Four Replications of a User-Defined Gesture Study with Smart Rings

[Jean Vanderdonckt](#)
Université catholique de Louvain
Louvain Research Institute in Management and Organizations (LouRIM)
Louvain-la-Neuve, Belgium
jean.vanderdonckt@uclouvain.be

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

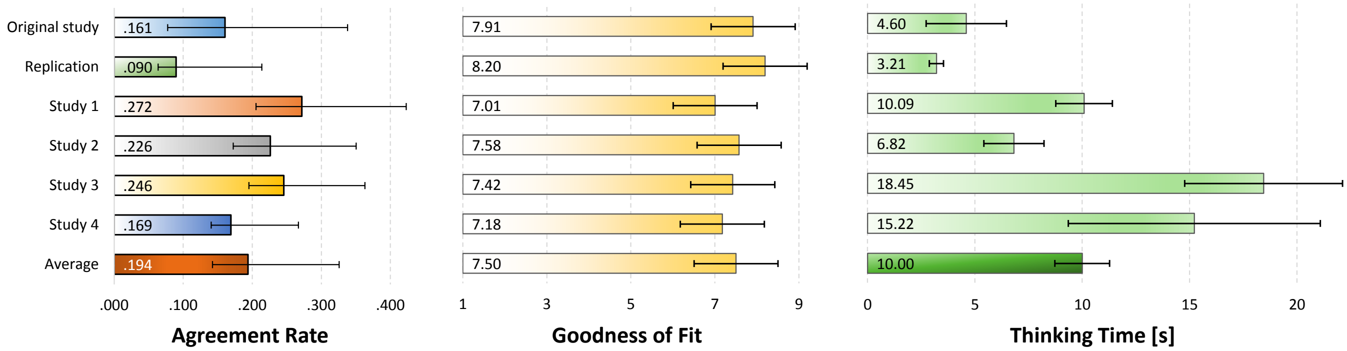

Figure 1: Four replications of an end-user gesture elicitation study involving smart rings are reported in this paper. These charts show the findings of the original study [11], a prior published replication [10], and our four replications in terms of agreement rate (left), goodness of fit (middle), and thinking time (right) measures. *Note:* error bars show 95% confidence intervals.

Abstract

Empirical findings in gesture-based interaction often stem from highly controlled experimental settings, which raises concerns about their generalizability. To explore how variations in such settings influence discoveries on user-defined gestures, we selected an end-user elicitation study involving smart rings that had been replicated at least once. By reusing the same stimuli, equipment, and data collection method, we conducted four new replications of the original study, involving a total of 120 participants across four different research teams. Our results show that smart ring gestures elicited in these replications overlap only partially, with differences in agreement rate, thinking time, and goodness of fit with corresponding system functions. We argue that systematic replication of gesture elicitation studies is essential for generalizable gesture sets.

CCS Concepts

• **Human-centered computing → Gestural input**; *User interface design*; *Participatory design*; *Empirical studies in interaction design*; **Mobile devices**; **Empirical studies in ubiquitous and mobile computing**.

Keywords

Gesture elicitation, Reproducibility, Replicability, Smart rings

ACM Reference Format:
Jean Vanderdonckt and Radu-Daniel Vatavu. 2025. Four Replications of a User-Defined Gesture Study with Smart Rings. In *Adjunct Proceedings of the 27th International Conference on Mobile Human-Computer Interaction (MobileHCI '25 Adjunct), September 22–25, 2025, Sharm El-Sheikh, Egypt.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3737821.3749582

1 Introduction

As mobile computing increasingly integrates gesture input, the intuitiveness and generalizability of user-defined gestures has become a central focus in mobile and wearable interactions [10,11,24]. In a larger context, Xia *et al.* [35] recommended that gesture set design should consider thirteen quality factors across four categories: situational (*e.g.,* context and modality), cognitive (*e.g.,* intuitiveness), physical (*e.g.,* efficiency), and system-related. Through this lens, both the generalizability and transferability of gesture sets across diverse interactive devices, user groups, and contexts of use impact all quality categories above.

