Journal of Language Evolution, 2025, **10**, lzaf006
https://doi.org/10.1093/jole/lzaf006
Advance access publication 1 December 2025
Research Article

Get data early, get data often, iterate constantly: an explorative, participatory approach to studying language evolution

Seán G. Roberts[1,*], Kateryna Krykoniuk[1,2], Mark Brown[3], Fiona M. Jordan[4]

[1]School of English, Communication and Philosophy, Cardiff University, Cardiff, United Kingdom
[2]School of Languages, Arts and Societies, University of Sheffield, Sheffield, United Kingdom
[3]Game Maker's Toolkit, United Kingdom
[4]Department of Anthropology and Archaeology, University of Bristol, Bristol, United Kingdom

*Corresponding author. School of English, Communication and Philosophy, Cardiff University, John Percival Building, Cardiff CF10 3EU, UK. E-mail: robertss55@cardiff.ac.uk

Associate Editor: Dan Dediu

We argue that, as well as an empirical approach borrowed from experimental psychology, studies of language evolution can also benefit from an explorative, participatory approach. This is based on a reflection on an experimental semiotics study where the process of arriving at an effective experimental design was equally valuable for developing the theory as the final results of the experiment. We suspect that this process is commonplace in many studies, but there is no formal method for documenting or exploiting any insights gained. We present methods from video game design and ethnography as candidates for addressing this gap and suggest they can be used in a hybrid approach that combines an exploratory phase of cyclic iteration with a final, more traditional linear phase. We illustrate these methods with two case studies and argue that a participatory approach can harness the creative power of our participants and help us reveal important aspects of our theories.

Keywords: experimental semiotics; video game design; exploratory research; thick description.

1. Introduction

How do researchers arrive at key insights for developing a theory? Recent approaches to language evolution using empirical methods have used a psychology paradigm, running experiments with human participants where the input to the process of interest is manipulated in order to test some differences in output. This process is followed by most 'experimental semiotics' studies into language evolution (Roberts 2017; Nölle and Galantucci 2022). This empirical turn has been productive for the field, particularly for studies of the cultural evolution of linguistic features (Tamariz and Kirby 2016; Müller and Raviv 2025). It has allowed researchers to make significant progress on developing hypotheses, engage with research across disciplines, and distance themselves from historical associations with untestable 'just so' stories (Pinker and Bloom 1990: 714, 715). However, in this paper, we argue that some of the central questions about the origins of language have been sidelined because of the uncritical adoption of psychology methods that may not always fit the aims of language evolution studies.

The issue lies in critical differences between the aims and research methods of language evolution and psychology (perhaps related to the relative age of the fields). In most psychology experiments, the object of research interest (processes in modern human brains) is also the object in the experiment (a participant's brain). This is often not the case for language evolution experiments, where the object of study is often pre-linguistic humans and the object in the experiment is a modern human who already speaks a language. Furthermore, the aim of language evolution experiments is not necessarily to obtain evidence to support a prediction but to find problems with the theoretical assumptions.

We suggest that a missing ingredient for an empirical study of the origins of language is an explicit embracing

© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

of explorative, iterative design. It is likely that elements of this approach are already used by researchers and indeed are similar to the typical practice of 'piloting' in psychology experiments. However, the valuable aspect of a psychology study is often presented as being the final results, not the journey of getting to the design. While the final results are also valuable in language evolution studies as evidence of an effective manipulation, we suggest that important insights may be gained in the process that led the researchers to arrive at the final design. This suggests that more transparent communication about the journey to the design would be productive.

However, there is currently no standard method for conducting and documenting this exploratory, iterative process. Therefore, in the vein of ethnographic work on 'laboratory life' (Latour and Woolgar 1979), we reflect on the actual processes that language evolution researchers follow. From this, we suggest that one solution lies in embracing the iterative design process of video game design alongside the ethnographic method of 'thick description' (Ryle 1968; Geertz 1973) to harness the creative power of ordinary participants against the confirmation bias of the researcher. We suggest this approach would be useful for researchers taking a synthetic approach to investigating general principles of language emergence, as in experimental semiotics (see below), rather than studies of non-human animal communication, typological studies of language change, or genetic or neuroscientific studies of the cognitive basis for language.

We lay out our argument as follows. In the first section, we contrast the way that empirical language evolution studies are presented with a more iterative process that seems to occur in reality, revealing a methodological gap. We then review exploratory methods in other fields and argue that iterative 'playtesting' in video game design may have applications in language evolution research. We illustrate this with two case studies, arguing that thick description from anthropology is another missing piece of the puzzle. Finally, we explain the advantages of a participatory approach. Throughout the paper, we make it clear that we are calling for more transparency, not less rigour.

