall answer to the question *"When is κ big enough?"* seems to be that *"'No one value of κ can be regarded as universally acceptable"*; see Bakeman [10, p. 1]. This answer has come up frequently in the literature [5, 11, 12]. However, despite contrasting recommendations in the statistics literature, interpreting κ coefficients and effect sizes based on Cohen's guidelines represents common practice; even Cohen [23] noted that *"although the definitions [for small, medium, and large effect sizes] were made subjectively, with some early minor adjustments, these conventions have been fixed since the 1977 edition of SPABS [Statistic Power Analysis for the Behavioral Sciences] and have come into general use"* (p. 156).

In this context, what should practitioners of end-user elicitation do? Should they not interpret any longer the magnitudes of agreement rates? Or, are there perhaps other margins more suited to this purpose? Our view is that the margins .10, .30, and .50 are nevertheless convenient and useful,[25] but their adoption should not be done blindly, but rather used in an informed manner, just like the clustering criteria by which proposals are grouped into signs are selected in an informed manner. The bottom line is that researchers and practitioners need to make responsible decisions for their studies based on the magnitudes of the agreement values they measure, decisions that should be consistent with the particulars of their studies and analysis approaches, e.g., the criteria used to cluster similar proposals into signs [114]. At this point, we have accumulated sufficient information to provide the following clarification for research question [$RQ_{2.3}$]:

[25]Note that Vatavu and Wobbrock [120] did not just derive their margins from Cohen's .10, .30, and .50 thresholds, but they also surveyed the magnitudes of agreement rates reported by 18 gesture elicitation studies published until 2015 and used the results to support these margins. For example, the average *AR* values of the studies examined was .261; only 3 out of the 18 average *AR* values (16%) were above .40, 12 out of 18 (66%) were under .30, and 2 were near .10.

Research Question [$RQ_{2.3}$]: How to interpret the magnitude of agreement in end-user elicitation studies?

Clarification: The margins .10, .30, and .50 proposed by Vatavu and Wobbrock [120] for *low*, *medium*, and *high* agreement are reasonable guidelines for interpreting the agreement results of end-user elicitation by connecting to the practice of interpreting effect sizes for a variety of test statistics [23]. Nevertheless, these margins should be used with care, as indicated by Tsandilas [108], to prevent drawing incorrect and misleading conclusions. Especially when it is the same practitioners that set the criteria to judge the similarity of elicited proposals and that employ those criteria to calculate and report agreement. The margins .10, .30, and .50 can be used as rough guidelines for the qualitative interpretations of agreement rates, but not necessarily to draw final conclusions about the outcomes of end-user elicitation studies or to compare the agreement rates reported by different studies. Those conclusions should consider the particularities of the specific investigations, application domains, contexts of the studies (e.g., participants, criteria, bias), and agreement analysis approaches.

We continue our discussion regarding how to calculate and report agreement by identifying models for agreement analysis in end-user elicitation studies.

8 MODELS FOR AGREEMENT ANALYSIS

In a 2008 paper, Stern et al. [103] intuited that acquiring gestures from participants is not a trivial process, and considered three methods, which they called "Direct Video Capture" (i.e., gestures performed by participants are recorded on camera), "Gesture Image Database" (gestures are picked by participants from an existing catalog), and "Coded Gesture Entry" (participants generate gestures by manipulating some parameters, such as the flexion of the fingers to form a specific hand pose). From these approaches, Stern et al. [103] decided to go with *"the coded gesture entry method [..] as one combining reasonable time demands, and accuracy in gesture labeling"* (p. 98). Later, in a 2014 paper on replicating gesture elicitation studies, Nebeling et al. [82] found value in discriminating between the "human" and "system recognizers", i.e., *"dividing the elicitation process into two parts, first using a human recogniser and then a system-based recogniser. This mixed-initiative design allowed participants to first think without any technological constraints, but then also get a feel for the technology and perhaps reconsider their interaction proposals to make them feasible for implementation"* (p. 23). Thus, there are several ways in which prior work has approached agreement analysis, some of which we have already hinted to in our discussion so far. In this section, we identify precisely and discuss models for agreement analysis in end-user elicitation studies.

8.1 The Expert, Codebook, and Computer Models for Agreement Analysis

Based on the original guessability maximization method of Wobbrock et al. [140], the approaches of Stern et al. [103] and Nebeling et al. [82], Tsandilas' [108] formalization of bias [76], and previous work on expert [77] and crowd [3] similarity judgments and automatic computation of agreement [114], we propose the following models for agreement analysis in elicitation studies:

8.1.1 The "expert" Model. In this model, the knowledge about which proposals are similar is defined by an "expert." The expert evaluates the proposals and determines agreement or disagreement [3]. An interesting recent work showed that experts' evaluations can be reliably reproduced from the similarity judgments of a crowd [3]. Grijincu et al. [38] is another example where the crowd was employed, this time to code individual characteristics of the elicited gestures, e.g.,

whether the hand pose used during the articulation of a gesture was spread, flat, and so on. In such cases, the expert model still applies as it is the crowd, as a whole, that embeds the tacit knowledge regarding which proposals are similar and which are not.

8.1.2 The "codebook" Model. In this model, experts are not available and the task of evaluating the similarity between proposals and determining agreement or disagreement is left to a non-expert member of the research or design team (e.g., a study facilitator or data analyst). This model is by far the most ubiquitous in published elicitation studies, where the authors do the agreement analysis themselves.

The first step in applying this model consists of creating a *codebook* that exhaustively enumerates all the properties relevant for the application domain for which proposals are elicited. This is usually implemented in the form of a taxonomy. For example, Ruiz et al. [93] classified motion gestures by their physical characteristics using three criteria: (1) kinematic impulse (low, moderate, and high), (2) dimension (single-axis, tri-axis, six-axis), and (3) complexity (simple and compound). Each category had a specific definition, e.g., a "moderate" impulse denoted a gesture with jerk between 3 and $6\,m/s^3$ [93, p. 202]. Other taxonomies have been proposed [18, 26, 27, 36, 139], but they essentially represent variations of the *form-nature-binding-flow* taxonomy of the first hand-gesture elicitation study published by Wobbrock et al. [141]. A simpler alternative to using a full taxonomy is to define a set of characteristics. For example, the coding manual of Troiano et al. [107] contained definitions of gestures according to the actions that compose the gestures and the number of fingers used for performing those actions.

In the second step of the codebook model, the experimenter characterizes the properties of the elicited proposals according to the codebook. Proposals are classified according to the categories of the taxonomy, usually by two or more persons, and an inter-rater reliability coefficient, such as κ from Equation (4), is calculated [18] to confirm the reliability of independent ratings. (Note that this is an intermediate step and it doesn't relate to agreement calculation between proposals. At this stage, researchers learn about the traits of their participants' proposals, i.e., our step 6 from Figure 1. For example, Chen et al. [18] discovered that nearly 60% of the gestures proposed by their participants for ear-based input were metaphorical in nature, whereas only 12% were symbolic.) The codebook can be *finite* or *infinite* in terms of the number of distinct proposals that it can accommodate. For example, the physical characteristics taxonomy of Ruiz et al. [93] mentioned above allow a number of $3 \times 3 \times 2 = 18$ distinct characterizations for motion gestures. However, if the first criterion of their taxonomy (kinematic impulse) were numerical instead of ordinal (i.e., actual jerk values retained), the number of distinct characterizations would be essentially infinite. Other examples of criteria measured on the interval or ratio scale generating infinite codebooks include the production time of a gesture or the amplitude of the gesture movement. A specific case of a finite codebook employs binary characteristics only, e.g., whether the left hand was used to produce the gesture, whether the gesture shape is symmetric, and so on. In that case, a proposal can be represented as a binary array [66, 67].

These two steps allow variations. For example, while it is more sensible to define the codebook before the study, some studies have created it during the analysis process [73, 107], where the coding manual was updated each time a new proposal did not match the existing definitions, akin to open coding in grounded theory [104]. Yet in other studies, the codebook was reduced to a set of rules of thumb, e.g., Piumsomboon et al. [87] defined similar gestures as *"having consistent directionality although the gesture[s] had been performed with different static hand poses"* (p. 958). Other approaches revisit the original classification as more information becomes available during the process, e.g., after constructing an affinity diagram, Chen et al. [18] noted that more factors can impact ear-based interactions and regrouped the elicited gestures.

8.1.3 The "computer" Model. In this model, a computer program is employed to automatically compute the dissimilarity between any two elicited proposals [114]. This approach saves considerable time for the analysis of proposals elicited from participants compared to the previous two models, and also helps reduce human error and subjectivity in deciding which proposals are similar and which are not. This model requires proposals to be captured in a numerical, computational form (e.g., whole-body gestures captured using Kinect [114], hand poses captured with the Leap Motion controller [123] or with a numerical glove [122], or strokes captured by a touch-sensitive surface [143]) as well as a dissimilarity function to be defined and implemented for the specific representation in which proposals are captured (e.g., DTW [106] and point-cloud matching [116] represent popular examples of dissimilarity functions employed for gesture classification).

The *computer* model also applies to situations where testing for agreement is straightforward due to the way in which eligible proposals are defined, but the acquisition of proposals is implemented by a computer. For example, in the original EdgeWrite system [140, 142], a stroke gesture was defined by an objective sequence of corners within a square, e.g., top-left, top-right, and bottom-right corresponded to the letter "t". Thus, stroke gestures could be represented as a sequence of integers corresponding to those corners, e.g., 124, and testing for agreement was reduced to merely comparing two integers. In that case, gesture recognition and calculation of agreement were straightforward once the gesture had been captured, and agreement in this case was strict integer equality, i.e., ϵ is exactly zero. The *computer* model applies here not because two numbers can be compared by a computer in order to recognize a gesture, but because the acquisition device (formally referred to in this model as the "computer") detects when a specific corner has been reached. And this detection process takes place automatically, without any human intervention.

According to the observations from Nebeling et al. [82], what we refer to as the *computer* model implements the best way to transfer the results of an elicitation study into an actual user interface or interactive system, since the same "computer" (read: sensor, detection method, dissimilarity function, algorithm, or device) is used for both agreement calculation during the design, prototyping, and testing stages, but also in the final system that is deployed to end users.

Note that our formalization of the classification step from Section 2.3.2 does not impose any requirements on the δ function other than non-negativity. Therefore, δ can equally represent the outcome of an expert's tacit reasoning, the use of a codebook, and the values computed by a dissimilarity function devised for gesture recognition in the computer model. Based on the our discussion thus far, we can now address research questions [$RQ_{3.1}$] and [$RQ_{3.2}$]:

> **Research Question [$RQ_{3.1}$]:** Are there viable models for the analysis of elicited proposals?
>
> **Clarification:** There are three possible models for evaluating the results of end-user elicitation studies: the *expert*, the *codebook*, and the *computer* model. Of these, the *codebook* model has been the most used in published gesture elicitation studies.
>
> **Research Question [$RQ_{3.2}$]:** Which model should one adopt for the analysis of elicited proposals?
>
> **Clarification:** We recommend the *computer* model for reasons of efficiency and replicability of results, but also due to straightforward transfer of the results of an end-user elicitation study into an actual system. However, we acknowledge that the *computer* model is not always economical, as it might require, for example, the implementation of a working system in order to effect this model.

8.2 Bias in End-User Elicitation Studies

An interesting discussion concerns bias that has been observed in end-user elicitation studies [76, 94, 108], e.g., when participants rely on their past experience with user interfaces to propose commands for new interactive systems and technologies ("It's a Windows World" [141]), that may affect the magnitude of reported agreement among participants' proposals because of agreement occurring by chance [108]. Morris et al. [76] coined the term "legacy bias" to refer to situations where users' proposals in gesture elicitation studies in particular, and in end-user elicitation in general, are influenced (biased) by their experience with other interfaces and technologies, such as WIMP interfaces. They noted: *"legacy bias limits the potential of user elicitation methodologies for producing interactions that take full advantage of the possibilities and requirements of emerging application domains, form factors, and sensing capabilities"* [76, p. 42]. To address this aspect, the authors proposed production, priming, and partners as three techniques aimed at reducing legacy bias, which were evaluated in [43]. Beyond legacy bias, Ruiz et al. [94] identified "performance bias" corresponding to the situation where the artificial setting in which the end-user elicitation study is conducted could prevent users' consideration of long-term aspects of the interactions they propose, such as fatigue for gesture input, and argued for the use of "soft constraints" during elicitation to make users aware of such aspects. Finally, in the attempt to isolate bias, Tsandilas [108] connected bias with chance agreement and considered bias as having a central role in his approach based on inter-rater reliability methods: *"a key argument of our analysis is that any kind of bias can deceive researchers about how participants agree on signs. The agreement measures that we recommend remove the effect of bias"* (p. 18:7).

