RepliGES and GEStory: Visual Tools for Systematizing and Consolidating Knowledge on User-Defined Gestures

Bogdan-Florin Gheran
MintViz Lab, MANSiD Research Center
Ştefan cel Mare University of Suceava
Suceava, Romania
bogdan.gheran@usm.ro

Santiago Villarreal-Narvaez
LouRIM
Université catholique de Louvain
Louvain-la-Neuve, Belgium
santiago.villarreal@uclouvain.be

Radu-Daniel Vatavu
MintViz Lab, MANSiD Research Center
Ştefan cel Mare University of Suceava
Suceava, Romania
radu.vatavu@usm.ro

Jean Vanderdonckt
LouRIM
Université catholique de Louvain
Louvain-la-Neuve, Belgium
jean.vanderdonckt@uclouvain.be

ABSTRACT

The body of knowledge accumulated by gesture elicitation studies (GES), although useful, large, and extensive, is also heterogeneous, scattered in the scientific literature across different venues and fields of research, and difficult to generalize to other contexts of use represented by different gesture types, sensing devices, applications, and user categories. To address such aspects, we introduce RepliGES, a conceptual space that supports (1) *replications* of gesture elicitation studies to confirm, extend, and complete previous findings, (2) *reuse* of previously elicited gesture sets to enable new discoveries, and (3) *extension* and *generalization* of previous findings with new methods of analysis and for new user populations towards consolidated knowledge of user-defined gestures. Based on RepliGES, we introduce GEStory, an interactive design space and visual tool, to structure, visualize and identify user-defined gestures from a number of 216 published gesture elicitation studies.

CCS CONCEPTS

• **Human-centered computing → Gestural input**; *User interface design*; *Participatory design*; *Empirical studies in interaction design*.

KEYWORDS

Gesture elicitation studies, replicability, reproducibility, generalization, repurposing, visual tools

ACM Reference Format:
Bogdan-Florin Gheran, Santiago Villarreal-Narvaez, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2022. RepliGES and GEStory: Visual Tools for Systematizing and Consolidating Knowledge on User-Defined Gestures. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI 2022), June 6–10, 2022, Frascati, Rome, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3531073.3531112

AVI 2022, June 6–10, 2022, Frascati, Rome, Italy

ACM ISBN 978-1-4503-9719-3/22/06…$15.00
https://doi.org/10.1145/3531073.3531112

1 INTRODUCTION

A critical aspect of designing gesture user interfaces is represented by the mapping between gestures and system functions. A large body of scientific knowledge exists on this topic represented by (i) findings from experiments examining the factors that impact the user experience of gesture-based interaction, such as social acceptability [44], memorability [37], perceived difficulty [61], and naturalness [22], and (ii) insights into users' preferences for gestures elicited in gesture elicitation studies (GESs) [58,69,70]. GESs, as a specialized instance of participatory design [51], are a powerful method in the toolbox of HCI researchers and practitioners to unveil users' preferences for and mental models of gesture-based interaction. In these studies, participants propose gestures they would like to use to effect the system functions of interactive systems, such as zooming in on a map with touch input [70] or turning on/off the TV set with a smart ring [16]. The practitioner analyzes the elicited gestures to develop an understanding of users' preferences for gesture commands that are perceived to be intuitive [69], low effort [70], and memorable [37] and, consequently, likely to reflect the preferences of the larger user population. With time, this understanding turns into consolidated knowledge [29,65,66], on which the community can capitalize to inform the design of gesture-based interactions reflective of end users' preferences.

Since the introduction of the method [69], a large number of GESs have been conducted [65] and have reported findings and design recommendations on touch input [70], mid-air hand gestures [59], interactions in AR [41], radar gestures [33], smartphone motion gestures [46], deformable displays [55], smartwatches and smartglasses [12], smart rings [16], public displays [45], and in-vehicle input [9]. These results constitute a substantial body of scientific and design knowledge about users-defined gestures. However, this rich body of knowledge has several shortcomings that limit the wide application and exploitation of the findings of GES.

- S_1. *Scattered literature*: GESs are published at various venues, from human-computer interaction [70] to intelligent transportation [9] to computational intelligence and design [22], making their findings difficult to identify and access.
- S_2. *Heterogeneous findings*: GESs employ different methods of analysis, not always consistent with each other.

S_3. *Partial coverage of context*: Some GESs lack a full description of the context of use in which the gestures are elicited. The context is sometimes unspecified (no information about the target users, the intended environment, or other contextual conditions) or partially specified (some dimensions are explicitly reported whereas others are missing or implied).

S_4. *Partial evidence*: Some GESs do not provide any indication whether the user-defined gestures they report are representative of consensus among different contexts or user populations, or if the consensus emerged during that study alone.

S_5. *Difficulty of cross-checking findings*: Findings from different GESs are difficult to structure and put in correspondence because of the different methods, measures, target end-user populations, devices, contexts of use, or missing information.

These shortcomings can be addressed in the context of reproducible research. However, replications of GESs have been scarce and, when conducted, they surfaced replication problems. For instance, when reanalyzing data collected by prior work, Tsandilas [56] noted that he "could not reproduce the agreement values that the [original] authors reported because the similarity criteria that they used to classify gestures were ill-defined" (p. 18:43), which is an example of failed method reproducibility [19]. In their GES about interactions with 3D objects on a public display, Du et al. [13] concluded that "repeating the study with participants from other groups (e.g., children, older people, or people with little technology experience) would lead to a more complete picture of users' perceptions and needs regarding interaction with 3D objects shown on public displays" (p. 201), which suggests that the reproducibility of the results [19] may depend on the category of users.

Furthermore, while discussing whole-body gestures, Vatavu [60] showed how the choice of similarity criteria that practitioners could use to group the elicited gestures into classes of similar types has a direct impact on the magnitude of agreement rates reported in GESs, i.e., both the reproducibility of the results and the inferential reproducibility [19] are threatened.

Replications of GESs can be found in just a handful of papers, of which just four [24,38,47,52] have considered replication as their primary goal, to the best of our knowledge. Possible causes for this state of things are the degree that replications are in general small in HCI [26], and the concept of *replication* itself, which may be confusing to researchers due to the many definitions[1] in the scientific literature [1,2,8,19,26,42]. Furthermore, there is no formalization of how replications could be implemented for the specifics of GESs or how existing GES findings could be extended or generalized. In this context, the contributions of our work are manyfold:

(1) We formalize replication, extension, generalization, and repurposing of GES findings with RepliGES, a conceptual space that specifies five types of replications, three types of extensions and generalizations, and two types of repurposing studies that reuse previous GES findings for new discoveries; see Figure 2.
(2) Based on RepliGES, we introduced GEStory, an interactive design space meant to structure, visualize, and identify user-defined gestures from an electronic database of gestures reported by 216 published GESs; see Figure 3.
(3) Using GEStory, we identify a set of twelve GES replications in the scientific literature, which we characterize using the dimensions of our RepliGES conceptual space; see Figure 6.
(4) We propose a set of replicability criteria to foster replications, extensions, generalizations, and repurposing of findings about user-defined gestures; see Figure 1.