2. Standard approaches to empirically informed language evolution studies

This section discusses the 'ideal' approach to language evolution studies and differences with the practical process that researchers actually use. To illustrate these points, we use a case study of a (currently unpublished) paper we wrote, which we will call 'PaperX'.

PaperX focuses on the 'problem of motivation' (Hurford 2007: 131)—identifying the key properties of an environment that motivates the emergence of complex communication systems. The authors conduct a standard 'experimental semiotics' experiment to explore how symbolic referential signals emerge. While there are many theories of the evolution of semanticity from anthropology and other fields, there are few ways of testing the theories against each other. The solution in PaperX is to run an experiment where modern human participants had to complete a task in a virtual world, in this case the video game *Minecraft* (though the use of a video game to conduct the experiment is incidental to our main argument about adopting video game design methodologies). Participants were seated at computers in the same room but could not see each other due to a divider placed between them. Participants were given a task to complete within the virtual world of Minecraft. For example, one condition gave them 20 minutes to work together to build a structure out of different coloured blocks (Fig. 1). Each participant had half the plan they needed to follow, and each had access to different block colours, requiring them to communicate to each other the position and colour of blocks. The participants were prevented from communicating using modern language by an experiment instructor, who informed them that they were not allowed to speak. Instead, they could communicate by knocking on the table or using gestures through their avatars in the virtual world. Thus, participants had the opportunity to develop a symbolic referential system during the experiment (e.g. one knock for a red block and two knocks for a green block).

However, the authors discovered that a task based around collaborative building did not motivate participants to create a symbolic referential language, as they were able to rely on trial-and-error or simple pointing gestures. For example, they could raise their avatar's arm and direct it towards a specific point in space or rotate their head in a desired direction and stare at the target area they wanted their partner to focus on (Fig. 2). In contrast, participants frequently created a symbolic referential language when given a different task based around maintaining a fire. This condition was designed so that the fuel they needed to refer to was distant from the participants at the time they needed to refer to it, and there was no iconic object in their immediate environment to represent the concept of fuel.

The write-up of PaperX is structured according to an ideal series of logical steps, which we will call the 'linear approach':

- Identify the relevant theories about the target phenomenon
- Identify the causal claims in the theories and generate a testable hypothesis (using tools from causal inference)

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Figure 1. The testing ground for the building experiment, as seen from the first participant's perspective, is enclosed by a fence, beyond which a mountain rises in the distance. The second participant (labelled 'ExperPlayer3') observes the completed structure made from colourful blocks.

- Design an experiment to test the hypothesis using principles from experimental semiotics and a common task framework
- Examine the results of the experiment to determine whether the hypothesis is supported or rejected
- Use the experimental results to support the core causal components of the theory (a robustness approach, Irvine et al. 2013)

The process is visualized in Fig. 3, which is based on the traditional 'waterfall' model of software engineering (Sommerville 2011: 30), which conceptualizes each stage of the process as self-contained and feeding forward into the next stage. It captures the central idea that running an experiment is often seen as a separate, self-contained phase from the other steps (of course, there is iteration at a longer timescale between the final evaluation of a theory and the next series of investigations). This is the normal scientific method and a clear way to present the logic and results in PaperX.

However, it omits a lot of what actually happened during the project. Specifically, the researchers followed a much more *ad hoc* or iterative process to arrive at the final experimental design. To be clear, we are not criticising the methods or results of PaperX nor claiming that anything was maliciously concealed, but we suspect that other studies use both iterative and non-iterative processes, and we aim to discuss the differences between them. These differences can be categorized into three types.

First, the ideal process suggests a linear progression from theory to hypothesis to experiment design to results. However, for PaperX, the actual process was much more iterative, largely due to a lack of foresight on the researchers' part. For example, in a pilot (from Irvine and Roberts 2016), instead of creating a new language, the participants' virtual avatars were immediately killed by an enemy that inhabits the Minecraft world. This was not anticipated by the researchers, so the next trial was modified to remove enemies. However, in this trial, a virtual cow wandered into the arena, distracting participants, so all creatures were removed. Next, participants attacked each other's virtual avatars, so the experiment instructions were modified to emphasize cooperation. Next, the participants took so long that the virtual day/night cycle moved on, leaving them in the dark. So the experiment was modified yet again to remove Minecraft's day/night cycle.

At the time, we saw these events as a standard part of 'piloting' psycholinguistic experiments: tuning incidental parts of the design to focus on the main hypothesis. However, as we will argue below, these modifications were not always incidental parts of the design but had important analogies to the phenomenon we were investigating.