It appears that the community has been concerned that legacy bias (1) could limit the effectiveness of the end-user elicitation method in discovering proposals in whatever new interactive context is being examined [76], but also (2) that bias may be related to agreement occurring by chance and, therefore, it should be isolated and removed from the computation of agreement measures, just like chance agreement is corrected for in inter-rater reliability studies [108]. While the first concern is appropriate and addressable with changes in the elicitation procedure [76, 94, 136], the second concern is debatable. Legacy bias means that users rely on their past experience when acting in a new context, whereas agreement by chance means that users disregard the referent but still propose something, in the lack of a suitable proposal, that may turn out to be the same answer as another user doing the same, i.e., the *"overall tendency of some signs to appear more frequently than others independently of the actual referents"* [108, p. 18:6] (emphasis ours). Therefore, while legacy bias identifies exactly the source of a proposal as a match for the prompted referent, agreement by chance ignores any source and considers the effect of chance alone. This distinction is important because legacy bias is not necessarily a bad thing (remember that its downside is that it may limit the effectiveness of discovering new commands, but in some cases, familiar commands might be desirable). In fact, Köpsel and Bubalo [56] argued that designers could benefit from legacy bias to smooth users' transition toward new forms of interaction with computing systems. And, in fact, Dingler et al. [26] noted: *"instead of reducing such biases and expectations as proposed by Morris et al. [76] we can take them into consideration when unifying a gesture set"* (p. 10). In that case, legacy bias was useful to the researchers to discover gestures that were consistent and, thus, transferable across several types of devices. Williams and Ortega [136] argued that legacy gestures reduce the learning curve for interactions with new technologies, and pinpointed that legacy gestures are not effective when they do not match the capabilities of the new interaction space. To address this downside, the authors proposed "evolutionary gestures," a technique that supports the design of gestures that build on the benefits of legacy bias. In their survey of gesture elicitation studies for mid-air interaction, Vogiatzidakis and Koutsabasis [127] noted: *"It becomes apparent that tackling*

or not the legacy bias is a matter of design decision and it mainly depends on whether the end product/system is meant to be a walk-up-and-use system or a system that would take full advantage of the novel interaction techniques" (p. 8). In this context, should legacy bias be connected to agreement simply occurring by chance (i.e., the participant that is blind to the referent [108]) and its effect removed from the agreement measure as in inter-rater reliability studies [108]? Or should legacy bias be seen as an intrinsic part of how users genuinely agree about their proposals for suitable interactions, even when those proposals are influenced by the users' past experiences with other interactive systems?

To address this debate, let us consider what happens with each referent in the original elicitation procedure [141]: after the user *confirms* that they understood the effect of the referent, the user *proposes* a command for that referent, and is asked to *rate*, among other things, the *suitability* of the proposed command for the effect they have just witnessed as a goodness-of-fit rating [141]. Sometimes, the experiment protocol collects other kinds of ratings, such as about usability and social comfort [18] or subjective satisfaction [68]. Suppose that the command is legacy biased, but the users rated it highly suitable for the referent and the same command and rating were observed for other users as well. Should this effect of legacy bias be removed from the computation of agreement? Of course not, because those users genuinely agree and believe in the goodness of fit between their proposals and the referent. And other users, in the future, may likely come up with the same command, which would make the system intuitive and readily usable, i.e., "design for guessability" [140, 141]. Suppose now that the users rate their proposals a poor fit for the effect they witnessed, but still their proposals match. At this point, the experimenter can decide to find out more about the poor fit ratings and encourage the participants to do better, e.g., via production or partnering [76], and, in the case when such methods fail to increase the goodness of fit, the experimenter will need to make a decision about whether these low quality proposals are actually worth considering for their study or not. The experimenter could ponder at whether the agreement between the two low quality proposals represents the effect of chance alone [108], augmented perhaps by legacy bias. *However, imagine the reversed situation: Two participants appear to disagree in their proposals, but the two proposals are rated low again so there is little confidence, from the part of both participants, in how good a fit those proposals are to effect the referent. Could this be a situation of disagreement by chance? How could the experimenter tell, without being deceived by an agreement that appears too high [108] or, as we suggest, one that appears too low?* One reasonable solution would be ignoring low quality proposals according to users' own ratings of how well they did, but this would affect not only situations of chance, either agreement or disagreement, but also proposals that are simply a less good fit. Instead, focusing on procedural changes to the elicitation method by encouraging production of multiple variants of proposals for the same referent [76] or the evolutionary gesture procedure [136] that leverages the legacy bias to arrive at new proposals, better suited for the new context and technology under study, is the best way to elicit proposals that users believe represent a good fit for referents. And when goodness of fit is high, there can be no argument for correcting the level of agreement, since that agreement is likely to generalize beyond the sample to the larger user population.

In this context, procedural bias (e.g., not giving the correct instructions, an artificial environment to collect proposals) or classification bias (e.g., experimenters being biased in selecting some categories from their codebook) should probably receive more focus. Grijincu et al. [38], Ruiz and Vogel [94], Tsandilas [108], and Vatavu [114] discussed such types of biases. For example, Grijincu et al. [38] acknowledged classification bias and proposed a crowd-sourcing approach to prevent it: *"In most previous studies that classify user-defined gestures the authors themselves classified the gesture sets... This approach does not scale well as the number of gestures go up and is also subject*

to a possible strong author bias" (p. 27). Next, we connect these aspects to our three models of agreement analysis: expert, codebook, and computer.

8.3 Modeling Bias

Tsandilas [108] was the first to model bias numerically in end-user elicitation studies. His discussion considered various sources of bias, including legacy bias [75], performance bias [93] (as one kind of bias from a larger category called procedural bias [108]), and bias that may be induced during the classification of proposals into signs. He concluded that *"biases are additive, so overall bias will be observed as an imbalance in the distribution of signs across all referents"* [108, p. 18:6]. To this end, Tsandilas employed two probability functions, the Zipf–Mandelbrot and Discrete Half-Normal, to model bias for or against specific signs, while affirming that other functions, such as generated by the interpolation of the former, could be equally employed. A bias function, as used by Tsandilas, gives the probability of selecting the kth most probable sign when ignoring or having no information about the referent. Thus, Tsandilas' approach to modeling bias applies directly to signs as the final result, i.e., the *"overall tendency of some signs to appear more frequently than others, independently of the actual referents"* (p. 18:6). However, our model of end-user elicitation (Figure 1), enables localization of these sources of bias: end-user bias [75] occurs at the levels of the mental model and proposal articulation, from where it should be further picked up in the description of the proposal by the recording apparatus employed in the study. Procedural bias [94, 108] affects the proposal step (e.g., because of the artificial, out-of-context setup, incomplete or confusing instructions delivered to participants, inadequate sampling of participants) and the description of the proposal (e.g., because to the choice of a recording apparatus with insufficient capability to capture the full subtlety of the participant's proposal). Ultimately, the type of model to be used for agreement analysis (expert, codebook, and computer) may introduce specific kinds of bias, which we acknowledge as being distinct in their nature, as follows:

(1) In the *expert* model, the criteria employed for evaluating the similarity between proposals are embedded *within* the expert, and getting access to those criteria may be difficult. For example, expert knowledge may prove difficult to articulate explicitly or, in the case of crowd workers [3], information about their similarity judgments may not be available at fine levels of detail, but rather in the form of simple yes/no similarity votes [3]. As no other information regarding agreement formation is available in the *expert* model except for the final set of signs, expert bias toward specific signs can be modeled with monotonically decreasing probability distribution functions, such as Zipf–Mandelbrot or others, as proposed by Tsandilas [108].
(2) In the *codebook* model, bias can be present especially if the codebook is finite and has few categories. In that case, the end-user elicitation study can be modeled as an inter-rater reliability study [40], since the data analyst practically "rates" each proposal according to a predefined set of categories or properties. For each category, bias applies to some extent and, thus, can be again modeled using monotonically decreasing functions as the ones used by Tsandilas [108] for the final signs.
(3) However, when the codebook is not finite or when proposals are captured in a numerical, computational form, as in the *computer* model, end-user elicitation studies are fundamentally different from inter-rater reliability studies since there is no rating taking place from predefined categories.

Figure 10 illustrates the various sources of bias in end-user elicitation in relation to Figure 1, the three models of agreement analysis, and the two types of outcomes of an end-user elicitation study, signs and traits. Next, we introduce simulation procedures for each model of analysis, extending

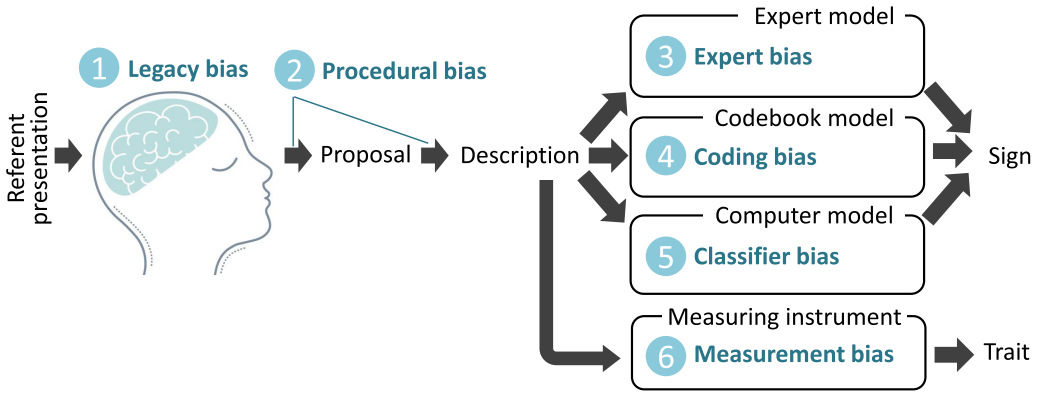

Fig. 10. Various sources of bias in end-user elicitation studies and relation to the end-user elicitation model (Figure 1) with signs and traits as outcomes.

Tsandilas' [108] approach from final signs to bias affecting actual proposals (Figures 1 and 10) with a formalization that accommodates the use of dissimilarity functions:

(1) *The expert model.* Following Tsandilas [108], we model bias using the Zipf–Mandelbrot distribution with the exponent $s = 2$ and using nine distinct values for the B parameter, which were derived experimentally by Tsandilas to correspond to agreement rates AR ranging from .10 to .20, .30, . . . , and .90, respectively; see [108, p. 18:10].

(2) *The codebook model.* We consider a codebook containing k criteria, where each criterion takes categorical values, e.g., the kinematic impulse of smartphone motion gestures can be either low, moderate, or high [93], or the number of fingers touching the screen for gestures performed on deformable displays can be either one, two, or at least three [107]. Using this approach, a proposal can be formalized as an array of k categories, $\{\xi_1, \xi_2, \ldots, \xi_k\}$, where ξ_i is the category corresponding to the ith criterion. We implement bias by using the same Zipf–Mandelbrot distribution, but this time applied for each criterion individually. The difference with the previous model is that we now simulate bias for proposals instead of the final signs, and each criterion can be affected by a different amount of bias corresponding to a different value of the B parameter of the Zipf–Mandelbrot distribution. We evaluate the dissimilarity between two proposals using a distance function based on the probabilistic information-theoretic approach proposed by Lin [64], which seems to be one of the best performing distances for categorical data [14].

(3) *The codebook model (variant).* In this variant of the *codebook* model, we represent proposals as binary arrays, i.e., $\xi_i \in \{0, 1\}$. This specific representation enables us to model bias by sampling from another distribution type, Bernoulli, characterized by the probability of success. We use Lin's distance [64] to evaluate the dissimilarity between proposals.

(4) *The computer model.* We model each proposal as a feature array of size k with values randomly sampled from normal distributions with means zero and varying SDs. We employ the nine SD values derived by Tsandilas [108] to correspond to AR from .10 to .20, . . . , and .90, respectively. In this model, bias is the tendency to sample values closer to the mean zero. We evaluate the dissimilarity between two proposals using several dissimilarity functions: (1) Euclidean, (2) Manhattan, (3) Minkowski with $p = 0.5$, and (4) squared Euclidean. This selection of distances enables a wide range of simulation options that correspond to real-world scenarios, e.g., the $1 gesture recognizer [143] employs the Euclidean distance; Minkowski

Table 3. Agreement Rate (*AR*) Absolute Errors and Peak Non-Transitivity Rate (NTR_1) Results for Our Simulations with 13 Distinct Combinations of Models of Analysis, Bias, and Dissimilarity Functions

Model/Distribution/Dissimilarity	k	$\lvert AR_{target} - AR \rvert$ mean (sd)	Peak NTR_1 mean (sd)
Expert/Zipf–Mandelbrot/Identity	1	.038 (.010)	0.0% (0.0%)
Codebook/Zipf–Mandelbrot/Lin	5	.003 (.006)	32.1% (1.9%)
Codebook/Bernoulli/Lin	5	.025 (.013)	34.3% (2.1%)
Computer/Normal/Minkowski (p = 0.5)	5	.000 (.000)	31.9% (1.1%)
Computer/Normal/Minkowski (p = 1)	5	.000 (.000)	28.8% (1.5%)
Computer/Normal/Minkowski (p = 2)	5	.000 (.000)	26.5% (1.9%)
Computer/Normal/Squared Euclidean	5	.000 (.000)	26.9% (1.7%)
Codebook/Zipf–Mandelbrot/Lin	10	.000 (.000)	34.6% (1.4%)
Codebook/Bernoulli/Lin	10	.001 (.001)	37.7% (1.2%)
Computer/Normal/Minkowski (p = 0.5)	10	.000 (.000)	33.6% (0.8%)
Computer/Normal/Minkowski (p = 1)	10	.000 (.000)	30.1% (1.4%)
Computer/Normal/Minkowski (p = 2)	10	.000 (.000)	27.3% (1.6%)
Computer/Normal/Squared Euclidean	10	.000 (.000)	27.6% (1.6%)

In this table, k is the number of criteria from the codebook for the codebook analysis model or the number of features to represent a gesture in the computer model.

represents a generalization of both the Euclidean ($p = 2$) and Manhattan ($p = 1$); Euclidean and Manhattan are metrics (i.e., they respect the triangle inequality), while Minkowski with $p = 0.5$ and squared Euclidean are not.