[1]To keep consistent with RepliCHI terminology [14,26,68], we use *replication* in this paper as an umbrella term to denote all possible instances of repeating, reconducting, reproducing, reimplementing and repurposing a GES. As we advance with our discussion, we characterize specific instances of replicating GESs (Figure 2).

2 RELATED WORK

We start our discussion of related work with a brief overview of the gesture elicitation method [69,70], and we describe recent developments and tools to support its implementation. We also discuss research reproducibility in Computer Science and HCI.

2.1 Gesture Elicitation Studies

A GES presents participants with system effects and elicits gestures that trigger those effects [69,70]. For example, Vatavu and Zaiţi [64] elicited mid-air gestures to control various functions of a smart TV, such as turning the TV on and off, going to the next and previous channels, etc. These system functions, or the results of user actions performed with interactive systems, are called "referents" in the GES scientific literature [70]. User-proposed input to invoke the referents takes the form of gesture commands in GESs but, for generic end-user elicitation [58,62,63], it can represent any symbol, action, or manifestation of user preference relevant to the study and interactive system, such as key presses, voice input, preferences for vibrotactile patterns, etc. The results of a GES are usually represented by a consensus gesture set, observations about user behavior with respect to the system under investigation, and design recommendations for interactive devices, applications, and systems controllable with gestures. Consensus among the elicited gestures has been calculated using agreement scores [69,70], agreement and coagreement rates [62,63], max-consensus and consensus-distinct ratios [35], chance-corrected coefficients of agreement [56], and the growth rate of the dissimilarity-consensus curve [60]; see Vatavu and Wobbrock [58] for an overview of these measures, clarifications, and recommendations about agreement calculation in end-user elicitation studies. Also, several tools are available to help practitioners implement GES and analyze the gestures elicited [5,6,34,62,63].

2.2 Research Reproducibility

Reproducibility is the cornerstone of cumulative knowledge and a requirement to confirm scientific truth [19,48]. Reproducibility is supported by clear descriptions of procedures, methods, experimental designs, and publicly available resources. The lack of public data hinders the reproducibility, repeatability, and verifiability of experiments since recollecting new data may be a tedious endeavor [43]. The "reproducibility crisis" [50] denotes the situation in which scientific studies are difficult, if not impossible, to reproduce. Many efforts have been made to raise awareness of the importance of reproducibility in scientific research and structure it [19,39,40], including in Computer Science [18] and HCI [20,26,68]. For example, Peng [40] considered that scientific articles are not replicable unless

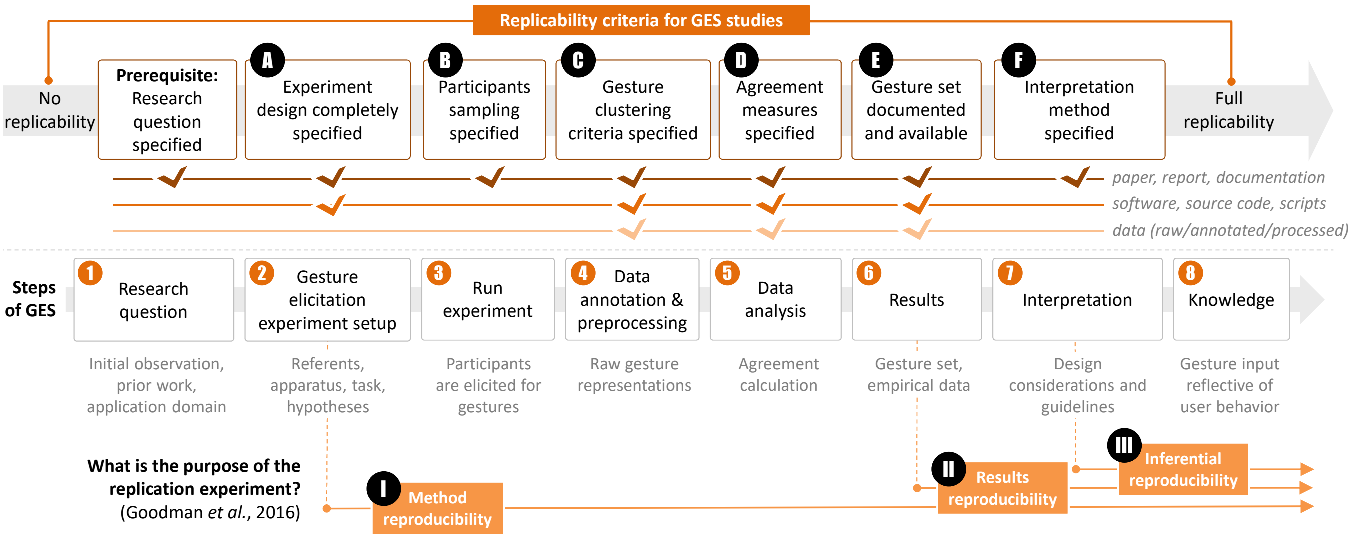

Figure 1: The RepliGES methodological framework illustrating replication configurations and replicability criteria (top) corresponding to the steps of a GES (middle) and the goals of research reproducibility from Goodman et al. [19] (bottom).

their artifacts are available. The next steps are to release the source code and data [14]. Patil et al. [39] proposed a visual notation to express to what extent reproducibility and replicability are addressed in a scientific article. The HCI community has also acknowledged the need for reproducible research with the RepliCHI initiative [68] but, according to the findings reported by Hornbæk et al. [26], a replication rate of just 3% exists in HCI. (This rate is very close to the 4% that we found for replications of GESs.) Consequently, the HCI community should embrace replication to advance the field, especially when the outcomes are likely to question previous results [25].

Following the RepliCHI initiative, the term *replication* has been widely adopted in HCI, while *reproducibility* has been used in other fields [19]. Although an ontological confusion exists because of different terminology, this confusion is actually generalized across the entire spectrum of scientific research; see Gomez et al. [18] that reported 27 different frameworks to classify reproducibility. A relevant example is the ACM [1,2] artifact review and badging system that specifies: "A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms." Thus, in 2018, ACM [1] proposed definitions for *repeatability*, *reproducibility*, and *replicability* for experimental Computer Science, only to revise them in 2020 by swapping *reproducibility* and *replicability* to harmonize ACM terminology with that used by the broader scientific community [2].

In this context, we choose to remain consistent with RepliCHI terminology [26,68] and use *replication* as the umbrella term to refer to any form of reproducible research on interactions between users and computer systems, and *generalization* to refer to reproducible research that covers other contexts. Furthermore, we identify *repurposing* as a new scenario, connected to replication, where datasets collected in prior GESs are used (repurposed) to generate new discoveries, as we have found in the GES literature [17,53]. In the next section, we discuss in detail specific types of replications, extensions, generalizations, and repurposing relevant for GESs.