Gesture Elicitation Studies (GESs) [34], where participants are prompted to propose their own gestures to effect specific system functions, materialized by referents, have been widely applied to derive gesture sets [31] based on user intuitive behavior [26]. While these studies produce rich insights, the aspects of their replication can pose a real threat [9,13] to generalizability and practical applicability of consensus gesture sets as well as regarding the

MobileHCI '25 Adjunct, Sharm El-Sheikh, Egypt

ACM ISBN 979-8-4007-1970-7/2025/09
https://doi.org/10.1145/3737821.3749582

completeness of the collected gestures. In general, a reproducibility problem [20,21] has been acknowledged in HCI research [14,28,33].

Studies involving mobile interactions remain highly context-dependent [5,6,16,17,24], influenced by many factors, such as device type (*e.g.,* smartphone, watch, glasses) [8], body pose (*e.g.,* seated *vs.* walking), environmental constraints (*e.g.,* public *vs.* private) [23], and sociocultural aspects (*e.g.,* social acceptance) [22]. A gesture set considered intuitive in a lab setting may reveal as impractical or little acceptable in a social context involving real-world mobile interaction. Therefore, replicating GESs across user populations, interaction devices, and diverse environments [12] helps determine which gestures exhibit cross-context stability and transferability compared to those that are just situational artifacts. These aspects complement previous concerns about the legacy bias in end-user gesture elicitation studies [19] and the availability of tools to support their practical implementation [3,18].

As mobile computing is expected to reach diverse end-user populations [32] and contexts of use [31], overfitting gesture design recommendations to narrow samples risks excluding broader user preferences. Consequently, more replications involving demographically diverse and situationally varied samples are necessary. In this context, this paper addresses the following research question:

> What is the effect of replicating a gesture elicitation study, originally conducted in a specific context, in new studies involving the same stimuli, task, and analysis method, but with different participants in different settings and different research teams?

Variation in the setting enables us to examine whether the elicited gestures are influenced by the localization of the study, while variation in the research team allows us to assess the consistency in applying the same analysis method. We evaluated these aspects in the replication studies reported in this work using agreement rate, thinking time, and goodness-of-fit measures [10,11,34]. After reviewing prior work in gesture elicitation with a focus on mobile contexts in Section 2, we present our experimental protocol in Section 3, identically applied for all our four replications, and report findings across our selected set of measures.

2 Related Work

ACM's [1,2] 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."* In our case, data artifacts are represented by the gestures elicited from the participants during the implementation of the GES. Consequently, we differentiate between reviewing the same data (*i.e.,* verifying the analysis) collected in an original study, 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) [27]. We refer readers to the RepliGES space [12], specifically designed for informing replications of end-user gesture elicitation studies with different approaches, including repeatability, repurposability, and extensibility. The Ampliatum framework [28], introduced for contextual reproducibility in human-computer interaction research, builds on these conceptual approaches to perform replications of interactive prototypes in various contexts of use involving specific user groups, devices, and environments.

To achieve generalizability, a research team runs the same study by employing the original method with a sample of participants drawn from another population to check whether the original findings generalize across populations. For example, Rust *et al.* [25] replicated Wobbrock *et al.*'s [34] original study by eliciting gestures from a different population, represented by children, and Gheran *et al.* [10] replicated a previous study [11] with new participants from the same population, consisting of young adults.