Except for the *expert* model, where dissimilarity functions are not explicitly defined but rather implicitly embedded in the expert, each approach can be used to generate a dissimilarity matrix $\mathbb{A}$ for a population of proposals. For the *expert* model, this matrix is actually an adjacency matrix and the agreement rate can be computed directly; see Table 2. For the other models, the matrix can be transformed into an adjacency matrix by using a tolerance level ϵ *equal to the quantile of the values from* $\mathbb{A}$ *that corresponds to the target AR*. We provide an example to illustrate our approach. Consider that we wish to generate a population of 100 proposals with a target *AR* of .735 by using the codebook model. Assume that our codebook contains a number of $k = 5$ criteria, for which the probabilities of individual categories are modeled using a Zipf–Mandelbrot distribution specified by given α and B's, e.g., $\alpha = 0.5$ and $B = 0.306$. We generate a proposal by randomly drawing samples from each distribution to generate each of the proposal's k categories. We repeatedly draw proposals until we reach our target population size. The next step is to compute a dissimilarity matrix based on these proposals using a dissimilarity function, in this case one that applies to nominal data. (The next section presents results for various such dissimilarity functions.) The dissimilarity matrix is then used to generate an adjacency matrix in order to compute the agreement rate, as we showed above. To obtain an exact value of the agreement rate, we select ϵ to be the 735th quantile of the values from the dissimilarity matrix, which assures us than 73.5% of the values will be less than ϵ. Computing the agreement rate on the resulting adjacency matrix will result in $AR = .735$ exactly. The computation is similar for the other models of agreement analysis, where other probability distributions and dissimilarity measures are used, as indicated above.

Table 3 shows the results of applying each simulation method to generate populations with *AR* from .10 to .20, . . . , and .90, and using $k = 5$ and $k = 10$ categories for the *codebook* and *computer* models (and $k = 1$ for the *expert* model). For *codebook*, these choices for k are a compromise between "too few" and "too many" criteria for a practical end-user elicitation study. For instance, the majority of taxonomies [36, 94, 141] used to characterize the properties of elicited proposals employ a number of dimensions around $k = 5$. For the *computer* model, $k = 10$ approaches the number of $k = 13$ features employed by the popular Rubine gesture recognizer [92].

Results given in Table 3 are averaged from 11,700 simulation runs: 13 (models) × 9 (target AR values) × 100 repetitions. Table 3 shows that our simulation procedures are exact (the absolute errors with respect to the target agreement rates are zero), except for the *expert* model using the Zipf–Mandelbrot distribution, which we borrowed from Tsandilas [108] for comparison purposes, and except the *codebook* and the Bernoulli distribution with $k = 5$, for which there were too few categories ($2^5 = 32$) that generated ties, affecting our quantile method; Bernoulli with $k = 10$ was not a problem, however; see Table 3. Besides being exact, our simulations generate populations for which the non-transitivity of the agreement relation, measured using Peak NTR_1[26] falls between 23.2% and 38.4%, values which are similar to those observed in actual data; see Table 1.

Our procedure can be applied for any value $|\mathcal{P}|$ of the size of the population that one wishes to generate with an exact agreement rate, e.g., $AR = .72$, $AR = .725$, or $AR = .7253$. The precision of the target AR, expressed with its number of decimals, depends on how many distinct values are available in the dissimilarity matrix to compute the quantile corresponding to the required level of precision, i.e., the 72nd quantile out of 100 for $AR = .72$ or the 7253rd quantile of 10,000 for $AR = .7253$. As $|\mathcal{P}|$ gets larger, finer precision can be attained. For practical needs represented by agreement rates of .10, .20, . . . , and .90, respectively, the results from Table 3 show that $|\mathcal{P}| = 100$ is enough to generate these exact agreement rates. However, the theoretical space and time complexity to generate the dissimilarity matrix, compute the quantile, and generate the corresponding adjacency matrix is $O(|\mathcal{P}|^2)$ and the simulations that we present in the next section need samples to be drawn from infinite populations, $|\mathcal{P}| \to \infty$. Nevertheless, there is a simple way to apply our method with good results for populations that are infinite. In that case, we estimate the values of the quantiles that determine exact agreement rates in the infinite population, e.g., $AR = .20$, with the quantiles derived for small populations, for which computations are manageable, such as $|\mathcal{P}| = 100$, as given in Table 3. In practice, we found this approach to work well; see Table 4 that shows the mean agreement rates computed for two independent samples of $N = 20$ proposals drawn repeatedly (117,000 times = 9 target AR's × 13 models × 1,000 repetitions) from infinite populations corresponding to the 13 models from Table 3. We expect even better estimations of the quantiles for the infinite population and, correspondingly, AR values closer to the targets as well as lower SD for the random samples, when informing those estimations from computations for finite $|\mathcal{P}|$ values larger than 100, but Table 4 already shows good results for practical purposes.[27] The next section details our simulation procedures for Type I error rates and power of statistical tests for agreement rates and within-subjects and between-subjects designs.

9 STATISTICAL TESTS FOR AGREEMENT RATES

Vatavu and Wobbrock [120, 121] were the first to discuss statistical inference for end-user elicitation studies, and introduced the V_{rd} [120] and V_b [121] tests for within- and between-subjects designs. In end-user elicitation, statistical tests are useful to help the researcher understand the observed difference in the magnitude of two or more agreement rates corresponding to several referents (for within-subjects designs) or the same referent and multiple user groups (for between-subjects designs). The V_{rd} test was proposed as an adaptation of Cochran's Q test [20] to elicitation data and addressed within-subjects designs, e.g., to be able to compare the agreement rates obtained by the participants of an elicitation study with two referents, such as "volume up" and

[26] Peak NTR_1 is defined as the maximum number of tuples of three participants, out of all distinct tuples $\binom{N}{3}$ for which the transitivity property is met for AR_ϵ, when ϵ varies from 0 to 1; see Section 3.

[27] Note that the simulations from Tsandilas [108] do not employ populations with exact agreement rates either. However, for finite-sized populations, our method produces exact agreement rates.

Table 4. Mean Agreement Rates, AR_1 and AR_2, For Two Independent Groups of Size $N = 20$ Repeatedly Drawn from Infinite Populations with Various Distributions with Quantiles for Target Agreement Rates .100 to .900 Estimated from Populations of Finite Size 100 as in Table 3

Target AR	Num. trials	AR_1 mean (sd)	AR_2 mean (sd)
.100	13,000	.105 (.044)	.106 (.045)
.200	13,000	.201 (.059)	.206 (.059)
.300	13,000	.301 (.076)	.304 (.076)
.400	13,000	.402 (.086)	.404 (.087)
.500	13,000	.499 (.091)	.502 (.090)
.600	13,000	.595 (.093)	.602 (.095)
.700	13,000	.700 (.087)	.700 (.090)
.800	13,000	.803 (.076)	.802 (.079)
.900	13,000	.901 (.059)	.901 (.059)

Table 5. Overview of Statistical Tests Proposed for End-User Elicitation Data

Test	Reference	Scope	Example
V_{rd}	[120]	within-subjects designs	$V_{rd(3, N=80)} = 121.737$, $p = .001$
V_b	[121]	between-subjects designs	$V_{b(2, N=20)} = 74.938$, $p = .028$
Confidence intervals (jackknife)	[108]	within-subjects designs	$\Delta\kappa_F = .04$ 95%, CI= [−.06, .14]
Confidence intervals (bootstrapping)	[108]	between-subjects designs	$\Delta\kappa_F = .06$, 95% CI = [−.11, .16]

"volume down." The V_b test was devised in analogy with Fisher's exact test [33] and introduced for between-subjects elicitation study designs, e.g., for comparing the agreement rates achieved by two independent user groups for "volume up." Recently, V_{rd} and V_b were re-evaluated by Tsandilas [108] as having large Type I error rates and, as an alternative, Tsandilas suggested bootstrapping methods that *"can be used to produce variance estimates, standard errors and confidence intervals for almost any agreement index, including agreement rates, κ coefficients, and agreement specific to categories"* (p. 18:29). Table 5 presents an overview of these tests.

These contrasting results from the literature are very likely to confuse practitioners. Therefore, we present in this section *extensive evaluation results for 16 statistical tests* applied to the *AR* measure of agreement for both within- and between-subjects experimental designs. Since *AR* is an instance of AR_ϵ for a specific ϵ value (Equation (2)), our results apply to the AR_ϵ measure as well. Statistical tests for *AR* have been evaluated in previous work [108, 121] using Monte Carlo procedures, where populations of participants were simulated, random samples selected, and Type I errors estimated. In the following, we discuss practical aspects for such simulations in connection to our three models: *expert*, *codebook*, and *computer*. Also, beyond previous work [108, 121], we report results on statistical power as well. Note that, unlike Type I errors that are fixed based on an alpha threshold, power depends on the size of the sample or the difference that is expected to be detected.

9.1 Between-Subjects Elicitation Studies

Our Monte Carlo procedure for between-subjects designs is as follows:

(1) Let $\mathcal{M}$ be a simulation method represented by a specific model of analysis (e.g., *codebook*), a given distribution of the characteristics of the proposals (e.g., Bernoulli for proposals encoded as binary vectors), and a dissimilarity function to compare proposals (e.g., Lin [64])

that is used to generate random samples from an infinite population of proposals. We compute tolerance values ϵ's for specified AR's from the set $\{.10, .20, \ldots, .90\}$ based on a finite population[28] of size $|\mathcal{P}| = 100$ with proposals generated using method $\mathcal{M}$. These ϵ's will be used in the next steps as estimates for the ϵ's that determine exact agreement rates $\{.10, .20, \ldots, .90\}$ in the infinite population, as discussed in Section 8.3.

(2) Sample two independent groups of $N = 20$ proposals[29] from the infinite population specified by the model $\mathcal{M}$ and use the tolerance values ϵ's computed at step (1) to arrive at the agreement matrices for the two groups.

(3) Compute the agreement rates for the two samples, AR_1 and AR_2, and run statistical inference tests. For each test, count a Type I error when the reported p-value is less than the critical significance levels of .01 and .05, respectively.

(4) Repeat steps 1 to 3 for 100 times.

To evaluate statistical power, we modify step 1 by computing two sets of ϵ's from two finite populations of $|\mathcal{P}| = 100$ proposals each, having two distinct exact agreement rates with an absolute difference of at least .10. We modify step 2 by sampling one group of $N = 20$ proposals from each of the two infinite populations according to method $\mathcal{M}$.

Previous work evaluated only one statistical test [121] or two tests at most for AR, e.g., V_b vs. bootstrapping [108]. However, our revelation from Section 6 that AR is a mean enables us to consider and evaluate a wide range of statistical tests based on two samples $\{\hat{a}_i\}$, such as the popular t-test or its nonparametric analogs, as follows:

(1) *The independent-samples t-test.* Since the agreement rate AR represents the mean of $\{\hat{a}_i\}$ (Equation (7)), the t-test can be directly used to compare the two means AR_1 and AR_2. For this test, we used the `R` implementation of the `t.test` function.[30] Note that this test assumes that observations are independent, which does not always occur for the AR measure when interpreted as a mean, even in the presence of non-transitivity. Therefore, we expect larger Type I errors for this test, but we include it anyway in our simulations to verify such assumptions and to present them in a clear manner to readers.

(2) *Welch's variant of the independent-samples t-test* that does not assume homogeneity of variances. We used the `R` function `t.test` with the `var.equal` argument set to `FALSE`.

(3) *The Wilcoxon–Mann–Whitney test*[31] implemented by the `wmw` function from Rand Wilcox's "Rallfun" `R` library of robust statistics [133]. This test computes the probability that a randomly sampled observation $\hat{a}_i$ from the first group is less than a randomly sampled observation $\hat{a}_j$ from the second group; see Wilcox [132, p. 349].

(4) *The Brunner–Munzel* version of the Wilcoxon–Mann–Whitney test, implemented by the `bmp` function from the "Rallfun" library [133], and recommended for situations where tied values and heteroscedasticity might occur [132, p. 355].

(5) *Cliff's test*, implemented by the `cidv2` function from the "Rallfun" `R` library [133]. Cliff's test improves on Wilcoxon–Mann–Whitney with a heteroscedastic confidence interval that applies to situations when the wrong standard error is used by the latter, which results in poor power [132, p. 352].

[28] Since we consider one proposal elicited per participant, the size of the population of proposals is also the size of the population of all possible participants.

[29] Twenty participants has been the commonly accepted practice in the literature for running end-user elicitation studies.

[30] https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/t.test.

[31] Also known as the the Mann–Whitney U test [71].

(6) *The Kolmogorov–Smirnov test* [55, 100] implemented by the ks function from [133]. The test evaluates the hypothesis that the two groups have identical $\hat{a}_i$ distributions.

(7) *The V_b test* proposed by Vatavu and Wobbrock [121].