The second difference from the ideal process is that we investigated a wider range of experimental

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Figure 2. Use of pointing (top row) and gaze (bottom row) in PaperX. The second participant (labelled 'ExperPlayer3') taps the ground (top left) or gazes at a point (bottom left) to indicate a location (the crosshair indicates where the first participant is looking). The first participant then places a block in that location (top right and bottom right).

conditions than is reported in the main paper. This includes a condition with enforced breaks, during which participants were still allowed to interact with each other through their avatars in the virtual environment but were instructed not to engage in the experiment's building task. This condition was designed to test whether the presence of an opportunity cost (spending time on communicating rather than building) discouraged participants from creating a symbolic communication system. However, participants mostly used this time to fight each other in the virtual environment or, in one case, to build a prison for their partner in a virtual space located outside the designated experiment area. They were able to do so because, during the enforced break, they still had access to the building blocks provided for the experiment.

The final difference is that we did not have a clear idea of the core causal components before starting the empirical experiments. Instead, the causal graph which we present as the target for testing actually emerged from the interaction between our initial hunches and the results of the experiment. For example, initially the authors did not expect participants to shun a symbolic communicating system in favour of pointing and trial-and-error. In retrospect, it was an obvious response on the part of the participants and a clear prediction that the researchers could have made. However, humans are not good at perspective taking (Sulik and Lupyan 2018), and it sometimes requires an outsider's perspective to identify weaknesses in a design. Still, we had no formal method for integrating the observations from our results into our hypothesis.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

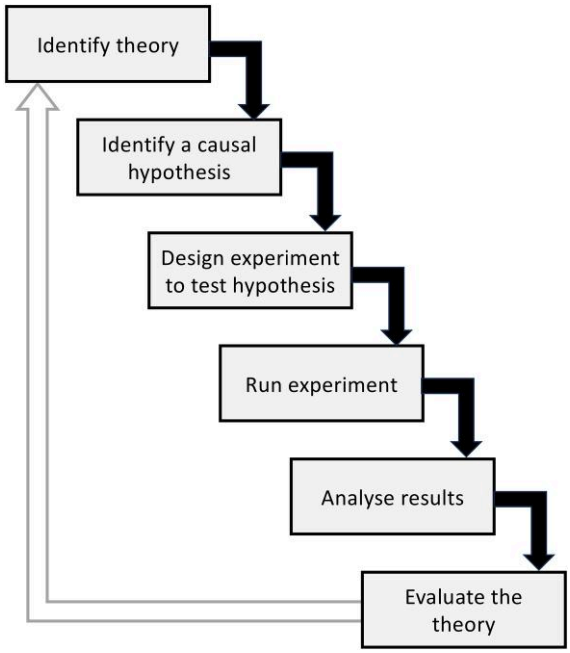

Figure 3. The linear 'waterfall' model of empirical research, based on Sommerville (2011: 30).

In summary, the real process of conducting the research (henceforth referred to as the 'cyclic approach') differed considerably from the logical progression of an ideal 'linear' study and from what was reported in the published report. We suggest that our experience is not uncommon, especially in more explorative research, including experimental semiotics and computational simulation. Still, we argue that there is no standard, formal process for guiding researchers in this type of research. In the next sections, we discuss possible methods to fill this gap.

3. Iterative research in practice: lessons from different disciplines

To situate the linear and cyclic approaches within a broader epistemological framework, it is helpful to consider their relationship to two fundamental modes of reasoning: linear inferential reasoning (comprising deductive and inductive logic) and cyclical dialectical reasoning. Deductive and inductive reasoning, in their classical definitions, fall under the category of linear inference, where reasoning proceeds in a step-by-step manner from premises or data towards conclusions. In deductive reasoning, a conclusion is logically derived from theoretical premises, moving from the general to the specific (Pereyra 2018: 3-1). In contrast, inductive reasoning draws generalized conclusions from specific empirical premises, which strongly support, but do not guarantee, the conclusion, moving from the specific to the general (Lau 2011: 89). On the other hand, cyclical dialectical reasoning, as articulated by Hegel ([1807] 1977), involves the recurrent re-examination of prior issues from newly attained perspectives, often through the resolution of contradictions and the synthesis of opposing ideas (Rescher 2022: 109).

Accordingly, the 'linear' or 'waterfall' approaches to studies reflect the principles of linear inferential reasoning, whether deductive or inductive, where progress is measured by a forward movement from premise to conclusion. In contrast, cyclic or iterative approaches mirror dialectical reasoning: they embody a cyclical process characterized by revision, opposition, experimentation, and the gradual refinement of ideas through repeated engagement. In what follows, we examine how iterative research is conceptualized in different disciplines, with particular attention to piloting in psycholinguistics, simulation modelling, experimental archaeology, and video game design.