3 REPLIGES

To structure possible types of GES replications, we start from (1) the steps needed to conduct a GES, from the initial formulation of the research question to the interpretation of results, (2) Johansen's [27] time-space matrix for specifying possible configurations involving the same or different entities, which we connect to ACM's [2] dimensions of team and experimental setup, and (3) existing classifications of reproducible research [1,2,19,26,40,57].

3.1 Preliminaries

Figure 1, middle illustrates the steps involved by a GES. A precise specification of each step is necessary to understand *what* can be replicated and *where* replication applies. Based on the original description of Wobbrock et al.'s [69] method to maximize guessability, its first application to user-defined hand gesture input [70], and a recent model of end-user elicitation in HCI [58], we identify the following steps of a GES: (1) the research question is formulated based on initial observations, prior work, and/or the specifics of the application domain; (2) the experimental setup is specified in terms of hypotheses, context of use, referents, apparatus, tasks, and procedure; (3) the population of potential end users is sampled and participants are elicited about their preferences for gesture commands; (4) the raw data is annotated and preprocessed; (5) consensus analysis follows to identify gestures that are equivalent or substantially similar in accordance with the specifics of the application domain and the goals of the study; (6) results are compiled in the form of a consensus gesture set and/or observations about participants' preferences for gestures; (7) results are interpreted and conclusions drawn; and (8) conclusions contribute to the accumulation of knowledge in the community.

These steps are useful to identify ways in which replications of a GES can be conducted. For example, replication can be done by reusing the same data in step (5) to verify the results of the original study or by repurposing the data in step (4) for another goal, thus arriving at new discoveries. The study could be replicated with new participants at step (3) or with new referents at step (2) to learn whether the conclusions of the original study still hold. To structure these possible types of replication, we turn our attention to several classification schemes for reproducible research.

3.1.1 Artifacts. ACM's [2] artifact review and badging system defines an artifact as "a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself." Following this definition, artifacts in GESs are the data (i.e., the gestures elicited) collected from the participants. Thus, *data* and *participants* are key concepts for GES replication. Consequently, we differentiate among reviewing the same data (i.e., verifying the analysis), collecting new data from the same participants (i.e., repeating the study) and collecting new data by involving new participants (i.e., performing an exact replication [57]).

3.1.2 Team and experimental setup. ACM's [2] system employs two dimensions, *team* and *experimental setup*, to define *repeatability* (same team, same experimental setup), *reproducibility* (different team, same experimental setup), and *replicability* (different team, different experimental setup). Our previous discussion of GES artifacts and focus on gesture data and participants falls under the experimental setup dimension. Next in this section, we show how the method of analysis [57] completes this dimension to specify possible types of replication for GES. However, ACM's [2] team dimension has a limitation: the configuration "same team, different setup" is not covered, although relevant for GESs when the same authors wish to conduct a variation of their original study to expand their previous findings. For example, Soni et al. [53] analyzed mental models of gestures performed on spherical displays by the participants involved in one of their previous studies [52]. To address this limitation, we draw inspiration from Johansen's [27] time-space matrix, which we reframe with the two dimensions of ACM, team and experimental setup. However, the composition of the team[2] conducting the replication is less important than the actual experimental setup used during the replication, i.e., whether changes have been implemented in the original method or not.

3.1.3 Types of replications. Tsang and Kwan [57] classify replications on two dimensions: *sources of data* (same data set, same population and different population) and *changes in the method/procedure* (same measurement and analysis, different measurement and/or analysis). They discuss six types of replication: checking of analysis (which corresponds to ACM's [2] definition of repeatability), reanalysis of data (different procedures are used on the same data), exact replication (the same study conducted with new participants from the same population), conceptual extension (new procedure applied to new participants from the same population), empirical generalization (same procedure used as in the original study but on a new population), and generalization and extension (new procedure and new population). For Tsang and Kwan, the *team* that is involved in replication is not important, but *population* is. Since generalizations of findings from GESs to other populations (e.g., from adults to children [47,52,70]) are valuable and insightful to consolidate knowledge about user-defined gestures, we adopt the *population* dimension in our RepliGES space next to *data* and *participants* by considering the scenario where new data are collected from participants sampled from a population different from the one from the original GES; see the horizontal axis of Figure 2.

[2]In some cases, new researchers collaborate with the original team, while other researchers from the original team are not part of the follow-up study; see [47,52,53] for examples. Such situations, occurring in practice, are not covered by ACM's [1,2] categories of the same or a different team.

3.1.4 Types of reproducibility. Goodman et al. [19] highlighted three types of reproducibility: (I) *method reproducibility* strives to make the data and experimental setup fully accessible so that the experiment can be repeated; (II) if the experiment can be repeated with the same results, *results reproducibility* is accomplished; and (III) *inferential reproducibility* describes the situation where the same conclusion is drawn from the replication. These types of reproducibility apply to a GES at steps (2), (6), and (7), respectively, and are highlighted at the bottom of Figure 1.

3.2 The RepliGES Conceptual Space

Based on the above considerations, we identify the following types of replication for GESs (see Figure 2):

- ❶ *Repeatability.* The data from the original study is reanalyzed with the same method to verify whether the same results are obtained again. This type of replication (denoted with symbol ❶ in Figure 2) connects to Tsang and Kwan's [57] "checking of analysis" and ACM's [2] definition of "repeatability." One frequent use in GESs is when several researchers from the same team code the gestures independently, after which an inter-rater reliability test is applied to confirm the consistency of the coding. For example, in their study about user-defined gestures for deformable displays, Troiano et al. [55] described their procedure as follows: "After the coding manual was finalized, one author coded all the tasks, while a second author independently coded a sub-set of tasks (10% of the whole set). An inter-rater reliability analysis was performed using Cohen's Kappa statistic to determine consistency among raters" (p. 4). Checking of analysis can also happen at a later time by the same or a different team.
- ❷ *Reproducibility.* The elicitation data from the original study is revisited with a new method of analysis. This type of replication (symbol ❷ in Figure 2) employs the research question and data from the original study and the team verifies whether the results can be obtained again, but with new methods. Tsang and Kwan [57] note that "quite often the replication involves using more powerful statistical techniques that were not available when the original study was conducted" (p. 766). The team can be an independent one (according to ACM's [2] reproducibility), but also the same team may wish to revisit their data with a new method (i.e., Tsang and Kwan's [57] "reanalysis of data"). Relevant examples are Vatavu and Wobbrock [58,62,63] and Tsandilas [56].