(8) *The V_b test with the conditional margins assumption* implemented by the AGATe toolkit [121].[32] This test, which we will refer to as $V_b^\star$ to differentiate it from V_b, works on the same probabilistic formalism as V_b, but implements the conditional margins of Fisher's exact test [33, pp. 99–101]. (Specifically, the sum from Equation 8 from [121] is restricted to only those $e_1, e_2, \ldots, e_k$ that sum to $a_1 + a_2 + \cdots + a_k$, i.e., the marginal sums of the contingency table are conditioned.) We include this version in our evaluation since Tsandilas [108] evaluated only V_b,[33] but AGATe actually reports $V_b^\star$ by default.[34]

(9) *The percentile bootstrap method* described in Wilcox [132, pp. 332–335], which we implemented in R for $\hat{a}_i$ with B = 2000 bootstrap samples.[35] Tsandilas [108] also implemented and evaluated bootstrapping.

(10) *The bootstrap-t method* described in Wilcox [132, p. 399], which we implemented in R for $\hat{a}_i$ and B = 2000 bootstrap samples.

Overall, we report results from 117,000 simulation trials (10 tests × 13 simulation conditions × 9 *AR* target values × 100 repetitions) for Type I error rate estimations, and from 468,000 trials (10 tests × 13 simulation conditions × 36 pairs of *AR* values[36] × 100 repetitions) for estimations of statistical power.

Table 6 shows Type I error rates and power estimations for $\alpha = .05$ and $\alpha = .01$. Only the percentile bootstrap and bootstrap-*t* methods seem to control the Type I error rate, with the percentile bootstrap showing better performance. The other tests have Type I errors between .142 and .338, most likely explainable by the fact that $\hat{a}_i$ values (Equation (7)) are not completely independent, even when the transitivity of agreement is relaxed. Detailed simulation results are given in Table 7. With a minor exception occurring for the *expert* model, the percentile bootstrap controls the Type I error rate very well.

A special note concerns the V_b test [121], which is criticized in Tsandilas [108]. Our simulations for V_b show similar Type I error rates for the expert/Zipf–Mandelbrot tested in [108] and slightly better performance for the other simulation conditions. (This is not surprising, since the Discrete Half Normal distribution, also evaluated in [108], showed lower Type I errors for V_b as well.) However, $V_b^\star$ showed much better performance, ranking third across the 10 tests that we evaluated. $V_b^\star$, reported by default in AGATe [121], was unfortunately not evaluated by Tsandilas [108]. The conditional margins version $V_b^\star$ test has better control of the Type I error rate (.153 on average for $\alpha = .05$ and .142 for $\alpha = .01$), but at the cost of low power as well (only 17% and 12%, respectively); see Table 6.

9.2 Within-Subjects Elicitation Studies

Our simulation procedure to evaluate statistical tests for within-subjects designs of end-user elicitation studies is as follows:

[32] Available from http://depts.washington.edu/acelab/proj/dollar/agate.html.
[33] According to the implementation available from https://agreement.lri.fr/.
[34] See http://depts.washington.edu/acelab/proj/dollar/agate.html.
[35] Wilcox's [132] implementations use the default value B = 2000.
[36] For independent groups with different agreement rates, the *AR* of the first population, AR_1, varies from .10 to .80 in increments of .10. For the second population, AR_2 varies from AR_1+.10 to .90 in increments of .10. Overall, there are 36 combinations of (AR_1, AR_2) so that $|AR_2\text{-}AR_1| \geq .10$.

Table 6. Mean Type I Errors and Power for Between-Subjects Designs Reported across Combinations of Models, Distributions, and Dissimilarity Functions

	Test	Abbr.	Type I error rate		Power	
			$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
1.	Independent-samples *t*-test	t	.304	.169	.852	.793
2.	Welch's *t*-test	Wt	.302	.166	.852	.792
3.	Mann–Whitney	MW	.335	.200	.866	.805
4.	Cliff	Cliff	.327	.194	.863	.802
5.	Brunner–Munzel	BM	.340	.218	.843	.782
6.	Kolmogorov–Smirnov	KS	.314	.195	.849	.780
7.	V_b	V_b	.338	.198	.879	.817
8.	V_b with conditional margins	$V_b^\star$	.153	.142	.174	.127
9.	★**percentile bootstrap**	**pb**	**.042**	**.013**	.653	.533
10.	★**bootstrap-*t***	**bt**	**.061**	**.061**†	.675	.675†

For details, see Table 7. *Note:* stars ★ highlight the best performing statistical tests.
†For $\alpha = .05$ and $\alpha = .01$, the Type I error rates and power are identical for the bootstrap-*t*, since rejection is based on confidence intervals and no actual *p*-values are computed.

(1) Let $\mathcal{M}$ be a simulation method represented by a specific model of analysis (e.g., *computer*), a given distribution of the characteristics of the proposals (e.g., normal for proposals encoded as numerical features), and a dissimilarity function to compare proposals (e.g., Minkowski with $p = 2$) that is used to generate proposals for random referents from an infinite population of referents. We compute tolerance values ϵ's for specified *AR*'s from the set $\{.10, .20, \ldots, .90\}$ based on 100 finite populations of referents, for which proposals are generated using method $\mathcal{M}$. Unlike the procedure for between-subjects designs, our goal this time is to work with a population of referents for which the mean agreement rate across all the referents is fixed. The ϵ's generated at this step will be used next as estimates for the ϵ's that determine exact agreement rates $\{.10, .20, \ldots, .90\}$ in the infinite population, as discussed in Section 8.3.

(2) Sample two dependent groups of $N = 20$ proposals corresponding to two referents from the infinite population of referents.

(3) Compute *AR* for the two samples and run statistical tests. Count a Type I error when the reported *p*-value is less than the significance levels of .01 and .05, respectively.

(4) Repeat steps 1 to 3 for 100 times.

Note the difference with respect to the simulation procedure employed for between-subjects designs in the previous subsection. Instead of simulating populations of participants (an approach equally followed by Vatavu and Wobbrock [121] and Tsandilas [108]), populations of referents are simulated instead, since the goal now is to compare the same participants for two different referents, with *the assumption that the two referents come from the same population.* Since Vatavu and Wobbrock [120] did not evaluate their V_{rd} test and Tsandilas [108] implemented an evaluation procedure for V_{rd} that is similar to the one used for between-subjects designs, we find it important to stress this key difference between simulating within-subjects and between-subjects designs.

We evaluate the following tests for two dependent samples:

(1) *The paired-samples t-test.* Since *AR* represents the arithmetic mean of $\hat{a}_i$, the *t*-test can be directly used to compare two means. We used the standard R implementation of the `t.test` function with the `paired` argument set to `TRUE`. Note that the *t*-test assumes that observations are independent, which does not always occur for the *AR* measure when interpreted as a mean, even in the presence of non-transitivity. Therefore, we expect larger Type I errors for this test, but we include it anyway in our simulations to verify such assumptions and to present them in a clear manner to readers.

Table 7. Type I Errors (α = .05) and Power for Between-subjects Designs

	Model, distribution & dissimilarity	*k*	Type I error rate										Power									
			t	**Wt**	**MW**	**Cliff**	**BMP**	**KS**	$V_b^\star$	V_b	**pb**	**bt**	**t**	**Wt**	**MW**	**Cliff**	**BMP**	**KS**	$V_b^\star$	V_b	**pb**	**bt**
S_1.	Expert/Zipf–Mandelbrot/Identity	1	.384	.384	.633	.610	.601	.760	.036	.551	.109	.068	.766	.766	.890	.875	.869	.934	.116	.862	.538	.411
S_2.	Codebook/Bernoulli/Lin	5	.280	.278	.316	.294	.305	.426	.174	.234	.029	.044	.886	.886	.888	.883	.839	.920	.154	.890	.707	.762
S_3.	Codebook/Zipf–Mandelbrot/Lin	5	.287	.286	.316	.308	.322	.303	.151	.306	.031	.054	.862	.862	.871	.868	.846	.859	.170	.881	.663	.700
S_4.	Computer/Normal/Minkowski (p = 0.5)	5	.329	.326	.329	.324	.347	.252	.173	.363	.034	.060	.861	.861	.865	.862	.848	.824	.194	.884	.676	.704
S_5.	Computer/Normal/Minkowski (p = 1)	5	.322	.322	.323	.322	.341	.277	.157	.350	.052	.082	.848	.848	.853	.851	.835	.827	.176	873	.646	.673
S_6.	Computer/Normal/Minkowski (p = 2)	5	.276	.273	.309	.301	.316	.277	.163	.333	.042	.062	.845	.845	.857	.856	.839	.830	.187	.877	.642	.666
S_7.	Computer/Normal/Squared Euclidean	5	.329	.326	.352	.346	.367	.313	.181	.352	.048	.072	.843	.843	.859	.855	.840	.832	.182	.877	.632	.653
S_8.	Codebook/Bernoulli/Lin	10	.232	.232	.228	.220	.237	.210	.149	.167	.008	.043	.904	.904	.903	.901	.867	.881	.168	.893	.720	.799
S_9.	Codebook/Zipf-Mandelbrot/Lin	10	.274	.274	.288	.288	.299	.224	.149	.294	.027	.054	.857	.857	.857	.856	.834	.826	.174	.868	.672	.714
S_{10}.	Computer/Normal/Minkowski (p = 0.5)	10	.300	.298	.282	.281	.307	.200	.168	.334	.034	.059	.863	.863	.864	.861	.844	.829	.186	.887	.669	.698
S_{11}.	Computer/Normal/Minkowski (p = 1)	10	.299	.297	.306	.294	.321	.238	.160	.379	.042	.064	.847	.846	.845	.843	.826	.818	.187	.879	.645	.667
S_{12}.	Computer/Normal/Minkowski (p = 2)	10	.327	.322	.341	.333	.343	.300	.172	.381	.043	.063	.842	.842	.850	.847	.831	.825	.184	.874	.641	.666
S_{13}.	Computer/Normal/Squared Euclidean	10	.309	.309	.330	.324	.340	.298	.161	.352	.042	.070	.853	.852	.861	.856	.842	.828	.185	.879	.642	.661
	Mean		**.304**	**.302**	**.335**	**.327**	**.340**	**.314**	**.153**	**.338**	**.042**	**.061**	**.852**	**.852**	**.866**	**.863**	**.843**	**.849**	**.174**	**.879**	**.653**	**.675**

Note: Test names are abbreviated; refer to Table 6 for full names.

Table 8. Mean Type I Errors and Power for Within-Subjects Designs

	Test	Abbr.	Type I error rate		Power	
			$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
1.	Paired-samples *t*-test	t	.290	.154	.853	.785
2.	Wilcoxon	W	.301	.154	.856	.778
3.	V_{rd}	V_{rd}	.360	.234	.883	.835
4.	V_{rd} with corrected df	$V^{\star}_{rd}$	.063	.038	.685	.634
5.	★**percentile bootstrap**	**pb**	**.049**	**.015**	.672	.551
6.	★**bootstrap-*t***	**bt**	**.065**	**.065**†	.665	.665†

For details, see Table 9. *Note:* stars ★ highlight the best performing statistical tests.
†For $\alpha = .05$ and $\alpha = .01$, Type I error rates and power are identical for the bootstrap-*t* technique, since rejection is based on confidence intervals and no actual *p*-values are computed.

(2) *Wilcoxon's signed-rank test* [134] implemented by the R function `wilcox.test`[37] with the `paired` argument set to `TRUE`.

(3) *The V_{rd} test* proposed by Vatavu and Wobbrock [120].

(4) *The V_{rd} test with corrected degrees of freedom.* Tsandilas [108] found V_{rd} problematic because *"the agreement pairs are highly interdependent"* (p. 18:25). Since the test involves $N \cdot (N-1)/2$ observations and, for each participant, there are N-1 observations from which its normalized agreement score $\hat{a}_i$ is computed (Equation (7)), we corrected the number of degrees of freedom of the χ^2 distribution of V_{rd} from 1 (when comparing two groups) to $N/2$ (when N-1 observations are pseudo-replicated $N/2$ times).

(5) *The percentile boostrap method* described in Wilcox [132, p. 411], which we implemented in R for *AR* using B = 2000 bootstrap samples.[38] Tsandilas [108] preferred the jackknife technique for within-subjects designs because it was *"faster and easier to evaluate through Monte Carlo simulations"* (p. 18:29), but we stick to bootstrapping to be consistent with the between-subjects evaluation.

(6) *The bootstrap-t method* [132] (p. 268), which we implemented in R for dependent *AR*'s using B = 2000 bootstrap samples.

Overall, we report results from 70,200 trials (6 tests × 13 simulation conditions × 9 target *AR* values × 100 repetitions) for Type I errors, and from 280,800 trials (6 tests × 13 conditions × 36 pairs of target *AR* values × 100 repetitions) for statistical power.

Table 8 shows the Type I error rates and power estimations for critical levels of significance .05 and .01, respectively. Again, the percentile bootstrap delivered the best control over the Type I error rate. Our simulation results confirm large Type I errors for V_{rd}, between .188 and .556 (M = .360); see detailed results in Table 9. However, the $V^{\star}_{rd}$ test with corrected degrees of freedom delivered a Type I error rate of just .063 for $\alpha = .05$ and .038 for $\alpha = .01$, better than the bootstrap-*t* method and much better than the other tests. Again, the performance of the other tests can be explained by $\hat{a}_i$ values not being completely independent, even under non-transitivity. (Note that we do not provide any mathematical proof in this article regarding the correctness of the *df* correction employed by $V^{\star}_{rd}$, and we base it on the above intuition that *df* should be $N/2$, as well as on a failed attempt from Tsandilas [108] to correct the V_{rd} test in another way; see [108], pp. 18:25.)