The research area of mobile interaction includes several impactfull GESs, which offer upfront gesture sets for smartphones [16,24], smartwatches [8], smart rings [11], cushions [30], mobile radars [17], and microinteractions through microgestures [6]. Reutilization of previously collected data for a different research goal has been also performed, such as an analysis of gestures elicited in prior work [11] for a different purpose; see details in [12,28]. This distinctive type of a replication was not covered by the ACM [2] or among Tsang and Kwan's [27] replication types, but is useful in the context of gesture elicitation to inform gesture sets that work across devices [8] and contexts of use [31]. For example, Villarreal *et al.*'s [31] survey of gesture elicitation examined the most common gestures and referents used in such studies. Reusing data is relevant for consolidating acquired knowledge since it can lead to new discoveries, complements those from the original study.

3 User Studies

To address our research question, we considered a type-7 reproducibility, *i.e.,* addressing generalizability by employing the same method and new participants [12], to further examine the findings of an original study [11], partially replicated in [10]. To this end, we designed and conducted four new replications.

3.1 Participants

We randomly recruited four samples of thirty (n=30) participants each from internal mailing lists to implement the four GESs, which we denote in the following as Study 1 (15 female, 15 male, M=37.5, SD=15.3 years), Study 2 (16 female, 13 male, M=35.5, SD=17.6 years), Study 3 (12 female, 16 male, M=32.5, SD=12.7 years, one participant lefthanded), and Study 4 (9 female, 21 male, M=33.2, SD=14.0, two participants lefthanded). None of the participants had prior experience with the Ring Zero device, the exact same type as in [11], and all reported moderate to extensive use of desktop computers, smartphones, and tablets. In total, 120 participants were involved. Each study was conducted by a different research team, following the same set of instructions, consistent with the original GES [11].

3.2 Stimuli and Setup

The referents used in our replications were: Turn the TV on/off, Start/stop Player, Turn lights on/off, Turn AC on/off, Turn Alarm on/off, Increase/decrease volume, Go to next/previous item, Answer call/Yes, End call/No, Play music, Pause music, Stop music, Brighten/dim lights, Ask help/assistance, and Turn Heating on-off. Of these, eleven are shared with the original study [11], whereas four—Play music, Pause music, Stop music, and Ask help—are new to collect additional data. The set of referents was the same for all four replications, which were carried out in four different locations simulating a connected IoT home environment. A within-subjects design was implemented across all studies with each participant proposing one gesture per referent, according to the original method [34] and the study [11] chosen for replication.

3.3 Procedure and Task

Participants signed an informed consent form and completed a socio-demographic questionnaire asking about age, gender, handedness, and use of digital devices. Subsequently, the participants were presented with the referents, in randomized order, for which they proposed gestures representing a good fit to the referents, easy to articulate and recall. Each session lasted approximatively 25 minutes. We measured the following dependent variables:

- Agreement-Rate, a ratio variable, expressing the agreement among participants' gestures, calculated with AGATe [29].
- Thinking-Time, a ratio variable, defined as the time in seconds elapsed between the moment when a referent was first presented and the the moment when the participant confirmed having found a suitable gesture to effect it. Just like in the original study [11], a stopwatch was used.
- Goodness-of-Fit, an ordinal variable with values ranging from 1 (low) to 10 (high), indicating the extent to which the participants perceived the gestures they proposed as suitable for effecting the corresponding referents.

3.4 Results

3.4.1 Gesture proposals. Each GES resulted in 30 (participants) × 1 (gesture) × 15 (referents) = 450 gesture proposals, thus totaling 1,800 records across all four studies. Following the analysis of gesture similarity using criteria from [11], we identified twenty-six different gesture classes across all replication studies. For example, Table 1 presents a direct comparison between the consensus gestures. Note how some referents are mapped to the same gesture in multiple studies, such as Increase/decrease volume, confirming previous findings, while other gestures are similar but map to different referents, like Draw letter "V" used to play music in Study 2, while it was proposed to answer Yes in Study 4.