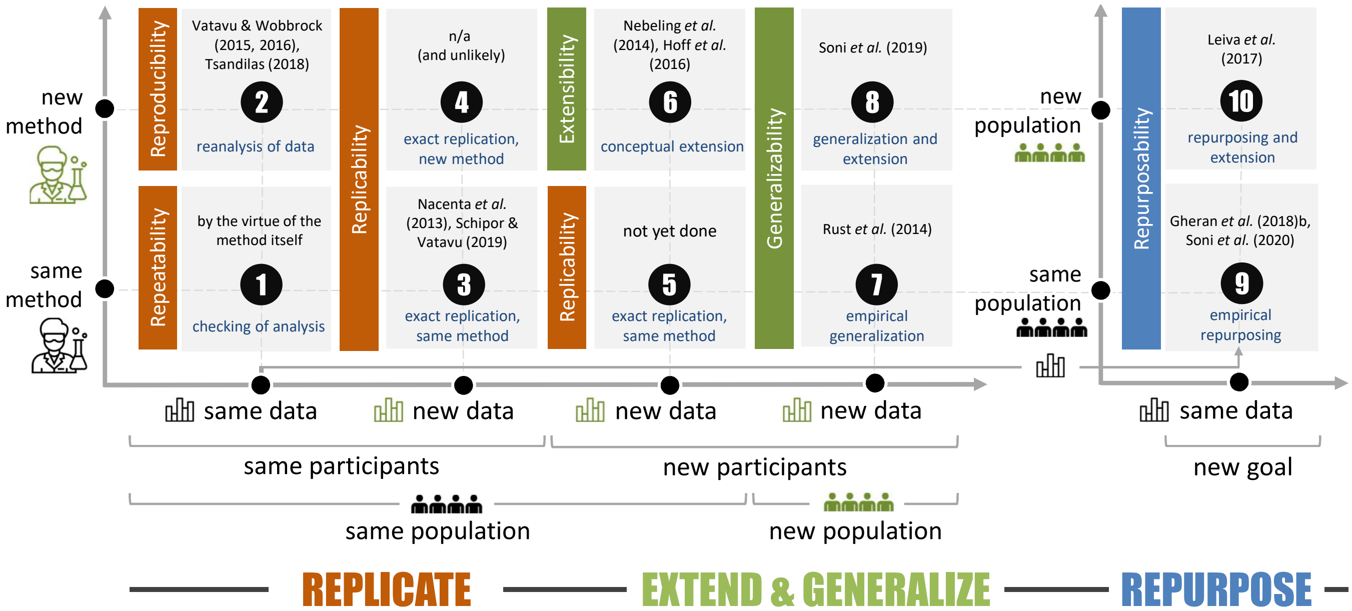

Figure 2: The RepliGES conceptual space for gesture elicitation studies.

❸❹❺ *Replicability.* The team collects new data, either from the same participants of the original study (❸) or new participants (❹❺). Examples of the former (❸) are Nacenta et al. [37] and Schipor and Vatavu [49] that recalled their participants after twenty-four hours. Replicability conducted by applying a new method at a later time with the same participants (❹) is probably the most challenging, if not impossible, type of replication since the participants may not be available any longer or, when they are, some aspect of the study will have probably changed over time. According to Tsang and Kwan [57], "Strictly speaking, the same study can never be repeated by a different researcher, or even by the same researcher" (p. 756). In another instance of replicability (❺), the research team collects new data from new participants to verify if the results from the original study still hold, i.e., an "exact replication" [57] or a "strict replication" [26].

So far we have focused on GES replications. Next, we build on Tsang and Kwan's [57] dimensions of the sources of data and the method to specify *extensions* and *generalizations* of GES findings as follows:

❻ *Extensibility.* The team employs a new method on data collected from new participants, i.e., they perform a "conceptual extension" [57] of the original study. For example, Nebeling et al. [38] replicated Morris' [35] "web on the wall" study with several variations of the original method: a system implementation, a custom software for recording and analyzing multimodal interactions, and a mixed-initiative elicitation procedure. Another example is Hoff et al. [24] that adapted and tested two techniques previously proposed to reduce legacy bias [36], i.e., production and priming.

❼ *Generalizability.* The research team runs the study by employing the original method, but with a sample of participants from another population to check whether the original findings generalize to the new population. One example is Rust et al. [47] that reimplemented Wobbrock et al.'s [70] study on tabletop interaction by eliciting gestures from children.[3]

❽ *Extensibility and generalizability.* The research team uses a new method on new gestures elicited from a new population. An example is Soni et al. [52] that examined gestures performed by children and adults on spherical displays and compared them to gestures performed on flat tabletops [47,70].

Besides the above categories, our survey of the GES literature revealed two studies [17,53] in which gestures elicited by previous work [16,52] were reused for another purpose. This specific type of reutilization of previously collected data for another goal is not covered by the ACM [2] or Tsang and Kwan's [57] classifications, the two main sources that have informed our RepliGES space so far. Nevertheless, reusing data is relevant for consolidating knowledge about user-defined gestures since it has the potential to generate new results, complementary to those from the original study that collected the data. Such studies can be performed by the same team or by an independent team. To reflect the change in goal without the effort of collecting new data, we use the term *repurposability*[4] and complete our RepliGES space as follows:

❾ *Repurposability for the same population.* The gesture data from a previous GES are employed for a new goal to discover something new about the same population. For example, Gheran et al. [17] reused ring gestures elicited in [16] to model bimanual gestures using temporal calculus. Soni et al. [53] reused the dataset from [52] to analyze users' mental models for interactions with spherical displays.

[3] Adaptations were adopted to make the experiment protocol work for children.
[4] Just like the use of the term in computer programming; see https://www.jefftk.com/p/programming-repurposeability.