[37] https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/wilcox.test.
[38] Wilcox's [132] implementations use the default value B = 2000.

Table 9. Type I Error (α = .05) and Power for Within-subjects Designs

Model, distribution & dissimilarity	k	Type I error rate						Power					
		t	W	V_{rd}	$V^{\star}_{rd}$	pb	bt	t	W	V_{rd}	$V^{\star}_{rd}$	pb	bt
S_1. Expert/Zipf-Mandelbrot/Identity	1	.354	.529	.556	.247	.127	.077	.777	.858	.876	.701	.587	.411
S_2. Codebook/Bernoulli/Lin	5	.264	.266	.270	.031	.043	.078	.887	.885	.886	.674	.713	.741
S_3. Codebook/Zipf-Mandelbrot/Lin	5	.282	.271	.324	.038	.038	.064	.868	.867	.887	.679	.689	.701
S_4. Computer/Normal/Minkowski (p = 0.5)	5	.277	.268	.334	.043	.044	.064	.854	.847	.881	.682	.671	.676
S_5. Computer/Normal/Minkowski (p = 1)	5	.303	.308	.381	.073	.053	.068	.853	.849	.888	.691	.667	.659
S_6. Computer/Normal/Minkowski (p = 2)	5	.278	.286	.361	.057	.054	.064	.850	.847	.884	.688	.665	.659
S_7. Computer/Normal/Squared Euclidean	5	.328	.331	.410	.054	.046	.062	.839	.838	.873	.690	.666	.659
S_8. Codebook/Bernoulli/Lin	10	.234	.228	.188	.006	.018	.046	.912	.905	.897	.682	.743	.786
S_9. Codebook/Zipf-Mandelbrot/Lin	10	.292	.278	.327	.031	.038	.063	.873	.871	.889	.678	.692	.713
S_{10}. Computer/Normal/Minkowski (p = 0.5)	10	.284	.269	.344	.041	.033	.051	.852	.850	.884	.687	.674	.681
S_{11}. Computer/Normal/Minkowski (p = 1)	10	.297	.300	.412	.062	.046	.068	.836	.833	.871	.677	.651	.647
S_{12}. Computer/Normal/Minkowski (p = 2)	10	.300	.299	.400	.071	.050	.071	.849	.847	.886	.692	.660	.657
S_{13}. Computer/Normal/Squared Euclidean	10	.280	.279	.376	.061	.041	.064	.839	.836	.873	.683	.657	.656
Mean		**.290**	**.301**	**.360**	**.063**	**.049**	**.065**	**.853**	**.856**	**.883**	**.685**	**.672**	**.665**

Note: test names are abbreviated; refer to Table 8 for full names..

9.3 Which Statistical Test to Use?

The available data from 936,000 simulations involving 13 combinations of models of agreement analysis, probability distributions for bias, and dissimilarity functions, lead to the conclusion that the percentile bootstrap controls the best the Type I error rate for both within- and between-subjects designs. At the same time, its power seems to be large enough (65%) to detect differences in agreement rates of at least .10 at the critical level of .05. We also hope that our evaluations clarify concerns from Tsandilas [108] about V_{rd} and V_b. Since $V^{\star}_{b}$, the default implementation of AGATe [121], delivers Type I error rates of 15% and power of just 17%, and while a mathematical proof is awaited for the very good performance delivered by $V^{\star}_{rd}$ (left for future work), *the percentile bootstrap represents our recommendation for practitioners to compare agreement rates for both within-subjects and between-subjects experimental designs.* Based on the results of our extensive evaluation, we can address research questions [$RQ_{4.1}$] and [$RQ_{4.2}$], as follows:

> **Research Question [$RQ_{4.1}$]:** Which statistical test should one use for analyzing agreement data for end-user elicitation studies with between-subjects experimental designs?
>
> **Clarification:** The percentile bootstrap [132] (pp. 332–335) seems to control the Type I error rate very well under a variety of testing conditions.
>
> **Research Question [$RQ_{4.2}$]:** Which statistical test should one use for analyzing agreement data for within-subjects experimental designs?
>
> **Clarification:** The percentile bootstrap [132] (p. 411) seems to control the Type I error rate very well under a variety of testing conditions.

10 DISCUSSION

In this section, we summarize clarifications for the SC expressed in the literature regarding the end-user elicitation method that we outlined in Section 1. We also use our theoretical and empirical results to provide recommendations for researchers and practitioners that wish to apply the end-user elicitation method in their own work.

10.1 Revisiting SC from the Literature Regarding the End-User Elicitation Method

We revisit the Specific Concerns [SC_1] to [SC_7], outlined in Section 1, that were formulated in the literature [75, 82, 103, 108, 114] regarding various practical aspects of conducting user studies with the elicitation method [140, 141] and its variations for calculating and analyzing agreement [32, 120, 121]. Our goal is to respond to these concerns with the support of the theoretical and empirical results presented in this article.

[SC_1] Stern et al. [2008] [103]: Claims that eliciting proposals by having participants actually performing them, as proposed in Wobbrock et al. [140, 141], may be a less suited approach compared to other ways to elicit end users' preferences for actions, commands, or symbols, such as the "coded gesture entry" method.

Clarification: The coded gesture entry method is a specific form of a codebook qualitative study, where participants specify the characteristics of gestures from a predefined list. Consequently, the results can and should be analyzed with the tools of inter-rater reliability [40], since chance agreement is a potential outcome. Contrary to coded gesture entry, the end-user elicitation method [140, 141] works without any predefined lists, categories, or constraints for participants from which proposals are elicited. Depending on the specific application of elicitation studies, there are three possible models for calculating agreement: the *expert*, *codebook*, and *computer* models. We recommend the use of the *computer* model for reasons of efficiency and reproducibility of results; see also our recommendations in the next subsection.

[SC_2] Nebeling et al. [2014] [82]: Claims that the end-user elicitation method should be extended toward reproducible and implementable user-defined interaction sets.

Clarification: We agree with this claim. Our recommendation for the future of end-user elicitation is toward reproducible results, for which we identify a five-level hierarchy of end-user elicitation studies. This hierarchy is discussed in the next subsection and represents one of the contributions of this article.

[SC_3] Morris et al. [2014] [76]: Claims that legacy bias, i.e., the potential pitfall of users' proposals to be biased by their experience with prior interfaces and technologies, is a limitation of the original end-user elicitation method [140, 141].

Clarification: We agree that legacy bias is a potential concern for limiting the effectiveness of the end-user elicitation method to discover new interactions that take full advantage of the possibilities offered by the areas under examination. To the best of our knowledge, Morris et al.'s proposed solutions of *production*, *priming*, and *partners* have received only few formal evaluations in the literature as to whether they reduce legacy bias; e.g., Hoff et al. [43] reported medium effect sizes and large variance between participants' proposals. More recently, Williams et al. [136] proposed a technique to address legacy bias in the form of an evolving set of interactions. Consequently, addressing legacy bias represents a promising avenue for future work. We note, however, that all legacy bias might not be bad. It is conceivable that for certain systems, designers may want to leverage preexisting knowledge or familiarity in order to make their new systems easier to guess or learn. Legacy bias could therefore be an advantage in some cases and, consequently, modeling legacy bias from the perspective of agreement occurring by chance and ignoring the corresponding amount of agreement is not recommended practice for end-user elicitation; see also the specific concern SC_4, next.

[SC_4] Tsandilas [2018] [108]: Claims that the established measures of agreement calculation, *A* and *AR*, advocated by Wobbrock et al. [140, 141], Findlater et al. [32], and Vatavu and Wobbrock [120, 121], do not take into account chance agreement.

Clarification: Since inter-rater reliability and end-user elicitation studies are fundamentally different, the measures of agreement A [140, 141] and AR [32, 120, 121] traditionally used in gesture elicitation studies incorporate both chance agreement and chance disagreement. Moreover, the agreement relation is a tolerance relation in end-user elicitation studies, which means that it is non-transitive. What would happen if one still used the κ and κ_F coefficients from inter-rater reliability for non-transitive agreement relations? Cohen's [21] assumptions of independent and mutually exclusive categories would break because of the conflict generated by non-transitivity. Therefore, κ and κ_F and non-transitive agreement relations are incompatible.

[SC_5] Tsandilas [2018] [108]: Claims that the guidelines proposed by Vatavu and Wobbrock [120] for interpreting the magnitude of agreement can lead to overoptimistic conclusions about the true level of agreement reached by the participants of end-user elicitation studies.

Clarification: We showed that the guidelines from Vatavu and Wobbrock [120] should be used with care to prevent drawing incorrect and misleading conclusions. Especially when it is the same practitioners that set the criteria to judge the similarity of elicited proposals and that employ those criteria to calculate and report agreement. The facts are that (i) different end-user elicitation studies use different criteria to evaluate the similarity of proposals elicited from the study participants, and (ii) the criteria influence the magnitude of the agreement rate. Nevertheless, the margins .10, .30, and .50 can be used as rough guidelines for the qualitative interpretations of agreement rates as *low*, *medium*, and *high* agreement, but not necessarily to draw final conclusions about the outcomes of the study or to compare results between studies. Instead, conclusions should consider the particularities of the specific investigations, application domains, contexts of the studies (e.g., participants, criteria, bias), and agreement analysis approaches.

[SC_6] Tsandilas [2018] [108]: Claims that the V_{rd} and V_b test statistics proposed by Vatavu and Wobbrock [120, 121] yield high Type I error rates.

Clarification: Our simulations for V_b show similar Type I error rates for the expert/Zipf–Mandelbrot tested by Tsandilas [108] and slightly better performance for the other simulation conditions. (This is not surprising, since the Discrete Half Normal distribution, also evaluated [108], showed lower Type I errors for V_b as well.) However, $V_b^\star$ showed much better performance, ranking third across the 10 tests under evaluation. $V_b^\star$, reported by default in AGATe [121], was unfortunately not evaluated by Tsandilas [108]. However, the relatively better control of the Type I error rate of $V_b^\star$ compared to V_b comes with the downside of a very low power (17% and 12%, respectively, for $\alpha = .05$ and .01), representing the lowest power among all the tests that we evaluated for the between-subjects condition). The $V_{rd}^\star$ test with corrected degrees of freedom delivered a Type I error rate of just .063 for critical level .05 and .038 for critical level .01, better than the bootstrap-t method and much better than the other tests that we evaluated. The performance of the other tests, e.g., the t-test, BMP, or KS, can be explained by $\hat{a}_i$ values not being independent, even under non-transitivity. In conclusion, our empirical results recommend the percentile bootstrap technique for analyzing agreement rates in end-user elicitation studies.

[SC_7] Vatavu [114]: Claims that the criteria used to evaluate the similarity of proposals elicited from the participants of end-user elicitation studies can make the magnitude of agreement scores irrelevant, because of the dependency between agreement and the criteria employed. Instead, a holistic approach in which agreement is interpreted as

a function of the criteria used to judge the similarity of elicited proposals should be preferred to using specific, possibly subjective criteria.

Clarification: Our measure AR_ϵ reconciles previous measures of agreement A [140, 141] and AR [32, 120, 121] with the C measure [114], showing that they are profoundly connected. Interesting future work lies ahead regarding structuring the criteria used to judge the similarity of elicited proposals to increase the replicability of the classification step.

Based on these clarifications, we provide answers to the four RQ outlined in Section 1 and show how the specific concerns SC relate to the research questions RQ, as follows:

[RQ_1] How do end-user elicitation studies compare to inter-rater reliability studies?

Answer: We showed that end-user elicitation and inter-rater reliability studies make fundamentally different assumptions. Consequently, adopting the theory and practice of inter-rater reliability (e.g., the concept of chance agreement and coefficients that correct for the influence of chance agreement) may not be valid in the context of end-user elicitation. This clarification addresses specific concerns [SC_1] and [SC_4].

[RQ_2] What is agreement in end-user elicitation?

Answer: We showed that the agreement relation is a mathematical relation of tolerance [145] that generates a tolerance space [102] over the set of distinct proposals elicited from end users. As a mathematical tolerance, *the agreement relation is not necessarily transitive,* a key aspect mistakenly taken for granted in prior work [108, 121], which has direct consequences on some of the measures that calculate agreement. This clarification addresses concerns [SC_2], [SC_4], [SC_5], and [SC_7], respectively.

[RQ_3] Can end-user elicitation be modeled formally?

Answer: Yes. We presented a comprehensive, six-step operational model for conducting *general* end-user elicitation studies in HCI, which include the popular gesture elicitation studies [140, 141]. Our model includes a formalization of the classification step for grouping elicited proposals into signs and highlights the step of characterizing proposals to report traits. In the framework of this formalization of end-user elicitation, we identify three distinct models of agreement analysis: *expert*, *codebook*, and *computer*. This clarification addresses concerns [SC_1], [SC_3], and [SC_4].

[RQ_4] What statistical procedures best apply to elicitation data?

Answer: The *percentile bootstrap* method [132] (pp. 332, 411) seems to be the best-performing statistical test. This finding results from our extensive simulations regarding the Type I error rate and statistical power of 16 statistical tests for within- and between-subjects experimental designs. This clarification addresses concern [SC_6].