3.4.2 Agreement rate analysis. Figure 2 presents Agreement-Rate values, computed with AGATe [29], for all referents and GESs. Also, Figure 1, left shows the average rates, which were .16 in the original study, .09 in the past replication, and .27, .23, .25, and .17 for in four replications, respectively. The past replication [10] revealed the lowest agreement, the original study [11] and Study 4 are comparable, and Studies 1 to 3 also show similarities with respect to this measure. Since the six GESs do not share the same number of referents exactly, we performed two rounds of statistical testing, using all referents and using the common referents only. Independent-sample t-tests revealed few statistically significant differences, as follows: between the original study and Study 1 (p=.027) with a small effect size (r=.36) across all referents; between Study 1 and Study 4 ($p\leq$.0001, large effect size r=.79) across the common referents; between Study 3 and Study 4 (p=.004, medium effect size r=.63); and between Study 2 and Study 4 (p=.042).

3.4.3 Thinking time analysis. Figure 3, left shows the average Thinking-Time values measured for individual referents in the six GESs, and Figure 1, right shows averages computed across all GESs. Significant variations were observed for the different replication studies. For example, the Turn TV on/off referent had the shortest thinking time (4.2s) in the original study [11] and the longest (24.7s) in Study 4. Overall, the past replication [10] and the four new GESs revealed different average thinking times, with larger values in Study 3 and Study 4. A Kruskal-Wallis test detected a statistically significant effect of study, and Conover-Iman [7] post-hoc tests various significant differences across individual pairs of referents. In most cases, the original GES [11] and the past replication [10] were significantly different from Studies 1 to 4. Study 2 also presented significant differences in participants' thinking times with respect to Studies 1, 3, and 4, which could be explained by variation in population sampling. A subsequent Kruskal-Wallis test ($H_{(5)}$=500.84, $p\leq$.0001) indicated that the average Thinking-Time of only the common referents was significantly different across the six studies. When significant differences were detected, Cohen effect sizes were small to medium with only the difference between the past replication [10] and Study 4 revealing a large effect size.

3.4.4 Goodness of fit analysis. Goodness-of-Fit was self-reported by participants as a rating between 1 (low) and 10 (high) regarding how well the proposed gestures matched the referents. Figure 3, right shows the results for all referents and studies, and Figure 1, middle presents averages across studies. A Kruskal-Wallis test indicated that Goodness-of-Fit ratings were statistically different across studies. Results of the Conover-Iman post-hoc tests are shown in Figure 3, right with their significance level overimposed on the charts but, overall, few statistically significant differences emerged. For example, a significant difference was detected between Study 1 and Study 2 for the Go to previous item referent. A subsequent Kruskal-Wallis test ($H_{(5)}$=99.21, $p\leq$.0001) indicated that the Goodness-of-Fit of common referents varied significantly across studies. However, the average ratings remained close, from 7.01 for Study 1 to 8.20 for the past replication [10].

3.4.5 Discussion. Returning to our research question, introduced in Section 1, we found that the various dependent variables presented significant variations across the replication studies, including at referent level. However, the practical differences were small, *e.g.,* the average Agreement-Rate values varied between .09 and .27, falling within the medium range [29]. Similarly, the average Goodness-of-Fit ratings ranged from 7.01 to 8.20, suggesting that all participants, both in the original study and the replications, were satisfied with their gesture-to-function mappings. From the perspective of these two measures, the variation in research team had little practical effect on the results. Thinking-Time varied considerably more across the studies, from an average of 3.2s to 18.5s—however, this measure is less critical compared to the previous ones. In light of these findings, the crucial question appears not to be the time it takes participants to think about suitable gestures so that researchers can reproduce findings on user-defined gestures, but whether participants agree on the same gesture types and to what extent. These aspects are key to designing consensus gesture sets that have high generalizability and transferability across interactive devices [8], use groups [10,25], and contexts of use [31].

Table 1: Examples of gestures reaching consensus among participants' proposals elicited in our replication studies.