3.1 Piloting in psycholinguistics

An iterative approach to experiment design comes close to the concept of 'piloting': 'a preliminary piece of research designed to 'road-test' various design elements (e.g. independent variables, dependent variables, details of procedure), in order to establish their viability and utility prior to the investment of time and money in a full study' (Haslam and McGarty 2019: 130), distinct from the sense of initial small-scale study that may be published in anticipation of a larger study. In our experience, piloting is seen as a normal part of the experimental process in psychology and psycholinguistics (e.g. Donnellan and Lucas 2013: 667, present it as one of the typical steps in primary research). However, there is little discussion of its role in theory building, or even methodological guidelines. Indeed, while there is a considerable literature on piloting for clinical research (see e.g. Thabane et al. 2010, or the journal *Pilot and Feasibility Studies*), we found little discussion in cognitive psychology (see supporting materials for a systematic search).

For example, across the 3,760 pages of 'Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience', piloting is only mentioned three times: twice in passing, then Wichmann and Jäkel (2018: 280, 281) suggest that piloting 'involves getting the flow of the experiment right', so participants are not exhausted. The definition at the start of this section is the only explicit one we could find in psychology textbooks (see Supplementary data), and we could find no advice on the methodology for pilots or their relation to theory.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

When piloting is mentioned, it is often in relation to avoiding 'mistakes' in the experimental procedure. This can include identifying stimuli that elicit unexpected responses (Kingston and Kramer 2013; Spector 2013) or avoiding floor and ceiling performance issues (Kantowitz et al. 2009: 273). Kekecs et al. (2023) suggest that piloting is one of the key methods for increasing a study's credibility by minimising 'the risk of mistakes, unforeseen events and protocol deviations' (Kekecs et al. 2023: 3).

Furthermore, piloting is characterized as distinct from developing theory or hypotheses (c.f. Ostrov and Hart 2013: 1, see Supplementary data). For example, Hales et al. (2019: 19) suggest that piloting 'enables a researcher to finalize their procedures, and also gives a clearer picture of potentially unexpected analytic issues that may arise. Researchers can then make these analytic decisions *a priori*, ensuring that they do not affect the conclusions that are drawn from the results'. Indeed, piloting is sometimes mentioned in relation to the replication crisis as a potential risk to research integrity. For example, Schott et al. (2019) discuss 'pilot dropping'—which they characterize as a kind of p-hacking or 'data peeking' where the decision of whether to continue with a study is made after the researcher sees the success or failure of the first few participants of an experiment. An author who 'changes her data collection plans based on observed data' is argued to run a risk of both Type I and Type II errors, as well as potentially needing to collect more data than would be necessary due to multiple false starts. The suggested solution is to use piloting to finalize the experimental method, then exclude the pilot data from the final analysis.

Overall, then, piloting is characterised as an attempt to avoid 'mistakes' and should be isolated from the real experiment. It seems that 'unexpected events' in an experiment are treated as undesirable, and changing a design on the basis of pilot results is seen as risky or unethical. We agree that clear separation between pilot data and final data is important and we are not suggesting that the decision to continue with a study should be based on the success of a pilot. Instead, we are arguing that, for language evolution, the adjustments to the pilot may provide useful insights into the assumptions behind the hypothesized mechanisms and should be discussed openly. This is in line with similar calls for transparency in piloting in psychology. For example, Rouder et al. (2019: 6) encourage a lab culture of recording and 'socializing mistakes', so that 'they are interpreted as reflecting a failure of systems rather than a failure of people'. In a similar vein, we want to encourage a culture of socializing insights that authors gain from piloting as an integral part of research.

3.2 Simulation modelling

Unexpected results are embraced in the 'simulation modelling' or 'living machines' approach (Webb 2000) which attempts to build robots based on insect biology, since 'the ultimate test of mechanistic understanding is the ability to build a machine that replicates the function' (Webb 2020: 1). The robot behaviour is compared with insect behaviour, with an emphasis on using physical implementations tested in real environments. Poor performance provides 'an opportunity to reject [hypotheses] and then improve performance by modifying the model' (Mangan et al. 2023: 2). This approach builds on 'synthetic psychology' (Braitenberg 1984), and Langton's 'synthetic' (versus more traditional 'analytic') approach to artificial life (Langton 1989), which try to identify the mechanistic ingredients necessary to create a phenomenon by building it from the 'bottom up'. That is, like artificial life studies seek to synthetically discover the ingredients for the emergence of emergent phenomena like flocking, PaperX was trying to find the ingredients that led to the emergence of semanticity. However, while artificial life aims to explore complex spaces that were not tractable with analytic methods/what is the best design, Webb's simulation modelling puts more emphasis on the problems we can find with theory along the way. Similar approaches have been used in language evolution, for example, the 'Talking Heads' experiments of Steels (1997, 2003; see also Smith 2014).