10.2 Recommendations for the Science of End-User Elicitation

Our theoretical and empirical examinations revealed that end-user elicitation data has different properties than believed until now, which we unveiled and used to address criticisms, in particular regarding the calculation of agreement [108]. Beyond that, however, how should practitioners make effective use of our findings to understand human behavior? To this point, we outline three recommendations for conducting end-user elicitation studies by adopting a numerical perspective for the data collected in those studies:

(1) *Whenever possible, employ computational acquisition and representation of proposals elicited from end-users.* For example, a gesture elicitation study examining free-hand commands

should capture gestures using an actual acquisition device, such as Leap Motion, Kinect, and so on, besides the experimenter's notes and videos. Direct advantages of capturing participants' proposals in this form are represented by higher fidelity of the data, higher efficiency in evaluating agreement by replacing human effort with numerical computations, and gaining a larger perspective on the relationship between the magnitude of agreement and the dissimilarity presented in the set of elicited proposals [114]. These practical advantages in terms of efficiency and perspective have direct implications on the researcher's time and effort spent to inform their user interface designs and prototypes. However, we acknowledge that technology to record proposals in a computational form might not always be available, e.g., when studying suitable gesture input commands for very new sensing technology, yet to be perfected, such as radar gestures [68], or even yet to be invented, such as always-available mid-air substrates of digital content [96]. Or when technology may not be readily accessible and/or affordable, such as sensing gestures in the terahertz domain [146]. In other cases, recording technology may be an unnecessary overkill for the limited scope of the study (e.g., a preliminary study to inform the design of a full experiment) or it may limit innovations because of the time and effort required for ethical clearances that the use of some technology on human subjects may require, e.g., gesture input for implanted user interfaces [46]. We encourage the community to consider computational recording of elicited proposals whenever possible and useful, while keeping in mind the scope and extent of their investigation.

(2) *Favor numerical tools to identify signs and compute agreement.* These tools include implementations of dissimilarity functions [114], clustering algorithms [3, 6], applications to acquire user input [82], and platforms [4, 69] that facilitate conducting elicitation studies. Even for cases when elicited proposals were not captured in a computational form, automated clustering methods could help practitioners visualize the dendrogram in order to support their decision regarding the most suitable partitioning of the set of proposals into signs.

(3) *Make data and/or analysis scripts publicly available to encourage replication of results.* Besides some prior work [38, 80, 114] and few cases where authors were kind enough to share their data when contacted,[39] the community lacks public elicitation datasets. This is unfortunate, especially considering that more than 200 gesture elicitation studies have been published to date [124].

Note that the dissimilarity function approach that we propose for end-user elicitation data analysis encompasses the new computational methods advocated in this section, but also the conventional method where researchers and practitioners manually label proposals into signs based on a codebook and their judgments about similarity. It is also important to note that both approaches may be affected by the subjectivity of the researcher, a problem discussed in length and exemplified in [114]. For instance, in the case of computational dissimilarity functions, subjectivity can surface in the form of how the raw data are preprocessed, filtered, normalized, or what kind of features are extracted from that data. However, even if subjectivity exists in both approaches, a computational method has the net advantage of being readily replicable: the code implementing the dissimilarity function and method can be run by another researcher to replicate previous results and/or to reproduce those results on another dataset.

These recommendations may be difficult to adopt in the short-term, but they are important in light of the new perspective on classification and non-transitivity offered in this article. To encourage adoption of our recommendations, we propose a classification of end-user elicitation studies by their level of data disclosure, inspired by RepliCHI efforts [138], a previous call to replication

[39]We are especially thankful to Gilles Bailly and Thammathip Piumsomboon.

in gesture elicitation [82], and ACM's [1] recommendations for repeatability, replicability, and reproducibility in experimental science:

Level-0: An elicitation study that does not record data computationally. Unfortunately, the majority of published end-user elicitation studies are of this kind. Not recording data computationally also makes it difficult to share the data.

Level-1: An elicitation study that records data computationally as well as releases the data, eventually with companion analysis scripts and software.

Level-2: A Level-1 study for which the data was validated and the results confirmed by independent researchers/authors.

Level-3a: A Level-2 elicitation study that was successfully reproduced with other participants.

Level-3b: A Level-2 elicitation study that is validated through an end-user identification study, which reverses the elicitation process. Specifically, the signs generated in an elicitation study are shown to a fresh set of participants, who then infer the intended referents, as described by Ali et al. [4].

Level-4: An elicitation study that satisfies both Level-3a and Level-3b. In other words, a study that is replicated as an elicitation study and validated as an identification study.

From Level-0 to Level-4, our confidence in the results reported from end-user elicitation studies increases, as well as our ability to replicate and reproduce those results toward consolidating knowledge in the community. At this point, the vast majority of studies are Level-0, Vatavu [114] is Level-1, and we are not aware of any Level-2 or 3 studies, although Nebeling et al. [82] did report on a replication of Morris [75] and, thus, comes close to a Level-3 elicitation study. Ali et al. [4] ran an identification study using the Crowdlicit platform, coming thus close to a Level-3b end-user elicitation study.

What about Level-0 studies? Are they still useful? We recommend conducting them internally to inform other, more complex studies, to confirm or disconfirm insights and initial hypotheses, or to gain more knowledge about a specific application domain. Thus, such studies are still useful for authors that employ them for the reasons enumerated above as well as for situations where acquisition technology may not be available, perfected, affordable, or reasonable to use on human subjects. In such cases, Level-0 studies, even without data in computational form, can nevertheless inspire innovation and develop knowledge regarding end users' preferences for interactive modalities, devices, and systems. However, Level-2 and Level-3 studies are our recommendation, when possible, in order to avoid replication issues, already remarked in other fields of research, regarding non-reproducible published results [24, 49] as well as to foster more replications in HCI [47].

10.3 Methodological Recommendations for Designers and Practitioners

End-user elicitation studies can be viewed as a form of participatory design [78, 97], where participants' ideas are taken directly as input into the design process, especially for generating new design ideas. Thus, our recommendations serve not only researchers who use the end-user elicitation method, but also design practitioners who seek to create interactive systems representative of users' thoughts and behaviors. From this perspective, we hope that our results will also be useful to designers, developers, and usability specialists, especially newcomers to end-user elicitation, to better understand recent theoretical developments, avoid pitfalls, and correctly apply measures of agreement and statistical tests. To this end, we summarize our theoretical elaboration and empirical findings in the following list of guidelines:

(1) Decide which model of agreement analysis (*expert*, *codebook*, *computer*) applies best to your particular study. To this end, answers to the following questions may be helpful: Are there

experts available? If not, is it sensible to use the crowd for the particular application domain? Can a codebook be clearly defined? Can a dissimilarity function be implemented in the form of a computer program?

(2) For the *expert* model, describe the expertise of the experts. When the crowd is used, describe the demographics of the crowd workers (e.g., gender, age, education, professional background), the recruitment procedure, and any validation checks performed on the similarity votes.

(3) For *codebook* models, clearly describe the codebook: the criteria (dimensions) and the categories of each criterion. If rules are used instead, clearly state all the rules employed for the analysis. Despite the common-sense of these recommendations, most published studies fail to describe how agreement was evaluated. In the absence of such information, replication of results may be difficult or even impossible. A key observation is that *agreement rates are irrelevant if they are not reliable, and to be reliable, they must be reproducible.* To this end, the codebook or set of rules must be clearly presented in the study description.

(4) For *computer* models, connect the outcomes of the end-user elicitation study with the application meant for end users. If the application is a new gesture or voice input recognizer, use the same dissimilarity function for clustering (i.e., agreement calculation) and discrimination (i.e., evaluation of the recognizer). Make it a priority to release source code and/or scripts that can be used by other researchers to reproduce the results.

(5) Once the data are collected, check if the agreement relation is transitive (likely true if the analysis model is *codebook* and the codebook is finite, but likely untrue for the other models). If the codebook is finite and there are few categories, consider whether chance agreement [108] may represent an issue for your study. Chance agreement can be an issue if the number of categories is small and the experimenter must select a category for each proposal, even when they are unsure about what to select [21, 40].

(6) Until future work further examines the $V_b^\star$ test and the corrected degrees of freedom for the $V_{rd}^\star$ test, we recommend using the percentile bootstrap method for statistical inferences about agreement rates.

(7) As a last guideline, we recommend the community to focus on applied results. Although unveiling users' preferences and behavior by conducting elicitation studies is valuable, too many published articles stop at reporting a consensus set of proposals, with no subsequent work developing an actual user interface, interaction technique, application, or interactive system. This approach fosters new publications, but hinders advances in applied results, which was the original goal of the method—to improve interface guessability [140].

11 RESOURCES

We provide companion code in R implementing measures of agreement, statistical tests for within-subjects and between-subjects experimental designs, and our simulation procedures for the *expert*, *codebook*, and *computer* models. We equally release our simulation data files with the note that we employed fixed seeds for random number generators in all our simulation experiments, so that our results from Tables 3 to 9 can be readily reproduced. Code resources are freely available for download from https://depts.washington.edu/acelab/proj/dollar/agate.html.

12 CONCLUSION

In this work, we re-examined fundamental issues in end-user elicitation studies, specifically concerning agreement procedures and calculations, and the objections articulately, but incompletely, raised by Tsandilas [108]. Our most important results are a model for general end-user elicitation

in HCI, experimental proofs for the non-transitivity of agreement, clarifications regarding chance agreement and the differences between inter-rater reliability and elicitation studies. Besides these contributions, we also provide support for a mathematical formalization based on dissimilarity functions and tolerance spaces to describe agreement formation for end-user elicitation. Our results suggest the need for revisiting current statistical tests for agreement data analysis [108, 120, 121], since these tests have either assumed transitivity, such as V_{rd} and V_b [120, 121], or used κ coefficients [108]. We leave this aspect for future work. Other connected aspects, such as the best ways to perform clustering to identify the consensus set of signs, are left for future work, due to constraints on space. More importantly, our results have implications on *our understanding of how agreement is formed and how it should be evaluated computationally*. It is our hope that this work will lead to more confident, formalized, rigorous, reliable, and reproducible end-user elicitation studies.

REFERENCES

[1] ACM. 2020. Artifact Review and Badging. Retrieved on October 2021 from https://www.acm.org/publications/policies/artifact-review-badging.

[2] Abdullah Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2021. "I am iron man": Priming improves the learnability and memorability of user-elicited gestures. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3411764.3445758

[3] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2018. Crowdsourcing similarity judgments for agreement analysis in end-user elicitation studies. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 177–188. DOI: https://doi.org/10.1145/3242587.3242621

[4] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2019. Crowdlicit: A system for conducting distributed end-user elicitation and identification studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3290605.3300485

[5] Douglas G. Altman. 1991. *Practical Statistics for Medical Research*. Chapman & Hall/CRC, Boca Raton, FL.

[6] Lisa Anthony, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2013. Understanding the consistency of users' pen and finger stroke gesture articulation. In *Proceedings of the Graphics Interface 2013*. Canadian Information Processing Society, Toronto, Ontario, 87–94. Retrieved from http://dl.acm.org/citation.cfm?id=2532129.2532145.

[7] Lisa Anthony and Jacob O. Wobbrock. 2012. $N-Protractor: A fast and accurate multistroke recognizer. In *Proceedings of the Graphics Interface 2012*. Canadian Information Processing Society, Toronto, Ontario, 117–120. Retrieved from http://dl.acm.org/citation.cfm?id=2305276.2305296.

[8] Gilles Bailly, Eric Lecolinet, and Yves Guiard. 2010. Finger-count & radial-stroke shortcuts: 2 techniques for augmenting linear menus on multi-touch surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 591–594. DOI: https://doi.org/10.1145/1753326.1753414

[9] 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*. ACM, New York, NY, 563–572. DOI: https://doi.org/10.1145/2470654.2470734

[10] Roger Bakeman. 2018. *KappaAcc: Deciding Whether Kappa is Big Enough by Estimating Observer Accuracy*. Technical Report No. 28. Georgia State University, Atlanta, GA. Retrieved from http://bakeman.gsucreate.org/DevLabTechReport28.pdf.

[11] Roger Bakeman and Vicenç Quera. 2011. *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press, New York, NY.

[12] Roger Bakeman, Vicenç Quera, and Augusto Gnisci. 2009. Observer agreement for timed-event sequential data: A comparison of time-based and event-based algorithms. *Behavior Research Methods* 41, 1 (2009), 137–147. DOI: https://doi.org/10.37582FBRM.41.1.137

[13] Bradley Boehmke and Brandon Greenwell. 2020. *Hands-On Machine Learning with R*. CRC Press, Taylor & Francis Group, Boca Raton, FL. Retrieved from https://www.crcpress.com/Hands-On-Machine-Learning-with-R/Boehmke-Greenwell/p/book/9781138495685.