Referent	Study 1	Study 2	Study 3	Study 4
Turn TV on/off	Press a button	Press a button on a remote control	Swipe one finger up/down	Finger forward
Start/stop player	Bend a finger	Press an imaginary button in mid-air	Swipe two fingers up/down	Finger forward, back and forward
Turn lights on/off	Close fingers	Snap fingers	Draw letter "L" in mid-air	Open/Close fist
Increase volume	Turn three fingers clockwise	Swipe up	Swipe clockwise	Draw hairspring clockwise
Decrease volume	Turn three fingers counterclockwise	Swipe down	Swipe counterclockwise	Draw hairspring counterclockwise
Play music	Draw letter "M" in mid-air	Draw letter "V" in mid-air	Swipe up	Draw triangle
Pause music	Press a button	Forward gesture with outstretched palm	Swipe down	Draw inverse of letter "N"
Stop music	Press a button	Forward gesture with outstretched palm	Swipe forward and back	Draw a square
Go to next item	Swipe right	Swipe right	Draw a circle clockwise	Swipe right
Go to previous item	Swipe left	Swipe left	Draw a circle counterclockwise	Swipe left
Answer call/Yes	Draw a check in mid-air	Thumbs up	Draw letter "Y" in mid-air	Draw letter "V" in mid-air
End call/No	Swipe up	Thumbs down	Draw letter "N" in mid-air	Draw letter "X" in mid-air

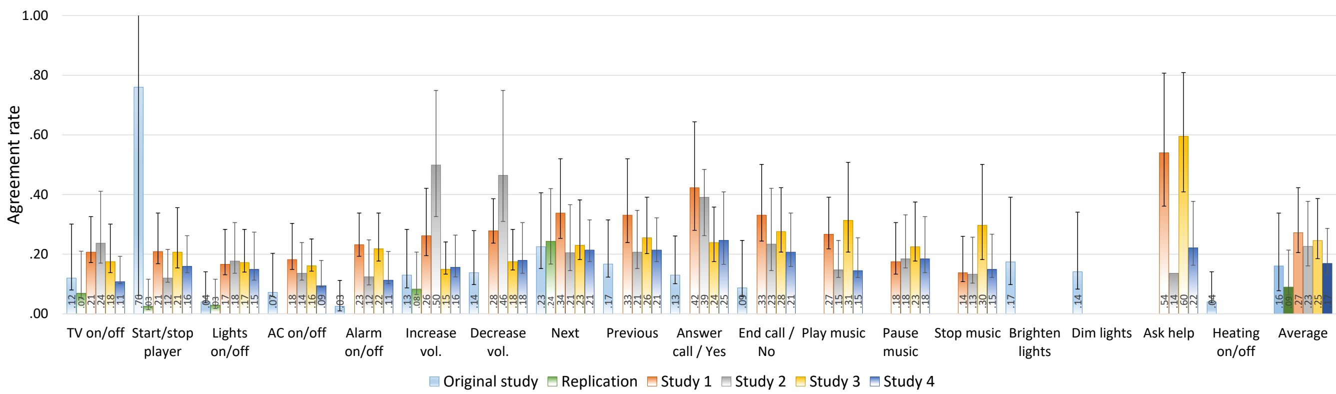

Figure 2: Comparison of Agreement-Rate values, all referents considered, from our four replications against the original study [11] and a past replication [10]. ***Note:*** **error bars show 95% confidence intervals.**