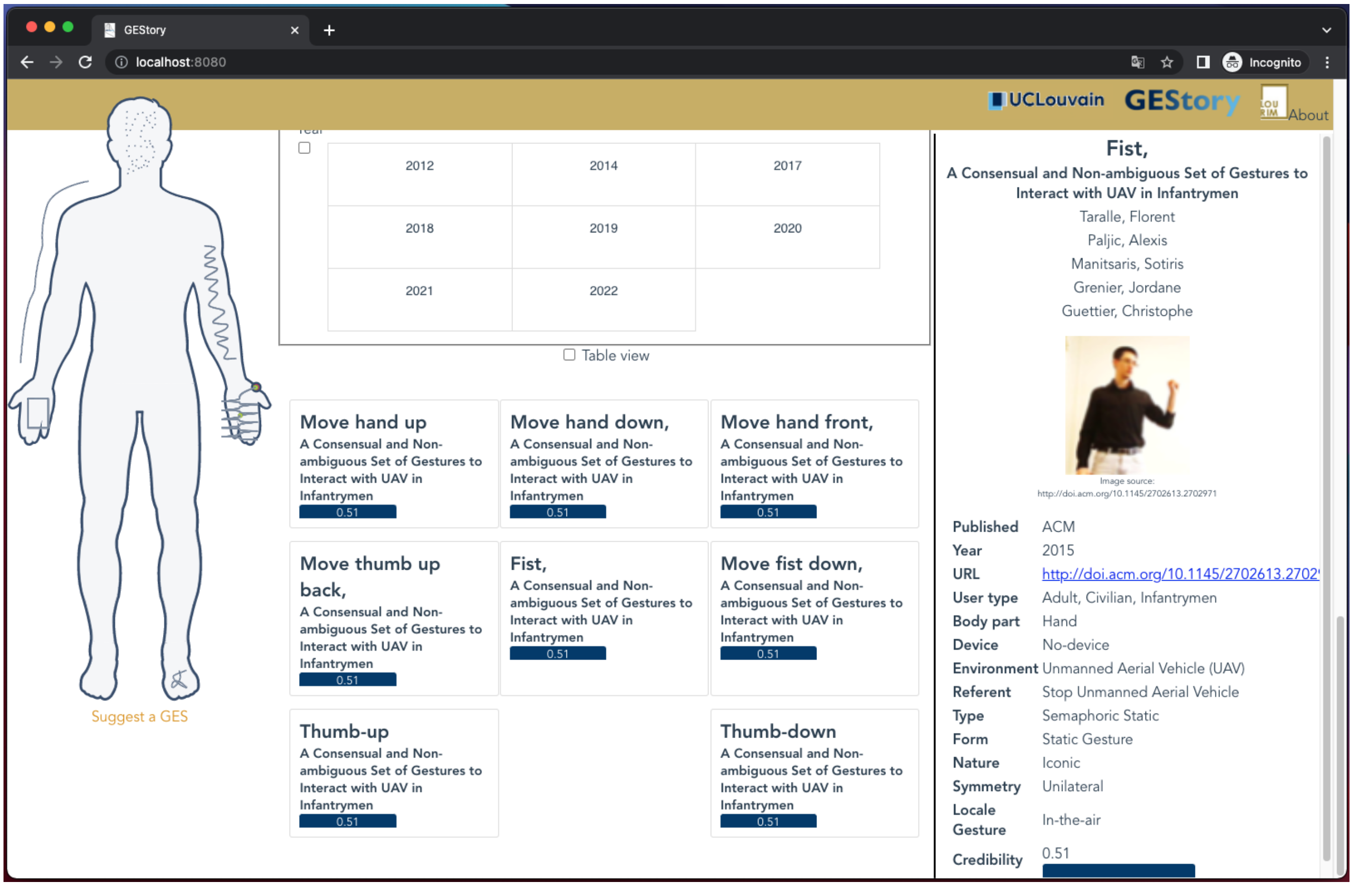

Figure 3: The GESTORY interactive design space implemented as a web application.

❿ *Repurposability for a new population.* An interesting case of repurposability is when the team employs a previously collected gesture dataset, elicited from participants belonging to one population, to arrive at new discoveries that apply to a different population. We did not find any example of this scenario applied to GESs, but we point to Leiva et al. [31], who used the Kinematic Theory to synthesize stroke gestures with the articulation characteristics of users with visual impairments from gestures collected from people without visual impairments. Although their study was conducted on gestures collected in response to visual stimuli instead of being elicited for specific referents as in a GES, Leiva et al.'s work represents the closest example of repurposability of gesture data applied to a new population and, we believe, is inspiring to foster similar work in GESs.

4 GESTORY

GESTORY (Figure 3) is an interactive design space [23] that structures gestures according to a domain model reproduced in Figure 4. To support the visualization of gesture-related information, dynamic query filters are tightly coupled with a starfield display [3] of the human body. In this way, gestures are shown in correspondence to the body limbs used for their articulation. For example, Figure 3 shows a dynamic query of hand gestures. GESTORY returns all of the gestures reported by previously published GESs, and couples the results with a starfield display superimposed on the human body; see Figure 3. When a gesture is selected, the dimensions according to which that gesture is characterized are shown below.

- *Body part* specifies the part(s) of the human body that are involved in gesture articulation according to the limb classification introduced by Villarreal et al. [65] for GESs: finger, hand, writs, arm, shoulder, head, foot, and torso.
- *Device* specifies the device or sensor used for gesture acquisition. According to the sensing technique, devices fall into one of the following categories: touch or contact-based devices and contact-less or vision-based devices. When custom-built devices are used, the "Prototype" slot is highlighted in GESTORY. For example, a smartwatch can be used as a contact-based device or a contact-less sensor.
- *User* specifies the category or profile of the target users addressed by a GES, e.g., children [47,52].
- *Task* specifies the referents from a GES according to Lenorovitz et al.'s [32] classification of actions and Aigner et al.'s [4] classification of tasks: constructive action, scale, next object/action, previous object/action, draw an object, rotate an object, increase/decrease a value, activate or deactivate an

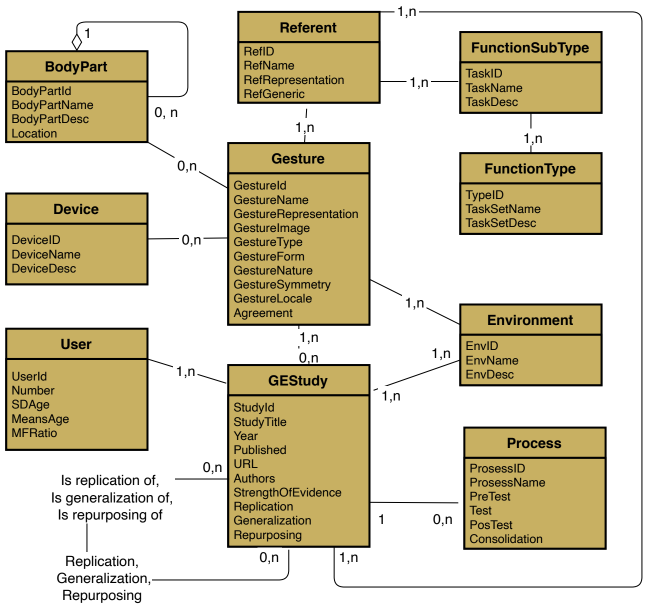

Figure 4: The domain model used for GEStory.

object or an action, and OK/cancel. This list can be extended to accommodate other types of tasks as well.

- *Gesture type* specifies the gesture category according to Aigner et al.'s [4] classification of gestures: pointing, static or static semaphoric, pantomimic, static or dynamic iconic and manipulative gestures. For example, pointing refers to gestures that indicate an object, person, or direction that are usually performed, but not always, with the index finger.
- *Gesture form* indicates whether the gesture is static, e.g., a hand pose, or dynamic, e.g., a mid-air hand gesture.
- *Gesture nature* specifies the nature of the gesture according to Kendon's [28] continuum.
- *Symmetry* indicates whether the gesture exhibits any form of body symmetry according to Guiard's [21] kinematic chain model: unilateral (involving one side), bilateral symmetric (involving two sides in a symmetric way), and bilateral asymmetric (two sides in a different way). For example, the gesture depicted in Figure 3 is unilateral.
- *Locale* specifies the centricity of the gesture: object-centric when the gesture is based on an object, body-centric, when it refers to a part of the human body, in-the-air, or mixed when several locales are combined.
- *Year* specifies the publication year of the GES results.