Language evolution lacks the same kind of ground truth to test the experimental results against. Still, if the experimental phenomenon is rigorously defined and a useful analogue of the real target phenomenon, we argue that it is still possible to use experimental semiotics to provide negative evidence and suggest aspects of a theory which require improvement (Irvine et al. 2013). However, this requires focusing on aspects that were unexpected or did not work, which is rare.

3.3 Experimental archaeology

Experimental archaeology (Wadley 2023) uses contemporary materials in realistic environments to test the plausibility of conditions for practices or the manufacture of objects to 'test out hypothetical scenarios using potentially authentic materials and conditions' and allow researchers to 'be inventive and develop new ways to enlarge our understanding by proxy' (Outram 2008: 2). For example, Bednarik's 'replicative archaeology' study of seafaring with Palaeolithic technology found that the amount of planning and communication involved, and its implications for the cognitive and linguistic abilities of Paleolithic humans, 'become evident in a way that remains entirely inaccessible to

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

conventional archaeology' (Bednarik 1998: 142). Similarly, various experiments have explored the need for verbal communication while teaching tool making (Ohnuma et al. 1997; Putt et al. 2014; Ruck 2014; Stade 2017). Contemporary participants were taught stone tool making by teachers using only gestures, only verbal communication or both. The actual tools made by participants were then assessed for effectiveness. Shilton (2019) concludes that there is no conclusive evidence that verbal language improves the transmission of tool-making skills, challenging prior theories. However, experimental archaeology is still not explicit at a high level on methods for arriving at these insights.

3.4 Video game design

While video game design may seem like a distant industry with very different goals, there are many parallels with experimental semiotics research. First, some experimental semiotics research literally involves video game design. For example, Nölle (2021) uses the Unity game engine to build experiments (the same engine as for the game *Overcooked*, which we discuss below), Namboodiripad et al. (2016) track gesture using Microsoft's Kinect for the Xbox, and Ravignani et al. (2013) uses Nintendo Wii controllers to measure primate rhythmic abilities. Games also provide model systems for cultural evolutionary dynamics (Strimling and Frey 2020; Allen et al. 2024). Virtual reality has been predicted to 'play a pivotal role in advancing our understanding of language evolution' (Nölle and Peeters 2025; Deffner et al. 2024).

However, we focus on more abstract principles of video game design. Practices have been developed to rigorously test games for player satisfaction, broadly labelled 'user testing' or 'playtesting' (Choi et al. 2016). The aim is to identify barriers to the player's experience rather than check the software's performance. Playtesting is a kind of iterative, user-centred interface design (Stone et al. 2005), which describes practices around building software using repeated phases of design, prototyping, and user testing. It recognizes that observing users engage with prototype systems is a more effective way of identifying problems in usability than the *a priori* predictions of designers (Nielsen 1993). Accordingly, in contrast to a linear waterfall model, it encourages stages of development to overlap and inform each other (Sommerville 1995: 7). The main outcome for user testing is 'a list of usability problems and suggestions for improvements in the interface' (Nielsen 1993), and ultimately a better qualitative understanding of 'how users behave and why' (Fessenden 2024).

Playtesting video games is an extension of user testing that seems particularly relevant for experimental semiotics since many games are embodied and many experiments have ludic qualities. Here, we focus on the practices of Valve, an award-winning video games developer responsible for several highly successful games selling tens of millions of copies each (e.g. the *Half-Life* series, the *Portal* series, *Counter-Strike*, *Team Fortress*, see Brown 2023). While most game developers use playtesting, Valve is known for using it intensively: *Half-Life* involved around 200 2-h playtest sessions, identifying thousands of issues (Birdwell 1999). Mike Ambinder, head of user experience for Valve with a background in experimental psychology, describes Valve's approach in experimental terms: 'We see our game designs as hypotheses and our playtests as experiments to validate these hypotheses' (Ambinder 2009; Bromley 2011).

That is, designs proposed by the developers are hypotheses about designs that would cause players to have fun. During playtests, the whole design team quietly observes players playing the game, noting quantitative and qualitative responses (e.g. game progress, body language, and explicit feedback). These evaluations provide empirical tests of the hypotheses and help set goals for modifications of the design which are tested in turn. The process is iterated until all barriers to player enjoyment have been identified and addressed. Iteration can take place at very short timescales (e.g. 1 week). In summary, 'get data early, get data often; iterate constantly' (Ambinder 2009).