4 Conclusion

To ensure the completeness, representativeness, and generalizability of user-defined gesture sets for mobile and wearable interaction, the scientific community must treat replications [12,28] not as optional but integral to end-user gesture elicitation, since different replications may produce significantly different results depending on the evaluation measure. Our findings indicate that agreement rate, goodness of fit, and thinking time can vary across referents and studies. Future work should address more variations in the experimental settings, including scenarios involving the user seated *vs.* walking or traveling, located inside [4] or outside [23] in smart buildings *vs.* outdoor environments of various kinds, as well as when engaging with gesture interaction in front of various audiences [24] in a public context for increased social acceptability [15]. Also, other evaluation measures could be considered for comparing findings about user-defined gestures across different replications, besides those used in our study, such as coagreement rates [29] computed between referents, settings, and user groups. The reliability of evaluating these measures across different research teams should also be taken into account in more depth, *e.g.,* thinking times were measured in our studies with a stopwatch, just like in the original study [11], but actual usage of the stopwatch with respect to the start and stop moments may have introduced variations across the teams. Such specific aspects can be addressed in future work with replication types further informed by dedicated frameworks, such as RepliGES [12] and Ampliatum [28]. With repeated validation in context, replication studies can support the generalizability and transferability envisioned for user-defined gestures across diverse mobile and wearable interaction scenarios for consistent user performance and experience. We still need to address the question of how many studies to replicate and how many participants to gather to obtain a set of gestures that maximize coverage and minimize variation, bearing in mind that the same referent may or may not give rise to the same consensus gesture across different studies. We need to define a cumulative elicitation study procedure that considers repeated gestures from one replication to the next.