GEStory is publicly available as a web-based application at the address https://sites.uclouvain.be/GEStory. The GitHub repository documents the implementation in HTML5, JavaScript, and Vue.js[5]. JSON files implement the database of GEStory. Each GES is also fully stored with its bibliographic reference extracted from the Zotero collection of Villarreal et al. [65]: title, authors, venue, editors, year of publication, and DOI. Dimensions *User*, *Task*, and *Device* cover the typical dimensions that characterize the context of use [10].

[5]https://vuejs.org/v2/guide

5 A SCIENTIFIC LITERATURE REVIEW OF GES REPLICATIONS

To identify GES replications, we turned to reviews of the scientific literature on GESs. Vuletic et al. [54] conducted a Systematic Literature Review (SLR) on hand gestures for user interfaces; Vogiatzidakis and Koutsabasis [66] conducted a review of GESs for mid-air interaction, followed by a SLR [29] with a corpus of 47 papers; and Villarreal et al. [65] conducted a SLR of 216 GESs with a Zotero collection[6] that was made publicly available. Using these sources, we found only ten papers, a mere 4% of the 216 studies examined in [65], which are highlighted in the RepliGES space in Figure 2 and detailed in Figure 6.

Vatavu and Wobbrock [62,63] reused the gesture datasets from [41, 59,64] to exemplify the use of the AGATe toolkit. Tsandilas [56] performed reanalysis of the data from [7,11,30,67] with different statistical methods. Nebeling et al. [38] replicated the "Web on the Wall" study by Morris [35] to highlight the importance of following up a GES with implementations of actual gesture recognizers, user interfaces, or interactive systems informed by the findings of the GES. Nacenta et al. [37] and Schipor and Vatavu [49] involved participants in subsequent sessions to assess the precision of their memory recall. Hoff et al. [24] addressed the aspect of legacy bias in gesture elicitation [36]. Gheran et al. [17] reused the smart ring gestures elicited by [16] to model bimanual gesture input with temporal calculus. And Soni et al. [52,53] examined gestures produced by children and adults on spherical and flat touch displays and discussed their findings with respect to prior work [47,70].

6 CONCLUSION AND FUTURE WORK

We introduced the RepliGES conceptual space to address several shortcomings (S_1 to S_6) about the body of knowledge accumulated on user-defined gestures in the scientific literature. Based on RepliGES, we developed GEStory, an interactive design space to structure, visualize, and query user-defined gestures reported by 216 GESs, which we demonstrated by identifying a set of replicated GESs. Our results have implications for the practice of end-user gesture elicitation, in support of which we provide replicability criteria and recommendations for researchers and practitioners. With these tools, the body of knowledge on user-defined gestures is concentrated in one single point of contact (S_1), where gestures are systematically specified with a unified format (S_2) in a consistent manner (S_3) and that is structured according to a domain model (S_5) to identify RepliGES links between different published GESs.

RepliGES and GEStory are mainly intended for designers and developers of gesture-based user interfaces, who can rely on the gesture database concentrated by GEStory and the replication opportunities identified by RepliGES to gain new insights and obtain new information in their specific application domain. RepliGES and GEStory are also intended for researchers to inform new experiments and studies toward the discovery of new gestures, but also to confirm or disconfirm existing GES findings. The interactive design space of RepliGES can be used to inform and support the design of specific gesture types since it incorporates dimensions that specify constraints and trade-offs relevant to gesture-based

[6]https://www.zotero.org/groups/2132650/gesture_elicitation_studies

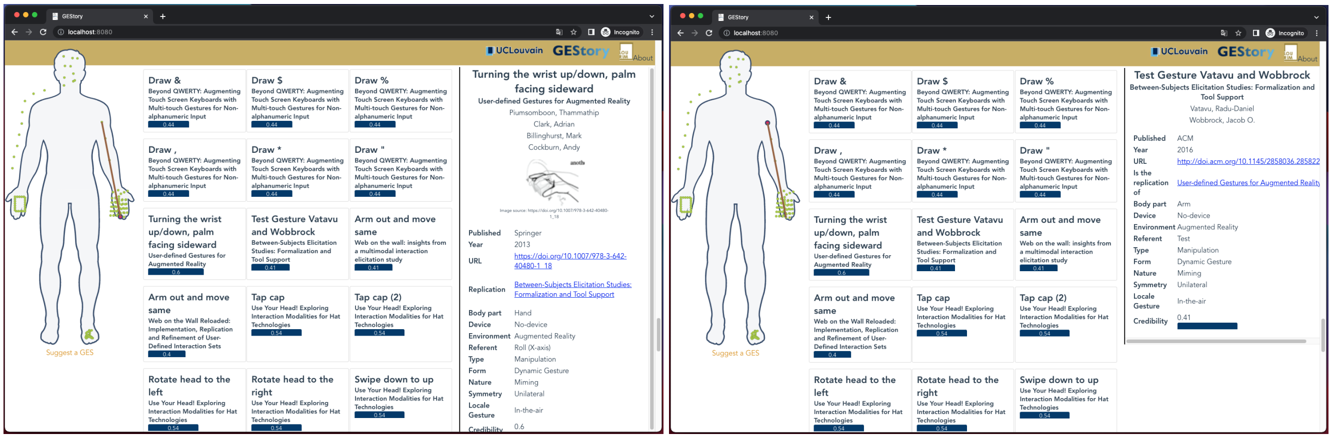

Figure 5: Identifying a GES replication in GESTORY: the source study (left) and the replication study (right).

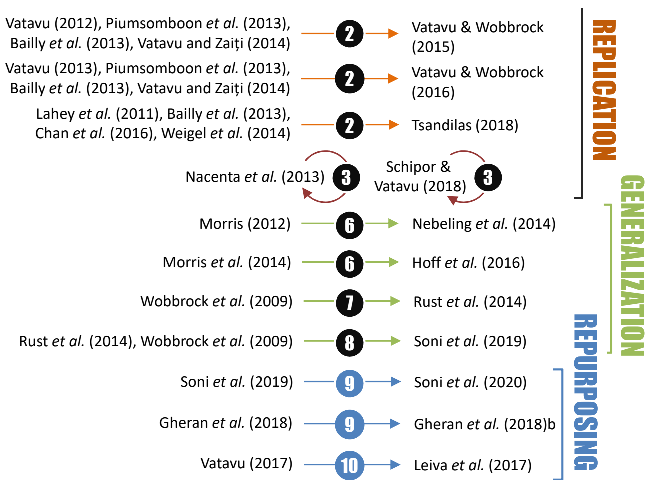

Figure 6: The papers [17,24,37,38,49,52,53,56,62,63] identified in our literature survey with replications, generalizations, and repurposing of previous GES results. *Note:* the numbers on the arrows correspond to those used in Figure 2.

interaction. Finally, GESTORY is also intended for historians of user interfaces since it captures the evolution of gestures reported by GESs. The repository of GESTORY is valuable to anyone interested in the evolution of gestures for interactive applications.