Brown (2023) describes cases where this process led to rapid development of important elements that were not obvious to the designers. For example, during playtesting of *Portal*, they observed players struggling to understand what was part of the puzzle and what was decoration. In response, they revised the aesthetic to be clean and uncluttered. When playtesting this new design, players did indeed solve the problem, but mistook the real game for a tutorial. The designers realized that the environment now looked too artificial, and there was a missing dramatic element. Consequently, they invented GLaDOS, an antagonist AI obsessed with testing puzzle-solving abilities. That is, this aesthetic and central character—now iconic in video game culture (Harkin 2020)—were a product of Valve's iterative game design process rather than the developers' initial vision. Ambinder describes the benefits of this process as aiding idea generation, identifying problems, and solving design arguments. Importantly, it creates a feedback loop between playtesting (experiment results) and revised design (hypotheses).

Game developers also use 'thinking aloud' methods where players verbalize their thoughts while playing (Hoonhout 2022). Similar methods were used in a language evolution study by Byun et al. (2022): users of

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Table 1. Comparison between psychological experiments and video game playtesting.

Psychological experiments	Video game playtesting
Is hypothesis supported?	Is hypothesis robust?
Participants are realistic learning mechanisms	Participants are game testers trying to 'break' the hypothesis
Participants are tested in batches	Participants tested incrementally
Interested in average behaviour	Interested in the range of behaviour
Attempt to generalize to a population	Attempt to find limits of the design space
Testing hypothesized predictions (evaluative)	Testing assumptions (diagnostic)

different sign languages met for the first time and play a communication game based around referring to colours. Some of the strategies were not transparent, so Buyn watched the recorded trials with the original participants and asked what they were thinking, helping to guide the data coding.

In summary, video game playtesting involves practical testing, rapid iteration and an explicit feedback loop between player behaviour and design 'hypotheses'. This process seems different in several ways to the way that psychology methods are presented. Table 1 summarizes some of the differences. For example, psycholinguistic experiments aim to test whether a hypothesis is supported. In contrast, video game playtesting aims to test the *robustness* of the game design. Psychology participants are treated as realistic learning mechanisms, while playtesters are trying to 'break' the hypothesis. Psychology experiments test participants in large batches, and researchers are supposed to be blind to results during testing, mainly because researchers are interested in average behaviour and generalising to a population. In contrast, the playtesting loop can be applied by developers after every trial, because they are interested in the *range* of behaviours and mapping the limits of the design space. Psychology sees pilot studies as ancillary to the main goal, and mainly useful for 'bug testing', while playtesting *is* piloting. Ultimately, psychological experiments are aimed at testing hypothesized predictions while playtesting is aimed at testing assumptions.

Reflecting on our own practice, we see several parallels between playtesting and studies of language evolution. We are also trying to identify a design (environment/arena) that will lead to a particular response in participants (to invent a symbolic referential system), and trying to evaluate these designs by running people through them. That is, the participant data were not the end-goal, but a way of evaluating the theory. As Schell (2015) puts it, 'the whole point of playtesting is to make clear to you that some of the decisions you were completely comfortable with are completely wrong' or, closer to the bone for us, Birdwell (1999) states 'nothing is quite so humbling as being forced to watch in silence as some poor playtester stumbles around your level for 20 minutes, unable to figure out the "obvious" answer that you now realise is completely arbitrary and impossible to figure out'. That is, experimental semiotics has tended to use evaluation late in the design process, and focus on measurement evaluation (how well it worked). We are advocating for evaluation to be iterative and diagnostic (looking for many problems, Stone et al. 2005: 22, 23).

3.5 Case study: key properties of the arena of language evolution from video game design

In this section, we review the development of *Overcooked* (Ghost Town Games 2016), which used iterative playtesting to identify several key properties that manipulate the need for communication. We show how these principles may also be relevant for theories of language evolution.

In *Overcooked*, 2–4 players must work together to run a kitchen (chopping, cooking, cleaning, and serving). The developers of the game explicitly wanted to design situations where players would be motivated to communicate with each other—similar to PaperX. Although players have identical abilities and information, aspects of the environment create asymmetries. For example, the layout of one kitchen has a long table separating the fridge and the cooker, meaning it is quicker for one player to fetch the food from the fridge and pass it over the table for another player to cook. Other levels make movement between different parts of the kitchen impossible, or only allow one player to pass at a time. Under a time pressure, these asymmetries motivate players to communicate to coordinate roles and complete tasks efficiently (Brown 2018).

However, during playtesting, the designers found environmental asymmetries alone did not lead to a persistent pressure to communicate: 'after the initial flurry of activity and communication players would start to settle, cooperation became fairly rote and cooperation unspoken' (Duncan 2016). The designers used iterative playtesting to develop several solutions, including:

- Having more tasks than players, forcing players to switch roles.
- Disruptions: Task demands change to prevent repetitive strategies (e.g. the layout of the kitchen

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Table 2. Key properties identified by the design of *Overcooked*, and whether various arenas of communication exhibit them.