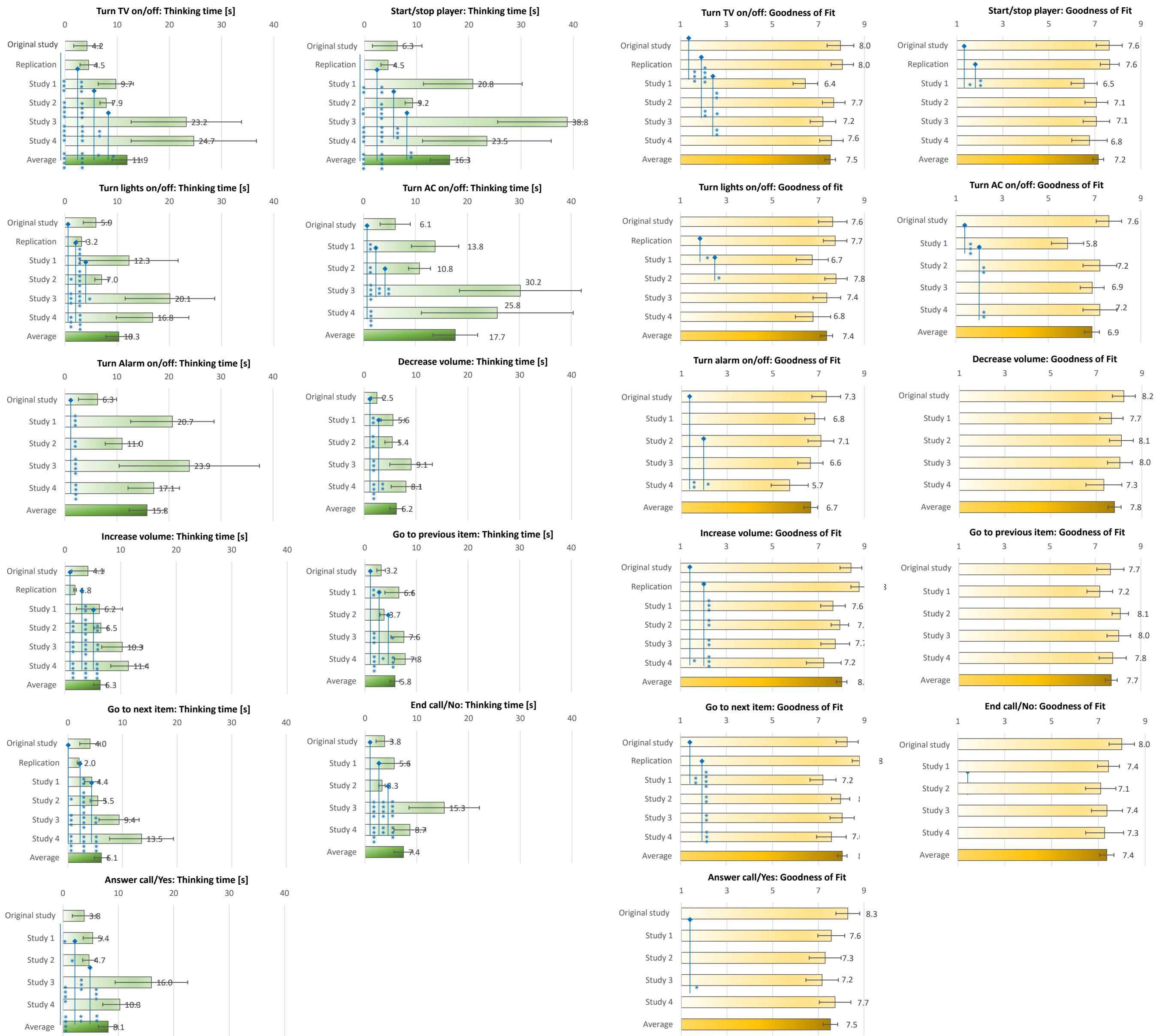

Figure 3: Comparison of Thinking-Time (left) and Goodness-of-Fit (right) average values of common referents in our four studies against the original study [11] and a past replication [10]. *Note:* error bars show 95% confidence intervals. The level of significance is defined as: $p \leq .05^{*}$, $p \leq .01^{**}$, **and** $p \leq .001^{***}$**.**

Open Science.

We share a public repository with the data from our studies. The spreadsheets contain the raw data collected from all studies, the original stydu, its first replication, and the four replications discussed in this paper, as well as the statistical tests along with their details. The charts reproduced in this paper are also included.

Acknowledgments

Jean Vanderdonckt is supported by the EU EIC Pathfinder-Awareness Inside challenge Symbiotik under Grant no. 101071147. Radu-Daniel Vatavu acknowledges support from the NetZeRoCities Competence Center funded by the European Union-NextGenerationEU and the Romanian Government under the National Recovery and Resilience Plan for Romania, contract no. 760007/30.12.2022 with the Romanian Ministry of Research, Innovation, and Digitalization through the specific research project P3, Smart and sustainable buildings.

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] 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 (UIST '18)*. ACM, New York, NY, USA, 177–188. doi:10.1145/3242587.3242621

[4] Wael Alsafery, Omer Rana, and Charith Perera. 2023. Sensing within Smart Buildings: A Survey. *ACM Comput. Surv.* 55, 13s, Article 297 (July 2023), 35 pages. doi:10.1145/3596600

[5] Michael Chamunorwa, Mikolaj P. Wozniak, Susanna Krämer, Heiko Müller, and Susanne Boll. 2023. An Empirical Comparison of Moderated and Unmoderated Gesture Elicitation Studies on Soft Surfaces and Objects for Smart Home Control. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 198 (Sept. 2023), 24 pages. doi:10.1145/3604245

[6] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User Elicitation on Single-hand Microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3403–3414. doi:10.1145/2858036.2858589

[7] W.J. Conover and Ronald N. Iman. 1979. *On Multiple-Comparisons Procedures*. Informal Report LA-7677-MS. University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico, USA. https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-07677-MS

[8] 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 (CHI '18)*. ACM, New York, NY, USA, 1–12. doi:10.1145/3173574.3173993

[9] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woundefinedniak. 2019. Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 455, 14 pages. doi:10.1145/3290605.3300685

[10] Bogdan-Florin Gheran, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2023. New Insights into User-Defined Smart Ring Gestures with Implications for Gesture Elicitation Studies. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 216, 8 pages. doi:10.1145/3544549.3585590

[11] 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 (DIS '18)*. ACM, New York, NY, USA, 623–635. doi:10.1145/3196709.3196741

[12] 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 ACM International Conference on Advanced Visual Interfaces (AVI '22)*. ACM, New York, NY, USA, Article 5, 9 pages. doi:10.1145/3531073.3531112

