Collecting Big Data Through Citizen Science: Gamification and Game-based Approaches to Data Collection in Applied Linguistics

Yoolim Kim, Vita V Kogan, Cong Zhang

Cite as:

Yoolim Kim and others, Collecting Big Data Through Citizen Science: Gamification and Game-based Approaches to Data Collection in Applied Linguistics, *Applied Linguistics*, 2023;, amad039, https://doi.org/10.1093/applin/amad039

Abstract

Gamification of behavioral experiments has been applied successfully to research in a number of disciplines, including linguistics. We believe that these methods have been underutilized in applied linguistics, in particular second-language acquisition research. The incorporation of games and gaming elements (gamification) in behavioral experiments has been shown to mitigate many of the practical constraints characteristic of lab settings, such as limited recruitment or only achieving small-scale data. However, such constraints are no longer an issue with gamified and game-based experiments, and as a result, data collection can occur remotely with greater ease and on a much wider scale, yielding data that are ecologically valid and robust. These methods enable the collection of data that are comparable in quality to the data collected in more traditional settings while engaging far more diverse participants with different language backgrounds that are more representative of the greater population. We highlight three successful applications of using games and gamification with applied linguistic experiments to illustrate the effectiveness of such approaches in a greater effort to invite other applied linguists to do the same.

INTRODUCTION

Applied linguistics is a highly interdisciplinary field that combines methodologies from linguistics, psychology, cognitive science, neuroscience, and language pedagogy. With many advantages that such interdisciplinarity grants, studies in applied linguistics have a number of limitations that often curtails the scope of research questions and results. First, most studies in the field focus largely on the acquisition of English as a first language (L1) and/or second language (L2), which limits our understanding of how typologically diverse (non-English) languages are acquired (e.g. languages with far richer morphology). Second, small participant samples that consist primarily of undergraduate student populations limit both reliability and generalizability of the study. Finally, most research is conducted in laboratory settings and lasts for an hour (at best) resulting in a set of data that does not reflect the complexity of the acquisition process and fails to capture its dynamic nature that can only be truly observed along a more longitudinal dimension. It would thus be beneficial for the field to expand the repertoire of available methodology to include experiments that involve greater linguistic diversity, a more representative sample of participants, and more sophisticated environments. A recent trend in the use of games and gamification in behavioral experiments (see Long et al. 2023 for a review of recent endeavors) might be a fruitful solution in addressing existing shortcomings in applied linguistics research.

Gamification involves using game mechanics such as competition and rewards in non-game environments (Deterding 2011; Werbach 2014), while game-based approaches refer to the design and

usage of standalone games or the repurposing of existing games. Both benefit data collection in applied linguistics because (i) we can engage everyday citizens in scientific inquiry to (ii) capture more ecologically valid data comparable to what can be obtained in controlled experimental settings, to a much larger scale, via remote means, (iii) making our observations more meaningful and robust and not necessarily an artifact of limited sampling of similar participants with similar linguistic and demographic profiles. Both approaches have been successful in other domains (Zichermann and Linder 2010; Iliev-Piselli et al. 2011; Caminal 2012), including those within linguistics (Cornish et al. 2013; Hartshorne et al. 2018; Leeman et al. 2019; Arvaniti et al. 2022), but they have yet to be fully utilized in applied linguistics. Long et al. (2023) note how games can improve the quality of data in behavioral science by diversifying the reach of participants beyond WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations and those comprising undergraduate student populations commonly sampled. Building on this momentum, we invite linguists to consider how games and gamification can enrich applied linguistic research, the benefits of which we believe would be enabling large-scale 'big' data that are robust and representative of the broader population of speakers and learners.

So far, the only use of gamification or games (to the best of our knowledge) within applied linguistics has been with research on children (e.g. Zwitserlood et al. 2022), but has yet to be further exploited. Our broad objective is to reverse this, and in doing so, call upon applied linguists to employ games and gamification more extensively. To illustrate the success of this experimental methodology, we introduce three cases of using gamification and games that can address the issues outlined above as they pertain to second-language (L2) acquisition, one of which is a proposed design for collecting speech perception data on a large scale that has yet to be properly achieved. We first discuss a case of using a standalone game to obtain a more naturalistic speaking sample and to reduce participatory anxiety. We also discuss the online applet 'Glyph' as a successful case of gamification of an experimental task to investigate writing systems, which has resulted in big data from tens of thousands of participants. Finally, our third case features gamification applied to socio-phonetic research, designed to showcase linguistic diversity. These three cases exemplify the ways in which the use of games and gamification have produced large-scale data comprising a diversified participant pool that extends beyond a traditional laboratory setting; such is seldom observed in applied linguistics but can be easily implemented in a similar fashion to yield comparable results.