Key property	Bee honey production	Fire maintenance	Collaborative hunting
More tasks than individuals to promote role switching	Worker bees may have a dozen different tasks (Johnson 2010)	No	No
Dead time to promote multitasking	Honey takes time to concentrate during storage	Fire burns without intervention	Keeping watch, waiting in ambush
Disruptions that change task demands to prevent ritualization	No	Rain, shifts in fuel structure which affect airflow.	Prey flight responses, injuries
Public goods to promote coordination	Produces food for the whole hive	Provides heat, light for all	Provides food, materials for all

changes so that players are separated and forced to change tasks).

- Dead time: Some tasks take time to complete without needing player activity (e.g. cooking rice). The efficient response is to multitask, also helping to switch roles and potentially requiring the recruitment of others to step in if things get out of hand.
- Public goods: Some activities are not directly rewarded and considered menial (e.g. washing dishes), but benefit everyone. Players must coordinate about who will do these.

These factors force players into asymmetric roles that change, motivating players to communicate continuously and from reciprocal perspectives. These factors also feature in some theories of language evolution (table 2). For example, fire maintenance (Twomey 2013) involves dead time and is a public good. Collaborative hunting (Sterelny 2012) involves role assignment and disruption. We note that honeybee communication about the location of nectar involves role switching, dead time, and public goods, but it is not clear there are disruptions and indeed, the communication system is genetically ritualized rather than learned. Furthermore, a combination of all these factors may be rare enough to explain why only humans have language (see also e.g. Sterelny 2012). We note that theories based on fire maintenance exhibit all the factors above.

More generally, the design of *Overcooked* suggests that the more abstract property of 'asymmetry' may be a key property of an arena that motivates the emergence of symbolic communication. For example, displaced reference (Kazakov and Bartlett 2004) involves asymmetry in information and within-group competition (Knight et al. 1995) involves asymmetry in reproductive abilities. Various predictions are testable, such as whether gamers communicate most about task demands when asymmetry is highest (cf. Peña and Hancock 2006; Ensslin 2011), or whether complex non-human communication arises in environments that have parallels of asymmetrical design (e.g. honeybees switch tasks, have dead time and public goods when caring for larvae; primate alarm calls are motivated by an asymmetry in information; ants have asymmetries in abilities). In summary, iterative playtesting generated insights that might be helpful for developing language evolution theory, suggesting that it is a productive methodology.

3.6 Case study: designing an arena for language evolution

In this section, we describe the application of the playtesting approach to a language evolution experiment. This relates to the experiment for the second arena in PaperX based on maintaining a fire. In this arena, one participant was trained as a 'miner' to fetch materials from a distant mine, and another participant was trained as a 'smelter' who refined the materials and knew which material was needed next. The expectation was that, since they could not point at the distant materials, participants would develop a symbolic system for the smelter to communicate to the miner which material was needed.

The authors took a cyclic, iterative approach. The methods were similar to the previous experiment, but a video game playtesting schedule was followed: One trial of the experiment was run, the behaviour was documented, and then the researchers met to discuss the results. During this discussion the researchers identified issues, related them to the theory and hypothesis, and changed the design. Those changes were implemented, and then another trial was run. This happened in the order of days, rather than the typical process of collecting a larger sample for a more limited number of designs over weeks. This playtesting process was used to refine the experiment design and explore the assumptions of the theory, but we emphasize that the final

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

results reported in PaperX come from a standard application of psychological experimental principles on a separate cohort of participants that was run after this more exploratory stage.

One practical question was how to document the participant behaviour to bridge the gap between evidence and theoretical insights. We suggest that this can be accomplished by 'thick description'. This is the anthropological practice described by Geertz (2005) for making detailed descriptions of human behaviour. Other qualitative, descriptive approaches are available (Bernard 2017), but thick description seemed to fit the experiment well. We applied this practice to observing and documenting participant behaviour, effectively using it to create a 'laboratory notebook' (Schnell 2015). The thick description included observations about participant characteristics, task behaviour, emotions, and responses to the post-task interview. For example:

> The participants were a Masters student and PS [Professional Services] staff who knew each other. The two participants went immediately to the mine. After some hesitation, the smelter pointed at one box and the miner opened it and brought it back to the furnace with the smelter. … After this, they would meet up next to the final deposit box. Because they'd already made a stock of ingots, the smelter used the ingots to indicate which ore to get (e.g. holding a gold ingot to signal to get a gold ore). In the interview, the smelter said that they tried to construct a language, and the miner stated they had not realised they were trying to do this. But they said they didn't need a signalling system after inventing the system of pointing to the ingots.