In future work, we plan replications of GESs for specific devices, e.g., electronic rings [16], or addressing specific user categories, e.g., people with motor impairments [15]. Findings from these studies and replications will also allow us to understand more closely aspects of the end-user gesture elicitation method, e.g., the impact of the analysis method on the results or the generalizability of the results for other user populations and contexts of use.

ACKNOWLEDGMENTS

R.-D. Vatavu acknowledges support from a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, PN-III-P4-ID-PCE-2020-0434 (PCE29/2021), within PNCDI III.

REFERENCES

[1] ACM. 2018. Artifact Review and Badging – Version 1.0 (not current). https://www.acm.org/publications/policies/artifact-review-badging
[2] ACM. 2020. Artifact Review and Badging - Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current
[3] Christopher Ahlberg and Ben Shneiderman. 1994. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proc. of CHI '94*. ACM, 313–317. https://doi.org/10.1145/191666.191775
[4] Roland et al. Aigner. 2012. *Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI.* Technical Report MSR-TR-2012-111.
[5] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2018. Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies. In *UIST '18*. ACM, 177–188. https://doi.org/10.1145/3242587.3242621
[6] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2019. Crowdlicit: A System for Conducting Distributed End-User Elicitation and Identification Studies. In *Proc. of ACM Int. Conf. on Human Aspects in Computing Systems (CHI '19)*. ACM, Article 255, 12 pages. https://doi.org/10.1145/3290605.3300485
[7] Gilles Bailly, Thomas Pietrzak, Jonathan Deber, and Daniel J. Wigdor. 2013. Métamorphe: Augmenting Hotkey Usage with Actuated Keys. In *CHI '13* (Paris, France). ACM, 563–572. https://doi.org/10.1145/2470654.2470734
[8] Fabien C.Y. Benureau and Nicolas P. Rougier. 2018. Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions. *Frontiers in Neuroinformatics* 11 (2018), 69. https://doi.org/10.3389/fninf.2017.00069
[9] Laura Bianca Bilius and Radu-Daniel Vatavu. 2021. A Multistudy Investigation of Drivers and Passengers' Gesture and Voice Input Preferences for In-Vehicle Interactions. *Journal of Intelligent Transportation Systems* (2021), 24 pages.
[10] Gaëlle Calvary, Joëlle Coutaz, David Thevenin, Quentin Limbourg, Laurent Bouillon, and Jean Vanderdonckt. 2003. A Unifying Reference Framework for multi-target user interfaces. *Interacting with Computers* 15, 3 (2003), 289–308. https://doi.org/10.1016/S0953-5438(03)00010-9
[11] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User Elicitation on Single-Hand Microgestures. In *CHI'16*. ACM, 3403–3414. https://doi.org/10.1145/2858036.2858589
[12] 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 *CHI '18*. ACM, Article 419, 12 pages. https://doi.org/10.1145/3173574.3173993
[13] Guiying Du, Auriol Degbelo, Christian Kray, and Marco Painho. 2018. Gestural Interaction with 3D Objects Shown on Public Displays: An Elicitation Study. *IxD&A* 38 (2018), 184–202. https://dblp.org/rec/journals/ixda/DuDKP18
[14] Florian Echtler and Maximilian Häuundefinedler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *CHI EA '18*. Article alt02, 8 pages. https://doi.org/10.1145/3170427.3188395
[15] Mingming Fan, Zhen Li, and Franklin Mingzhe Li. 2021. Eyelid Gestures for People with Motor Impairments. *Commun. ACM* 65, 1 (dec 2021), 108–115.
[16] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *DIS '18*. ACM, 623–635. https://doi.org/10.1145/3196709.3196741
[17] Bogdan-Florin Gheran, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2018. Ring X2: Designing Gestures for Smart Rings Using Temporal Calculus. In *DIS '18 Companion*. ACM, 117–122. https://doi.org/10.1145/3197391.3205422
[18] Omar S. Gómez, Natalia Juristo, and Sira Vegas. 2010. Replications Types in Experimental Disciplines. In *ESEM '10*. ACM, Article 3, 10 pages. https://doi.

org/10.1145/1852786.1852790

[19] Steven N. Goodman, Daniele Fanelli, and John P.A. Ioannidis. 2016. What Does Research Reproducibility Mean? *Science Translational Medicine* 8, 341 (June 2016), 341ps12. http://dx.doi.org/10.1126/scitranslmed.aaf5027

[20] Christian Greiffenhagen and Stuart Reeves. 2013. Is Replication Important for HCI?. In *Proc. of the CHI '13 Workshop on the Replication of HCI Research* (Paris, France). http://ceur-ws.org/Vol-976

[21] Yves Guiard. 1987. Asymmetric Division of Labor in Human Skilled Bimanual Action: The Kinematic Chain as a Model. *Journal of Motor Behavior* 19 (1987), 486–517. http://www.cogprints.org/625/1/jmb_87.html

[22] Zhifan He, Ruifo Zhang, Zheng Liu, and Zhengyu Tan. 2020. A User-Defined Gesture Set for Natural Interaction in a Smart Kitchen Environment. In *Proc. of ISCID '20*. 122–125. https://doi.org/10.1109/ISCID51228.2020.00034

[23] Florian Heller, Kashyap Todi, and Kris Luyten. 2021. An Interactive Design Space for Wearable Displays. In *Proc. of MobileHCI '21*. ACM, Article 4, 14 pages. https://doi.org/10.1145/3447526.3472034

[24] Lynn Hoff, Eva Hornecker, and Sven Bertel. 2016. Modifying Gesture Elicitation: Do Kinaesthetic Priming and Increased Production Reduce Legacy Bias?. In *TEI '16*. ACM, 86–91. https://doi.org/10.1145/2839462.2839472

[25] Kasper Hornbæk. 2015. We Must Be More Wrong in HCI Research. *Interactions* 22, 6 (Oct. 2015), 20–21. https://doi.org/10.1145/2833093

[26] 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 *CHI '14*. 3523–3532. https://doi.org/10.1145/2556288.2557004

[27] Robert Johansen. 1991. Groupware: Future directions and wild cards. *J. Organizational Computing* 1, 2 (1991), 219–227. https://doi.org/10.1080/10919399109540160

[28] Adam Kendon. 1997. Gesture. *Annual Review of Anthropology* 26 (1997), 109–128. https://www.jstor.org/stable/2952517

[29] Panayiotis Koutsabasis and Panagiotis Vogiatzidakis. 2019. Empirical Research in Mid-Air Interaction: A Systematic Review. *Int. Journal of Human-Computer Interaction* (2019), 1–22. https://doi.org/10.1080/10447318.2019.1572352

[30] Byron Lahey, Audrey Girouard, Winslow Burleson, and Roel Vertegaal. 2011. PaperPhone: Understanding the Use of Bend Gestures in Mobile Devices with Flexible Electronic Paper Displays. In *CHI '11*. 1303–1312. https://doi.org/10.1145/1978942.1979136

[31] Luis A. Leiva, Daniel Martín-Albo, and Radu-Daniel Vatavu. 2017. Synthesizing Stroke Gestures Across User Populations: A Case for Users with Visual Impairments. In *Proc. of CHI '17*. ACM, New York, NY, USA, 4182–4193. https://doi.org/10.1145/3025453.3025906

[32] David R. Lenorovitz, Mark D. Phillips, R.S. Ardrey, and Gregory V. Kloster. 1984. A taxonomic approach to characterizing human-computer interaction.. In *Human-Computer Interaction*. (Amsterdam, 1984). Elsevier Science Publishers, 111–116.