TABLETOP GAMES IN L2 ACQUISITION RESEARCH

Although collecting data in the lab is considered the gold standard in applied linguistics, it has its limitations. Participants who are usually language learners often act differently when completing experimental tasks in a foreign language: these differences might manifest as performance anxiety, limited or forced speaking output, or functioning at a proficiency level lower than what they demonstrate in class (Guiora et al. 1969; Ozdemir and Papi 2022). As a result, the collected data at times do not capture the measured abilities accurately. Games are widely praised for mitigating the effects of anxiety (Hwang et al. 2017), making them an ideal environment for spontaneous speech elicitation and more naturalistic data collection. In one study, Kogan et al. (in preparation) used a commercial roll-and-move tabletop game, similar to 'Monopoly', to obtain speech samples from 27 lower-intermediate/B1 learners of L2 Russian. Participants played a game in triads for 30 min and explained their moves and decisions aloud in Russian as they progressed through the game (Figure 1).

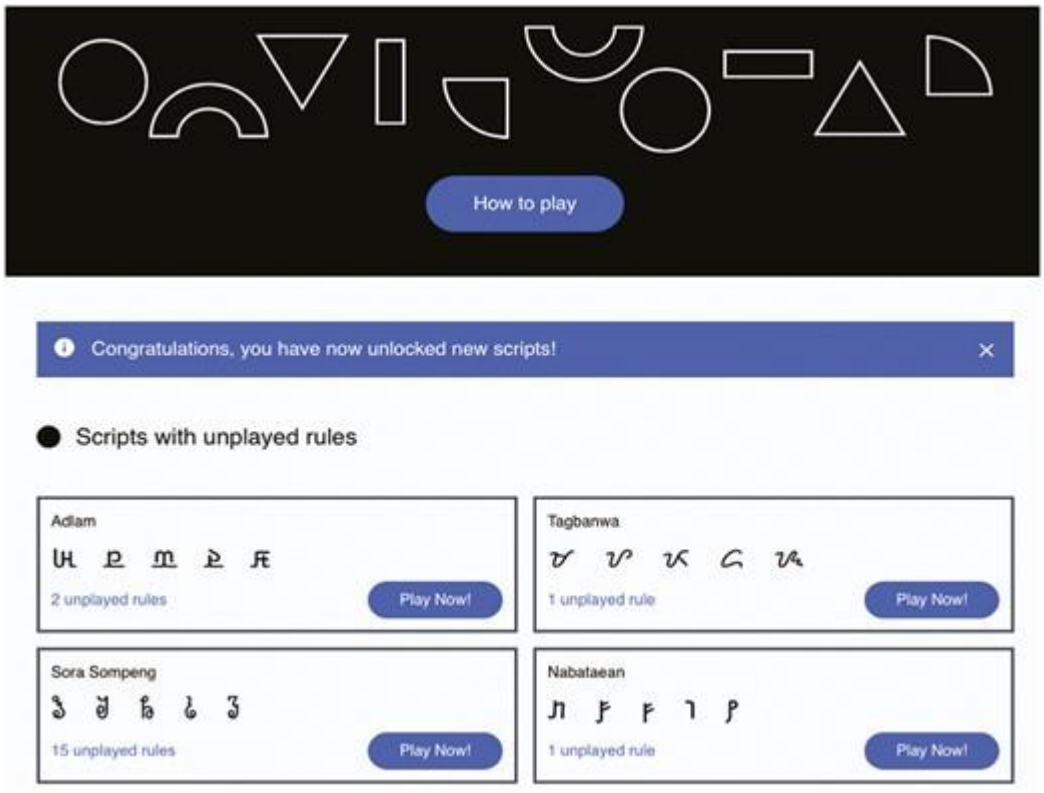
Figure 1: Participants (including a researcher) play LinguaPolis, a tabletop roll-and-play game with mechanics similar to Monopoly.

The recorded speech was analyzed for the use of verbs of motion—a challenging aspect of Russian morphosyntax, often avoided by even advanced learners of Russian. Compared to data generated from 19 participants who completed a typical speech elicitation task (picture-based storytelling), rather than playing the game, the data elicited from players produced significantly more verbs of motion per session. In a follow-up, participants who played the game reported feeling relaxed and engaged during the experiment, whereas the non-players described the task as being ‘difficult’ and ‘boring’. Kogan et al. show that using games as a methodological framework for data collection alleviates participatory anxiety and provides more ecologically valid L2 data.