[13] Christian Greiffenhagen and Stuart Reeves. 2013. Is Replication Important for HCI?. In *Proc. of the CHI '13 Workshop on the Replication of HCI Research*. CEUR-WS, Paris, France, 1–6. http://ceur-ws.org/Vol-976

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

[15] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. 2020. Social Acceptability in HCI: A Survey of Methods, Measures, and Design Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–19. doi:10.1145/3313831.3376162

[16] 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1303–1312. doi:10.1145/1978942.1979136

[17] 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. doi:10.1109/ACCESS.2019.2951349

[18] Nathan Magrofuoco and Jean Vanderdonckt. 2019. Gelicit: A Cloud Platform for Distributed Gesture Elicitation Studies. *Proc. ACM Hum.-Comput. Interact.* 3, EICS, Article 6 (June 2019), 41 pages. doi:10.1145/3331148

[19] 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:10.1145/2591689

[20] Manish Parashar. 2019. The Reproducibility Initiative. *IEEE Computer* 52, 11 (2019), 7–8. doi:10.1109/MC.2019.2935265

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

[22] Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 887–896. doi:10.1145/1753326.1753458

[23] Isabel B. 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, USA, 32–41. https://doi.org/10.1145/3132272.3134118

[24] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-Defined Motion Gestures for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 197–206. doi:10.1145/1978942.1978971

[25] Karen Rust, Meethu Malu, Lisa Anthony, and Leah Findlater. 2014. Understanding Childdefined Gestures and Children's Mental Models for Touchscreen Tabletop Interaction. In *Proceedings of the 2014 Conference on Interaction Design and Children (IDC '14)*. ACM, New York, NY, USA, 201–204. doi:10.1145/2593968.2610452

[26] Carmel Shavitt, Anastasia Kuzminykh, Itay Ridel, and Jessica R. Cauchard. 2021. Naturally Together: A Systematic Approach for Multi-User Interaction With Natural Interfaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 349 (Oct. 2021), 31 pages. doi:10.1145/3476090

[27] 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 (1999), 759–780. http://dx.doi.org/10.2307/259353

[28] Jean Vanderdonckt and Radu-Daniel Vatavu. 2025. Context is Key for Reproducibility of Empirical Studies in Human-Computer Interaction. In *Proceedings of the ACM Conference on Reproducibility and Replicability (REP '25)*. ACM, New York, NY, USA, 10 pages. doi:10.1145/3736731.3746142

[29] 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 ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1325–1334. doi:10.1145/2702123.2702223

[30] Santiago Villarreal-Narvaez, Arthur Sluÿters, Jean Vanderdonckt, and Efrem Mbaki Luzayisu. 2022. Theoretically-Defined vs. User-Defined Squeeze Gestures. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 559 (Nov. 2022), 30 pages. doi:10.1145/3567805

[31] Santiago Villarreal-Narvaez, Arthur Sluÿters, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2024. Brave New GES World: A Systematic Literature Review of Gestures and Referents in Gesture Elicitation Studies. *ACM Comput. Surv.* 56, 5, Article 128 (Jan. 2024), 55 pages. doi:10.1145/3636458

[32] 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 (DIS '20)*. ACM, New York, NY, USA, 855–872. doi:10.1145/3357236.3395511

[33] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: From a Panel to a New Submission Venue for Replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. ACM, New York, NY, USA, 1185–1188. doi:10.1145/2212776.2212419

[34] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1083–1092. doi:10.1145/1518701.1518866

[35] Haijun Xia, Michael Glueck, Michelle Annett, Michael Wang, and Daniel Wigdor. 2022. Iteratively Designing Gesture Vocabularies: A Survey and Analysis of Best Practices in the HCI Literature. *ACM Trans. Comput. Hum. Interact.* 29, 4 (2022), 37:1–37:54. doi:10.1145/3503537