[33] 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 (2019), 176982–176997. https://doi.org/10.1109/ACCESS.2019.2951349

[34] Nathan Magrofuoco and Jean Vanderdonckt. 2019. Gelicit: A Cloud Platform for Distributed Gesture Elicitation Studies. *PACMHCI* 3, EICS, Article 6 (2019), 41 pages. https://doi.org/10.1145/3331148

[35] Meredith Ringel Morris. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *ITS '12*. ACM, 10 pages. https://doi.org/10.1145/2396636.2396651

[36] 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.

[37] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-Designed and User-Defined Gesture Sets. In *CHI '13* (Paris, France). ACM, 1099–1108. https://doi.org/10.1145/2470654.2466142

[38] 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 *ITS '14*. ACM, 15–24. https://doi.org/10.1145/2669485.2669497

[39] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. 2019. A Visual Tool for Defining Reproducibility and Replicability. *Nat. Hum. Behav.* 3 (2019).

[40] Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. http://dx.doi.org/10.1126/science.1213847

[41] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-Defined Gestures for Augmented Reality. In *CHI EA '13*. 6 pages. doi.org/10.1145/2468356.2468527

[42] Hans E. Plesser. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Front. Neuroinf.* 11 (2018).

[43] Stefan Pröll, Kristof Meixner, and Andreas Rauber. 2016. Precise Data Identification Services for Long Tail Research Data. (01 2016). http://dx.doi.org/10.6084/M9.FIGSHARE.3847632

[44] Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proc. of CHI '10*. ACM, 887–896. https://doi.org/10.1145/1753326.1753458

[45] Isabel Benavente Rodriguez and Nicolai Marquardt. 2017. Gesture Elicitation Study on How to Opt-in & Opt-out from Interactions with Public Displays. In *Proc. of ISS '17*. ACM, 32–41. https://doi.org/10.1145/3132272.3134118

[46] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-Defined Motion Gestures for Mobile Interaction. In *CHI '11*. ACM, 197–206. https://doi.org/10.1145/1978942.1978971

[47] Karen Rust, Meethu Malu, Lisa Anthony, and Leah Findlater. 2014. Understanding Childdefined Gestures and Children's Mental Models for Touchscreen Tabletop Interaction. In *IDC '14*. ACM, 201–204.

[48] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* 9, 10 (2013). https://doi.org/10.1371/journal.pcbi.1003285

[49] 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. https://doi.org/10.1109/MPRV.2018.2873856

[50] Jonathan W. Schooler. 2014. Metascience Could Rescue the "Replication Crisis". *Nature* 515, 9 (Nov. 2014), 9. http://dx.doi.org/10.1038/515009a

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

[52] 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 *PerDis '19*. ACM, New York, NY, USA, Article 24, 7 pages. https://doi.org/10.1145/3321335.3324941

[53] Nikita Soni, Schuyler Gleaves, Hannah Neff, Sarah Morrison-Smith, Shaghayegh Esmaeili, Ian Mayne, Sayli Bapat, Carrie Schuman, Kathryn A. Stofer, and Lisa Anthony. 2020. Adults' and Children's Mental Models for Gestural Interactions with Interactive Spherical Displays. In *CHI '20*. ACM, New York, NY, USA, 1–12.

[54] Vuletic Tijana, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129 (2019), 74 – 94. https://doi.org/10.1016/j.ijhcs.2019.03.011

[55] Giovanni Maria Troiano, Esben Warming Pedersen, and Kasper Hornbundefinedk. 2014. User-Defined Gestures for Elastic, Deformable Displays. In *AVI '14*. 1–8.

[56] Theophanis Tsandilas. 2018. Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *TOCHI* 25, Article 18 (2018), 49 pages. http://doi.org/10.1145/3182168

[57] Eric W.K. Tsang and Kai-Man Kwan. 1999. Replication and Theory Development in Organizational Science: A Critical Realist Perspective. *The Academy of Management Review* 24, 4 (Oct. 1999), 759–780. http://dx.doi.org/10.2307/259353

[58] 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 (2022), 5:1–5:70. https://doi.org/10.1145/3476101

[59] Radu-Daniel Vatavu. 2012. User-Defined Gestures for Free-Hand TV Control. In *EuroITV '12*. ACM, 45–48. https://doi.org/10.1145/2325616.2325626

[60] Radu-Daniel Vatavu. 2019. The Dissimilarity-Consensus Approach to Agreement Analysis in Gesture Elicitation Studies. In *CHI '19*. Article 224, 13 pages. doi.org/10.1145/3290605.3300454

[61] Radu-Daniel Vatavu, Daniel Vogel, Géry Casiez, and Laurent Grisoni. 2011. Estimating the Perceived Difficulty of Pen Gestures. In *INTERACT'11*. Springer-Verlag, Berlin, Heidelberg, 89–106. http://dx.doi.org/10.1007/978-3-642-23771-3_9

[62] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *CHI '15*. ACM, 1325–1334. http://dx.doi.org/10.1145/2702123.2702223

[63] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2016. Between-Subjects Elicitation Studies: Formalization and Tool Support. In *CHI '16*. 13 pages. http://dx.doi.org/10.1145/2858036.2858228

[64] Radu-Daniel Vatavu and Ionut-Alexandru Zaiti. 2014. Leap Gestures for TV: Insights from an Elicitation Study. In *TVX '14*. ACM, 131–138.

[65] 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 *DIS '20*. ACM, 855–872. https://doi.org/10.1145/3357236.3395511

[66] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2018. Gesture Elicitation Studies for Mid-Air Interaction: A Review. *MTI* 2, 4 (2018).

[67] Martin Weigel, Vikram Mehta, and Jürgen Steimle. 2014. More than Touch: Understanding How People Use Skin as an Input Surface for Mobile Computing. In *CHI '14*. ACM, 179–188. https://doi.org/10.1145/2556288.2557239

[68] Max L. Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. RepliCHI - CHI Should Be Replicating and Validating Results More: Discuss. In *CHI EA'11*. ACM, 463–466.

[69] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI EA'05*. 4 pages. http://doi.org/10.1145/1056808.1057043

[70] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *CHI '09*. ACM, 10 pages. https://doi.org/10.1145/1518701.1518866