ONLINE APPLET AND ENGAGING CITIZEN SCIENTISTS FOR WRITING SYSTEMS RESEARCH

Glyph is an online gaming applet designed to investigate the shapes of letters. The underlying research interest is to better understand how letter shapes evolve negotiating constraints of distinctiveness and informativeness. Players have a choice of 45 different scripts with which to play (Figure 2). The scripts are selected to maximize diversity in a language family, region of use, and structural properties, and are intended to be those with which people have less familiarity. Players are asked to classify and sort the letters of a script into two groups according to a rule of their choices, for example, 'All characters that contain at least one enclosed space'. If players are able to successfully replicate their classification after a short time-lapse, they win points; they win as many points as the number of characters they classified. The game has no definite end, as players are welcome to create as many different rules for each of the different scripts as their creativity allows. Players' standings can be viewed on a global leaderboard, and additional points can be earned if players are the first to propose a rule.

Figure 2: The main page of Glyph that displays a short tutorial on how to play with abstract shapes and sampling of available scripts.



The resulting data can help to illuminate the field of learning a L2 orthography. If the L1 is written using an alphabetic system, then are there significant burdens for learning L2 with a structurally different (non-alphabetic) script? Understanding what visual properties people are sensitive to and building a visual morphology of writing systems may aid in this research. Data collection is ongoing and has so far generated participation from tens of thousands of users from a diverse age range, all around the world, with varying degrees of proficiency in spoken and written languages.

ONGOING RESEARCH IN L2 PHONETICS AND PHONOLOGY

Data collection for phonetics and phonology can be lengthy and tedious as the tasks usually involve hearing or producing many tokens of speech sounds. Gamifying the research process can benefit speech perception and production data collection by increasing data size and improving data spontaneity, which are critical aspects in the field of phonetics and phonology. Currently, we are gamifying socio-phonetic and L2 learning research on accent identification. We designed a game on an online experiment platform, Labvanced. Integrating a gamified task into an existing experiment platform is much more accessible and affordable, particularly compared to outsourcing developers. The participants can also use their own devices to play, enabling remote data collection. In the game, participants provide demographic information as well as language background, and then guess the accent of English they hear. It first starts with by-country identification (i.e. which country is the accent from?), but as the levels progress, the accents become specific to increasingly smaller regions of the country, such as counties and towns. The game will feature a leaderboard that is updated daily as well as different levels of varying difficulty. Modeling identification rate in relation to various demographic factors reveals key information about the cues listeners rely on for accent identification. By engaging with the game, participants are exposed to a variety of different accents with varying degrees of familiarity. Players reflect on the demographic information they associate with each accent that can enrich and challenge existing attitudes toward accents, and in the process, illustrate their great diversity.

POSSIBLE CAVEATS AND FURTHER CONSIDERATIONS

Despite the numerous benefits of gamified and game-based experiments, there are some caveats to consider, which we present here. Because of the entertaining nature of gaming, some participants might not approach a gamified task with as much seriousness as they otherwise would in a laboratory setting (Annamalai et al. 2021). Gamified experiments tend also to be administered online, which might exacerbate the issue for two reasons: (i) as online platforms tend to vet and recruit participants, it limits the information researchers have about the participants and results in a less nuanced understanding of the participant demographic, and (ii) there are generally fewer opportunities to interact, observe, and supervise how participants completed the experiment. One possible way to mitigate such issues concerning whether participants are paying attention is to introduce ‘attention check questions’ (e.g. what was the color of the previous item? or what was the animal in the top corner of the screen?). Platforms like PsyToolkit (Stoet 2017) allow researchers to adjust certain experiment settings to limit in part how participants engage and interact with the study. For example, participants can be barred from completing the study on a mobile phone, which can help to improve the validity of the data. Platforms such as CloudResearch and Prolific are rigorous about data quality and ban dishonest participants from their database (Peer et al. 2022).

It is also important to consider the ethical aspects of this methodology. Utilizing user-friendly consent interfaces that clearly explain the implications of data collection and provide individuals with meaningful choices and control over their data is crucial. Online data collection also involves promoting transparency, providing clear and easily understandable privacy policies and terms of service, and employing appropriate technologies to protect users, all of which may also place additional burdens for researchers when planning. Another consideration is that not all populations of participants will be able to access online and digital gamified experiments due to significant disparities in access to technology and the Internet across different regions, socio-economic groups, and demographics. Certain populations, such as marginalized communities, the elderly, people with disabilities, or individuals with limited digital literacy, may face barriers to participating in data collection activities. Sometimes it is possible to overcome these difficulties by employing targeted

outreach and partnerships with community organizations. In-person events can also mitigate these issues; researchers can hold public events with laptops, tablets, and other such devices, where members of the community can try the gamified studies or applets in a supervised setting, engaging directly with the researchers themselves. Gamified studies can also be designed specifically to target certain communities; Braun et al. (2021), for example, designed an app specifically for rural populations where dialectal varieties of German are used. Such varieties are difficult to test in urban settings where laboratories tend to be based, so gamification and online data collection offer viable solutions.

Finally, some participants might feel infantilized when asked to play a game. Explaining the rationale and purpose of the research study, articulating clearly to participants their role as citizen scientists, and making clear their contribution to the research process, would reduce feelings of infantilization. Participants are then more likely to view the gamified elements as enhancing the experience of data collection, rather than trivializing, and in the process, also feel empowered as being part of the scientific process.