[14] Shyam Boriah, Varun Chandola, and Vipin Kumar. 2008. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 8th SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 243–254. DOI: https://doi.org/10.1137/1.9781611972788.22

[15] Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41, 3 (1981), 687–699. DOI: https://doi.org/10.1177/001316448104100307

[16] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Amaury Trujillo. 2017. Analyzing visually impaired people's touch gestures on smartphones. *Multimedia Tools and Applications* 76, 4 (2017), 5141–5169. DOI: https://doi.org/10.1007/s11042-016-3594-9

[17] Mingyu Chen, Ghassan AlRegib, and Biing-Hwang Juang. 2012. 6DMG: A new 6D motion gesture database. In *Proceedings of the 3rd Multimedia Systems Conference*. ACM, New York, NY, 83–88. DOI: https://doi.org/10.1145/2155555.2155569

[18] Yu-Chun Chen, Chia-Ying Liao, Shuo-wen Hsu, Da-Yuan Huang, and Bing-Yu Chen. 2020. Exploring user defined gestures for ear-based interactions. *Proceedings of the ACM on Human–Computer Interaction* 4, ISS, Article 186 (Nov. 2020), 20 pages. DOI: https://doi.org/10.1145/3427314

[19] 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*. ACM, New York, NY, 1559–1568. DOI: https://doi.org/10.1145/2470654.2466207

[20] W. G. Cochran. 1950. The comparison of percentages in matched samples. *Biometrika* 37, 3/4 (1950), 256–266. DOI: https://doi.org/10.2307/2332378

[21] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. DOI: https://doi.org/10.1177/001316446002000104

[22] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (1968), 213–220. DOI: https://doi.org/10.1037/h0026256

[23] J. Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159. DOI: https://doi.org/10.1037/0033-2909.112.1.155

[24] David Colquhoun. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1, 3 (2014), 140216. DOI: https://doi.org/10.1098/rsos.140216

[25] Christoph Dalitz. 2009. Reject options and confidence measures for kNN classifiers. In *Document Image Analysis with the Gamera Framework*. Christoph Dalitz (Ed.), Shaker Verlag, Aachen, 16–38. Retrieved from https://lionel.kr.hs-niederrhein.de/~dalitz/data/publications/sr09-knn-rejection.pdf.

[26] Tilman Dingler, Rufat Rzayev, Alireza Sahami Shirazi, and Niels Henze. 2018. Designing consistent gestures across device types: Eliciting RSVP controls for phone, watch, and glasses. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3173574.3173993

[27] Guiying Du, Auriol Degbelo, and Christian Kray. 2019. User-generated gestures for voting and commenting on immersive displays in urban planning. *Multimodal Technologies and Interaction* 3, 2 (2019), 31. DOI: https://doi.org/10.3390/mti3020031

[28] Jane L. E., Ilene L. E., James A. Landay, and Jessica R. Cauchard. 2017. Drone & Wo: Cultural influences on human-drone interaction techniques. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 6794–6799. DOI: https://doi.org/10.1145/3025453.3025755

[29] David Ellerman. 2013. An introduction to logical entropy and its relation to Shannon entropy. *International Journal of Semantic Computing* 7, 2 (2013), 121–145. DOI: https://doi.org/10.1142/S1793351X13400059

[30] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (June 2006), 861–874. DOI: https://doi.org/10.1016/j.patrec.2005.10.010

[31] Yasmin Felberbaum and Joel Lanir. 2018. Better understanding of foot gestures: An elicitation study. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3173574.3173908

[32] Leah Findlater, Ben Lee, and Jacob O. Wobbrock. 2012. Beyond QWERTY: Augmenting touch screen keyboards with multi-touch gestures for non-alphanumeric input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2679–2682. DOI: https://doi.org/10.1145/2207676.2208660

[33] Ronald Fisher. 1954. *Statistical Methods for Research Workers*. Oliver and Boyd, London.

[34] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. DOI: https://doi.org/10.1037/h0031619

[35] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1737–1746. DOI: https://doi.org/10.1145/2207676.2208303

[36] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for smart rings: Empirical results, insights, and design implications. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, New York, NY, 623–635. DOI: https://doi.org/10.1145/3196709.3196741

[37] Joan Greenbaum and Morten Kyng (Eds.). 1992. *Design at Work: Cooperative Design of Computer Systems*. L. Erlbaum Associates Inc.

[38] Daniela Grijincu, Miguel A. Nacenta, and Per Ola Kristensson. 2014. User-defined interface gestures: Dataset and analysis. In *Proceedings of the 9th ACM International Conference on Interactive Tabletops and Surfaces*. ACM, New York, NY, 25–34. DOI: https://doi.org/10.1145/2669485.2669511

[39] Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* 61, 1 (2008), 29–48. DOI: https://doi.org/10.1348/000711006X126600
[40] Kilem L. Gwet. 2014. *Handbook of Inter-Rater Realiability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (4th ed.). Advanced Analytics, LLC, Gaithersburg, MD.
[41] Kim Halskov and Nicolai Brodersen Hansen. 2015. The diversity of participatory design research practice at PDC 2002-2012. *International Journal of Human-Computer Studies* 74, C (Feb. 2015), 81–92. DOI: https://doi.org/10.1016/j.ijhcs.2014.09.003
[42] Greg Hamerly and Charles Elkan. 2003. Learning the *k* in *k*-means. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, 281–288. Retrieved from https://dl.acm.org/doi/10.5555/2981345.2981381.
[43] Lynn Hoff, Eva Hornecker, and Sven Bertel. 2016. Modifying gesture elicitation: Do kinaesthetic priming and increased production reduce legacy bias? In *Proceedings of the 10th International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, New York, NY, 86–91. DOI: https://doi.org/10.1145/2839462.2839472
[44] Michael Hoffman, Paul Varcholik, and Joseph J. LaViola. 2010. Breaking the status quo: Improving 3D gesture recognition with spatially convenient input devices. In *Proceedings of the 2010 IEEE Virtual Reality Conference*. IEEE Computer Society, Washington, DC, 59–66. DOI: https://doi.org/10.1109/VR.2010.5444813
[45] J. W. Holley and J. P. Guilford. 1964. A note on the G index of agreement. *Educational and Psychological Measurement* 24, 4 (1964), 749–753. DOI: https://doi.org/10.1177/001316446402400402
[46] Christian Holz, Tovi Grossman, George Fitzmaurice, and Anne Agur. 2012. Implanted user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 503–512. DOI: https://doi.org/10.1145/2207676.2207745
[47] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3523–3532. DOI: https://doi.org/10.1145/2556288.2557004
[48] Heloise Hwawen Hse and A. Richard Newton. 2005. Recognition and beautification of multi-stroke symbols in digital ink. *Computers and Graphics*. 29, 4 (Aug. 2005), 533–546. DOI: https://doi.org/10.1016/j.cag.2005.05.006
[49] John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine* 2, 8 (2005), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124
[50] R. C. James. 1992. *The Mathematics Dictionary* (5th ed.). Chapman & Hall, New York, NY.
[51] Xu Jia, Kun-Pyo Lee, and Hyeon-Jeong Suk. 2011. Considerations of applying surface-based phone gestures to natural context. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. ACM, New York, NY, 545–546. DOI: https://doi.org/10.1145/2030112.2030205
[52] Robert Gilmore Pontius Jr and Marco Millones. 2011. Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing* 32, 15 (2011), 4407–4429. DOI: https://doi.org/10.1080/01431161.2011.552923
[53] 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*. ACM, New York, NY, 413–422. DOI: https://doi.org/10.1145/1978942.1979001
[54] Eamonn Keogh. 2002. Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Data Bases*. VLDB Endowment, 406–417. Retrieved from http://dl.acm.org/citation.cfm?id=1287369.1287405.
[55] A. N. Kolmogorov. 1933. Sulla determinazione empirica di una legge di distributione. *Giornale dell'Istituto Italiano degli Attuari* 4 (1933), 83–91.
[56] Anne Köpsel and Nikola Bubalo. 2015. Benefiting from legacy bias. *Interactions* 22, 5 (Aug. 2015), 44–47. DOI: https://doi.org/10.1145/2803169
[57] Panayiotis Koutsabasis and Panagiotis Vogiatzidakis. 2019. Empirical research in mid-air interaction: A systematic review. *International Journal of Human-Computer Interaction* 35, 18 (2019), 1747–1768. DOI: https://doi.org/10.1080/10447318.2019.1572352
[58] Klaus Krippendorff. 1970. Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70. DOI: https://doi.org/10.1177/001316447003000105
[59] Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA.
[60] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (Mar. 1977), 159–174. DOI: https://doi.org/10.2307/2529310
[61] Huy Viet Le, Sven Mayer, Maximilian Weiß, Jonas Vogelsang, Henrike Weingärtner, and Niels Henze. 2020. Shortcut gestures for mobile text editing on fully touch sensitive smartphones. *ACM Transactions on Computer-Human Interaction* 27, 5, Article 33 (Aug. 2020), 38 pages. DOI: https://doi.org/10.1145/3396233

[62] DoYoung Lee, Youryang Lee, Yonghwan Shin, and Ian Oakley. 2018. Designing socially acceptable hand-to-face input. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 711–723. DOI: https://doi.org/10.1145/3242587.3242642

[63] Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady* 10, 8 (1966), 707–710. Retrieved from https://mathscinet.ams.org/mathscinet-getitem?mr=0189928.

[64] Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 296–304. Retrieved from http://dl.acm.org/citation.cfm?id=645527.657297.

[65] Jiayang Liu, Zhen Wang, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. In *Proceedings of the 2009 IEEE International Conference on Pervasive Computing and Communications*. IEEE Computer Society, Washington, D.C., 1–9. DOI: https://doi.org/10.1109/PERCOM.2009.4912759

[66] Naveen Madapana, Glebys Gonzalez, Rahul Taneja, Richard Rodgers, Lingsong Zhang, and Juan Wachs. 2019. Preference elicitation: Obtaining gestural guidelines for PACS in neurosurgery. *International Journal of Medical Informatics* 130 (2019), 103934. DOI: https://doi.org/10.1016/j.ijmedinf.2019.07.013

[67] Naveen Madapana, Glebys Gonzalez, and Juan Wachs. 2020. Gesture agreement assessment using description vectors. In *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG'20)*. IEEE Computer Society, Los Alamitos, CA, USA, 40–44. DOI: http://dx.doi.org/10.1109/FG47880.2020.00043

[68] Nathan Magrofuoco, Jorge-Luis Pérez-Medina, Paolo Roselli, Jean Vanderdonckt, and Santiago Villarreal. 2019. Eliciting contact-based and contactless gestures with radar-based sensors. *IEEE Access* 7, 1 (2019), 176982–176997. DOI: https://doi.org/10.1109/ACCESS.2019.2951349

[69] Nathan Magrofuoco and Jean Vanderdonckt. 2019. Gelicit: A Cloud platform for distributed gesture elicitation studies. *Proceedings of the ACM on Human-Computer Interaction* 3, EICS, Article 6 (June 2019), 41 pages. DOI: https://doi.org/10.1145/3331148

[70] Meethu Malu, Pramod Chundury, and Leah Findlater. 2018. *Exploring Accessible Smartwatch Interactions for People with Upper Body Motor Impairments*. ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3173574.3174062

[71] H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 1 (1947), 50–60. Retrieved from https://www.jstor.org/stable/2236101.

[72] Fabrice Matulic, Brian Vogel, Naoki Kimura, and Daniel Vogel. 2019. Eliciting pen-holding postures for general input with suitability for EMG armband detection. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. ACM, New York, NY, 89–100. DOI: https://doi.org/10.1145/3343055.3359720

[73] Dan Mauney, Jonathan Howarth, Andrew Wirtanen, and Miranda Capra. 2010. Cultural similarities and differences in user-defined gestures for touchscreen user interfaces. In *Proceedings of the CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 4015–4020. DOI: https://doi.org/10.1145/1753846.1754095

[74] Keenan R. May, Thomas M. Gable, and Bruce N. Walker. 2017. Designing an in-vehicle air gesture set using elicitation methods. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, New York, NY, 74–83. DOI: https://doi.org/10.1145/3122986.3123015

[75] 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*. ACM, New York, NY, 95–104. DOI: https://doi.org/10.1145/2396636.2396651

[76] 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. DOI: https://doi.org/10.1145/2591689

[77] Meredith Ringel Morris, Jacob O. Wobbrock, and Andrew D. Wilson. 2010. Understanding users' preferences for surface gestures. In *Proceedings of the Graphics Interface 2010*. Canadian Information Processing Society, 261–268. Retrieved from https://dl.acm.org/doi/10.5555/1839214.1839260.

[78] Michael J. Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM* 36, 6 (June 1993), 24–28. DOI: https://doi.org/10.1145/153571.255960

[79] C. S. Myers and L. R. Rabiner. 1981. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical* 60, 7 (1981), 1389–1409. DOI: https://doi.org/10.1002/j.1538-7305.1981.tb00272.x

[80] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1099–1108. DOI: https://doi.org/10.1145/2470654.2466142

[81] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* 33, 1 (March 2001), 31–88. DOI: https://doi.org/10.1145/375360.375365

[82] Michael Nebeling, Alexander Huber, David Ott, and Moira C. Norrie. 2014. Web on the wall reloaded: Implementation, replication and refinement of user-defined interaction sets. In *Proceedings of the 9th ACM International Conference on Interactive Tabletops and Surfaces*. ACM, New York, NY, 15–24. DOI: https://doi.org/10.1145/2669485.2669497

[83] Uran Oh and Leah Findlater. 2013. The challenges and potential of end-user gesture customization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1129–1138. DOI: https://doi.org/10.1145/2470654.2466145

[84] G. P. Patil and C. Taillie. 1982. Diversity as a concept and itsmeasurement. *Journal of the American Statistical Association* 77, 379 (1982), 548–561. DOI: https://doi.org/10.1080/01621459.1982.10477845

[85] James F. Peters and Piotr Wasilewski. 2012. Tolerance spaces: Origins, theoretical aspects and applications. *Information Sciences* 195 (July 2012), 211–225. DOI: https://doi.org/10.1016/j.ins.2012.01.023

[86] Tran Pham, Jo Vermeulen, Anthony Tang, and Lindsay MacDonald Vermeulen. 2018. Scale impacts elicited gestures for manipulating holograms: Implications for AR gesture design. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, New York, NY, 227–240. DOI: https://doi.org/10.1145/3196709.3196719

[87] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-defined gestures for augmented reality. In *Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 955–960. DOI: https://doi.org/10.1145/2468356.2468527

[88] Henri Poincaré. 1895. L'espace et la géométrie. *Revue de la métaphyisique et de la morale* 3 (1895), 631–646. Retrieved July 4, 2019 from http://ekladata.com/n4Usy_CBSC1B051W5DX5uYxf8eM/Henri-Poincare-L-Espace-et-La-Geometrie.pdf.

[89] Henri Poincaré. 1905. *Science and Hypothesis*. The Walter Scott Publishing Co. Ltd., New York, NY. Retrieved from http://www.gutenberg.org/ebooks/37157.

[90] Roman Rädle, Hans-Christian Jetter, Mario Schreiner, Zhihao Lu, Harald Reiterer, and Yvonne Rogers. 2015. Spatially-aware or spatially-agnostic? Elicitation and evaluation of user-defined cross-device interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3913–3922. DOI: https://doi.org/10.1145/2702123.2702287

[91] Isabel Benavente Rodriguez and Nicolai Marquardt. 2017. Gesture elicitation study on how to opt-in & opt-out from interactions with public displays. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. ACM, New York, NY, 32–41. DOI: https://doi.org/10.1145/3132272.3134118

[92] Dean Rubine. 1991. Specifying gestures by example. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, New York, NY, 329–337. DOI: https://doi.org/10.1145/122718.122753

[93] 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*. ACM, New York, NY, 197–206. DOI: https://doi.org/10.1145/1978942.1978971

[94] Jaime Ruiz and Daniel Vogel. 2015. Soft-constraints to reduce legacy and performance bias to elicit whole-body gestures with low arm fatigue. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3347–3350. DOI: https://doi.org/10.1145/2702123.2702583

[95] Satu Elisa Schaeffer. 2007. Graph clustering. *Computer Science Review* 1, 1 (2007), 27–64. DOI: https://doi.org/10.1016/j.cosrev.2007.05.001

[96] Ovidiu-Andrei Schipor and Radu-Daniel Vatavu. 2018. Invisible, inaudible, and impalpable: Users' preferences and memory performance for digital content in thin air. *IEEE Pervasive Computing* 17, 4 (2018), 76–85. DOI: https://doi.org/10.1109/MPRV.2018.2873856

[97] Douglas Schuler and Aki Namioka (Eds.). 1993. *Participatory Design: Principles and Practices*. L. Erlbaum Associates Inc., Hillsdale, NJ.

[98] William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19, 3 (1955), 321–325. DOI: https://doi.org/10.1086/266577

[99] Mícheál O. Searcóid. 2007. *Metric Spaces*. Springer-Verlag, London. DOI: https://doi.org/10.1007/978-1-84628-627-8

[100] H. Smirnov. 1939. Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique (Matematiceskii Sbornik)* 6 (1939), 3–26.

[101] Nikita Soni, Schuyler Gleaves, Hannah Neff, Sarah Morrison-Smith, Shaghayegh Esmaeili, Ian Mayne, Sayli Bapat, Carrie Schuman, Kathryn A. Stofer, and Lisa Anthony. 2019. Do user-defined gestures for flatscreens generalize to interactive spherical displays for adults and children? In *Proceedings of the 8th ACM International Symposium on Pervasive Displays*. ACM, New York, NY, Article 24, 7 pages. DOI: https://doi.org/10.1145/3321335.3324941

[102] A. B. Sossinsky. 1986. Tolerance space theory and some applications. *Acta Applicandae Mathematica* 5, 2 (Feb 1986), 137–167. DOI: https://doi.org/10.1007/BF00046585

[103] H. I. Stern, J. P. Wachs, and Y. Edan. 2008. Optimal consensus intuitive hand gesture vocabulary design. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*. IEEE, Washington, D.C., 96–103. DOI: https://doi.org/10.1109/ICSC.2008.29

[104] A. Strauss and J. Corbin. 1994. Grounded theory methodology: An overview. In *Handbook of Qualitative Research*. N. K. Denzin and Y. S. Lincoln (Eds.), Sage Publications, Inc., Thousand Oaks, CA, 273–285. DOI: https://psycnet.apa.org/record/1994-98625-016.

[105] Yanke Tan, Sang Ho Yoon, and Karthik Ramani. 2017. BikeGesture: User elicitation and performance of micro hand gesture as input for cycling. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 2147–2154. DOI: https://doi.org/10.1145/3027063.3053075

[106] Eugene M. Taranta II, Amirreza Samiei, Mehran Maghoumi, Pooya Khaloo, Corey R. Pittman, and Joseph J. LaViola Jr. 2017. Jackknife: A reliable recognizer with few samples and many modalities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5850–5861. DOI: https://doi.org/10.1145/3025453.3026002

[107] Giovanni Maria Troiano, Esben Warming Pedersen, and Kasper Hornbæk. 2014. User-defined gestures for elastic, deformable displays. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/2598153.2598184

[108] Theophanis Tsandilas. 2018. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Transactions on Computer-Human Interaction* 25, 3, Article 18 (June 2018), 49 pages. DOI: https://doi.org/10.1145/3182168

[109] John S. Uebersax. 2019. The Myth of Chance-Corrected Agreement. Retrieved July 11, 2019 from https://www.john-uebersax.com/stat/kappa2.htm.

[110] Jean Vanderdonckt, Nathan Magrofuoco, Suzanne Kieffer, Jorge Pérez, Ysabelle Rase, Paolo Roselli, and Santiago Villarreal. 2019. Head and shoulders gestures: Exploring user-defined gestures with upper body. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments*. Aaron Marcus and Wentao Wang (Eds.), Springer International Publishing, Cham, 192–213. DOI: https://doi.org/10.1007/978-3-030-23541-3_15

[111] Radu-Daniel Vatavu. 2012. User-defined gestures for free-hand TV control. In *Proceedings of the 10th European Conference on Interactive TV and Video*. ACM, New York, NY, 45–48. DOI: https://doi.org/10.1145/2325616.2325626

[112] 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. DOI: https://doi.org/10.3233/AIS-130200

[113] Radu-Daniel Vatavu. 2017. Smart-pockets: Body-deictic gestures for fast access to personal data during ambient interactions. *International Journal of Human-Computer Studies* 100, 103 (2017), 1–21. DOI: https://doi.org/10.1016/j.ijhcs.2017.01.005

[114] Radu-Daniel Vatavu. 2019. The Dissimilarity-Consensus approach to agreement analysis in gesture elicitation studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3290605.3300454

[115] Radu-Daniel Vatavu. 2020. Quantifying the consistency of gesture articulation for users with low vision with the Dissimilarity-Consensus method. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, New York, NY. DOI: https://doi.org/10.1145/3406324.3410709

[116] Radu-Daniel Vatavu, Lisa Anthony, and Jacob O. Wobbrock. 2012. Gestures as point clouds: A $P recognizer for user interface prototypes. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 273–280. DOI: https://doi.org/10.1145/2388676.2388732

[117] Radu-Daniel Vatavu, Bogdan-Florin Gheran, and Maria Doina Schipor. 2018. The impact of low vision on touch-gesture articulation on mobile devices. *IEEE Pervasive Computing* 17, 1 (Jan. 2018), 27–37. DOI: https://doi.org/10.1109/MPRV.2018.011591059

[118] Radu-Daniel Vatavu and Ovidiu-Ciprian Ungurean. 2019. Stroke-Gesture input for people with motor impairments: Empirical results & research roadmap. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY. DOI: https://doi.org/10.1145/3290605.3300445

[119] Radu-Daniel Vatavu, Daniel Vogel, Géry Casiez, and Laurent Grisoni. 2011. Estimating the perceived difficulty of pen gestures. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part II*. Springer-Verlag, Berlin, 89–106. Retrieved from http://dl.acm.org/citation.cfm?id=2042118.2042130.

[120] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1325–1334. DOI: https://doi.org/10.1145/2702123.2702223

[121] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2016. Between-subjects elicitation studies: Formalization and tool support. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3390–3402. DOI: https://doi.org/10.1145/2858036.2858228

[122] Radu-Daniel Vatavu and Ionuț-Alexandru Zaiți. 2013. Automatic recognition of object size and shape via user-dependent measurements of the grasping hand. *International Journal of Human-Computer Studies* 71, 5 (2013), 590–607. DOI: https://doi.org/10.1016/j.ijhcs.2013.01.002

[123] Radu-Daniel Vatavu and Ionuţ-Alexandru Zaiţi. 2014. Leap gestures for TV: Insights from an elicitation study. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, New York, NY, 131–138. DOI: https://doi.org/10.1145/2602299.2602316

[124] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. A systematic review of gesture elicitation studies: What can we learn from 216 studies? In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, New York, NY, 855–872. DOI: https://doi.org/10.1145/3357236.3395511

[125] John Vines, Rachel Clarke, Tuck Leong, John McCarthy, Ole Sejer Iversen, Peter Wright, and Patrick Olivier. 2012. Invited SIG - Participation and HCI: Why involve people in design? In *Proceedings of the CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 1217–1220. DOI: https://doi.org/10.1145/2212776.2212427

[126] John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. 2013. Configuring participation: On how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 429–438. DOI: https://doi.org/10.1145/2470654.2470716

[127] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2018. Gesture elicitation studies for mid-air interaction: A review. *Multimodal Technologies and Interaction* 2, 4 (2018), 1–21. DOI: https://doi.org/10.3390/mti2040065

[128] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2019. Frame-based elicitation of mid-air gestures for a smart home device ecosystem. *Informatics* 6, 2 (2019), 23 pages. DOI: https://doi.org/10.3390/informatics6020023

[129] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2020. Mid-air gesture control of multiple home devices in spatial augmented reality prototype. *Multimodal Technologies and Interaction* 4, 3 (2020), 61. DOI: https://doi.org/10.3390/mti4030061

[130] Andrew Webb. 2002. *Statistical Pattern Recognition* (2nd. ed.). John Wiley & Sons Ltd., West Sussex. DOI: https://doi.org/10.1002/0470854774

[131] Ernesto Henrico Weber. 1834. *De pulsu, resorptione, auditu et tactu. Annotationes anatomicae et physiologicae*. Koehler, Leipzig. DOI: https://doi.org/10.3931/e-rara-70261

[132] Rand Wilcox. 2012. *Modern Statistics for the Social and Behavioral Sciences. A Practical Introduction*. CRC Press, Boca Raton, FL.

[133] Rand Wilcox. 2019. Rallfun-v37. Retrieved on October 2021 from https://dornsife.usc.edu/labs/rwilcox/software/.

[134] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. DOI: https://doi.org/10.2307/3001968

[135] Don Willems, Ralph Niels, Marcel van Gerven, and Louis Vuurpijl. 2009. Iconic and multi-stroke gesture recognition. *Pattern Recogn.* 42, 12 (Dec. 2009), 3303–3312. DOI: https://doi.org/10.1016/j.patcog.2009.01.030

[136] Adam S. Williams and Francisco R. Ortega. 2020. Evolutionary gestures: When a gesture is not quite legacy biased. *Interactions* 27, 5 (Sept. 2020), 50–53. DOI: https://doi.org/10.1145/3412499

[137] Andrew D. Wilson. 2007. Sensor- and recognition-based input for interaction. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (2nd ed.). Andrew Sears and Julie A. Jacko (Eds.), L. Erlbaum Assoc. Inc., Hillsdale, NJ, 177–199. DOI: https://doi.org/10.1201/9781410615862

[138] Max L. Wilson, Ed H. Chi, Stuart Reeves, and David Coyle. 2014. RepliCHI: The workshop II. In *Proceedings of the CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, NY, 33–36. DOI: https://doi.org/10.1145/2559206.2559233

[139] Markus L. Wittorf and Mikkel R. Jakobsen. 2016. Eliciting mid-air gestures for wall-display interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, New York, NY. DOI: https://doi.org/10.1145/2971485.2971503

[140] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the guessability of symbolic input. In *Proceedings of the CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 1869–1872. DOI: https://doi.org/10.1145/1056808.1057043

[141] 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*. ACM, New York, NY, 1083–1092. DOI: https://doi.org/10.1145/1518701.1518866

[142] Jacob O. Wobbrock, Brad A. Myers, and John A. Kembel. 2003. EdgeWrite: A stylus-based text entry method designed for high accuracy and stability of motion. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 61–70. DOI: https://doi.org/10.1145/964696.964703

[143] Jacob O. Wobbrock, Andrew D. Wilson, and Yang Li. 2007. Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 159–168. DOI: https://doi.org/10.1145/1294211.1294238

[144] Yukang Yan, Chun Yu, Xin Yi, and Yuanchun Shi. 2018. HeadGesture: Hands-free input approach leveraging head movements for HMD devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4, Article 198 (Dec. 2018), 23 pages. DOI: https://doi.org/10.1145/3287076

[145] E. C. Zeeman. 1962. The topology of the brain and visual perception. In *Topology of 3-Manifolds and Related Topics*. Jr. M. K. Fort (Ed.), Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[146] Zhi Zhou, Zongjie Cao, and Yiming Pi. 2018. Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures. *Sensors* 18, 1 (2018), 10. DOI: https://doi.org/10.3390/s18010010

Received December 2019; revised July 2021; accepted July 2021