CONCLUSION

In this contribution, we have highlighted the positive ways in which gamification and games have changed the quality and scale of data in language research. We believe that research in applied linguistics has yet to fully utilize such methods and stands to benefit from them as well. We have offered three cases where such impact is exemplified: within the contexts of L2 acquisition, whether it is learning a new symbol mapping in a writing system or the phonemes of a spoken language. We believe that greater use of games and gaming elements to applied linguistics will allow for more robust data collection yielding big data that are more representative of speaker and learner communities while engaging the everyday citizen.

NOTES AND ACKNOWLEDGEMENTS

 This project was partially supported by the Institute for Social Science, Newcastle University through a Pioneer Award to CZ.

REFERENCES

Annamalai, N., Kabilan, M. K., Rashid, R. A., Oleskevicience, G. V., & Vaičiūnienė, V. 2021. ‘English language learning through non-technology games: A case study of international students at a Lithuanian university.’ *The Qualitative Report*, 26/10: 3261–3278.

Arvaniti, A., Gryllia, S., Zhang, C., & Marcoux, K. P. 2022. ‘Disentangling emphasis from pragmatic contrastivity in the English H*~ L+ H* contrast,’ *Proceedings of Speech Prosody 2022*: 837-841.

Caminal, R. 2012. ‘The design and efficiency of loyalty rewards,’ *Journal of Economics & Management Strategy*, 21/2: 339–371.

Cornish, H., Smith, K., & Kirby, S. 2013. ‘Systems from sequences: An iterated learning account of the emergence of systematic structure in a non-linguistic task.’ *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35/35.

Deterding, S., Dixon, D., Khaled, R., & Nacke E. L. 2011. 'Gamification: Toward a definition,' *CHI 2011*.

Granic, I., Lobel, A., & Engels, R. C. 2014. 'The benefits of playing video games,' *American Psychologist*, 69/1: 66.

Guiora, A. Z., Taylor, L. C., & Brandwin, M. A. 1969. 'A contribution to the psychology of second language behavior,' *Proceedings of the XVlth International Congress of Applied Psychology*.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. 2018. 'A critical period for second language acquisition: Evidence from 2/3 million English speakers,' *Cognition*, 177: 263–277.

Hwang, G. J., Hsu, T. C., Lai, C. L., & Hsueh, C. J. 2017. 'Interaction of problem-based gaming and learning anxiety in language students' English listening performance and progressive behavioral patterns,' *Computers & Education*, *106*: 26–42.

Iliev-Piselli, M. M., Fadjo, C. L., & Lee, J. J. 2011. 'Bank-It: A mobile financial literacy game,' *Proceedings of the 7th International Conference on Games Learning Society Conference*, 260–262.

Klopfer, E., Osterweil, S., & Salen, K. 2009. *Moving Learning Games Forward*. The Education Arcade.

Kogan, V. V., Bondarenko, M., & Kapustin, S. 2023. The acquisition of Russian verbs of motion with a task-based language teaching approach. *Manuscript in preparation*.

Leeman, A., Derungs, C., & Elspaß, S. 2019. 'Analyzing linguistic variation and change using gamification web apps: The case of German-speaking Europe,' *PLoS ONE,* 14/12: eCollection 2019.

Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. 2023. 'How games can make behavioral science better,' *Nature,* 613: 433–436.

Ozdemir, E., & Papi, M. 2022. 'Mindsets as sources of L2 speaking anxiety and self-confidence: The case of international teaching assistants in the US,' I*nnovation in Language Learning and Teaching*, 16/3: 234-248.

Peer, E., Rothschild, D., Gordon, A., & Damer, E. 2022. 'Erratum to Peer et al. (2021) Data quality of platforms and panels for online behavioral research.' *Behavior Research Methods*, 54/5: 2618-2620.

Ryan, R. M., Rigby, C. S., & Przybylski, A. 2006. 'The motivational pull of video games: A self-determination theory approach,' *Motivation and Emotion*, 30: 344–360.

Stoet, G. 2017. 'PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments'. *Teaching of Psychology,* 44/1: 24-31.

Werbach, K. 2014. '(Re)Defining gamification: A process approach,' *Persuasive Technology,* 8462: 266–27.

Zichermann, G., & Linder, J. 2010. *Game-Based Marketing: Inspire Customer Loyalty Through Rewards, Challenges, and Contests.* John Wiley & Sons.

Zwitserlood, R., Harmsel, M. T., Schulting, J., Wiefferink, K., & Gerrits, E. 2022. 'To game or not to game? Efficacy of using tablet games in vocabulary intervention for children with DLD.'*Applied Sciences*, 12/3: 1643.