The process of writing these observations as well as reviewing them, and interpreting their meanings in totality during analysis provided an opportunity to reflect on the 'unexpected' behaviour of participants. Rather than treating these as a bug or 'mistake' to be avoided, we realized they provided insights into important analogies with the theories we were trying to test (Table 3). For example, the case where player avatars were killed by the enemies present in Minecraft presented an opportunity to reflect on the assumptions about predation risk. After all, the real world is not free from dangerous predators, so is one of the assumptions that our ancestors were in an environment with relatively low predation risk? This insight revealed links to other theories that explicitly include predation risk, such as the relation between predation, group size, and division of labour (Dunbar 2004), that collective vocalization deters predators (Kight and Lewis 2017), and various theories of reduced selection pressures due to self-domestication (Thomas and Kirby 2018). Similarly, for the case where one player built a 'prison' for the other, we updated the instructions to make it clearer that participants were cooperating. However, this also revealed to us our theoretical assumptions about cooperation (Knight et al. 1995; Dessalles 1998).

Table 3. Observations and insights from the case study.

Observation	Insight into our assumptions about …
Participants killed by enemies	Predation risk
Participant A built a jail for participant B	Cooperation
Participant A found difficult not to say 'please'.	Politeness
One knock for gold, three for emerald, to reflect syllable structure of English words.	The role of iconicity
Participant got bored going back and forth to the mine.	Motivation and attention
The miner switched roles with the smelter, meaning they didn't need to communicate	Role switching

Other unexpected behaviours raised revealed further assumptions. For example, one pair used one knock for gold ore and three knocks for emerald ore. They explained that this reflected the number of syllables in the target words. In another trial, participants pointed to a green block that was part of the training area to refer to emerald ore. While these might be interpreted as 'bugs' exploiting a prior linguistic system, it raises questions about the role that iconicity might play in the theory. The researchers assumed the experiment was immune to iconic affordances but, as is often the case, humans are good at finding patterns. The second case in particular suggests that even if the intended referent is distal, there still may be proximal objects with analogous properties that can be pointed to (e.g. pointing to something brown to refer to wood), removing the need to create linguistic symbols. This is an additional argument for why pointing inhibits the evolution of symbolic systems which did not occur to the researchers until they consciously reflected on the behaviour.

Even behaviour that seemed tangential or incidental was useful. For example, one participant reported getting bored walking back and forth to get materials, causing them to lose focus and engage less with the task. This prompted us to make the physical route to the mine more challenging, which seemed to fix the 'bug', but more importantly revealed a key cognitive assumption of the scenario: that the 'miner'/forager is

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

able to maintain memory, attention, and motivation (Brinck 2001; Corballis 2019), reminiscent of the importance of 'disruption' in the *Overcooked* example. Another participant reported difficulty inhibiting themselves from saying 'please' and 'thank you', suggesting that politeness could provide its own 'arena' for language evolution (Mühlenbernd et al. 2021).

Even in the main experiment phase, participants came up with unexpected solutions. For example, in one experiment the participants switched roles. This meant that the player who knew the order of materials was also the one fetching them, meaning that the pressure to communicate disappeared and they did not invent a symbolic system. Similarly, in the building arena, in one case a participant just placed random blocks for the other to leave or remove. These cases are very creative solutions that had not occurred to the researchers, and they raised questions about the assumptions about role switching and how fragile the conditions for motivating symbolic systems may be.

In summary, playtesting was useful not only in adjusting the conditions for the main experiment, but in developing the theory itself. The participants helped the researchers explore the design space of our experiments. The causal structure of the theory that is presented in PaperX is the product of participants helping us to explore the design space of our experiments and of serious reflection on their behaviour.

4. Discussion: a participatory approach

After reflecting on the video game approach to playtesting, we realized that it had changed our attitude to the role of participants. In classic psychological experiments, participants are passive subjects, often uniformed about the aim of the experiment, and assumed to be exhibiting normal behaviour. In contrast, video game playtesting treats participants as active participants or collaborators in trying to 'break' the hypotheses. Game testers are knowledgeable about the domain and are used as a source of creative thinking. Indeed, there is a whole subculture of 'speedrunning' games, where players find creative exploits and technical glitches to complete games in the shortest time possible. It is commonplace for a game that has undergone substantial play testing and bug testing to be released only for players to discover some exploit within days. For example, the creator of *Animal Well* expected some puzzles to remain permanently unsolved, but the community solved them within a week. Other games have been exploited to such an extent that players can 'beat' games that were designed to be 'unbeatable'. This is achieved through a process of knowledge collaboration paired with competitive practice (Escobar-Lamanna 2019). In a similar way, PaperX benefitted from participants 'breaking' the experiment by finding creative solutions that the researchers had not considered.

This shift in attitude reminded us of crowd sourcing or 'citizen science' approaches, and more broadly of the participatory ethnography movement (Blomberg and Karasti 2012), which 'deliberately and explicitly emphasizes collaboration at every point in the ethnographic process, without veiling it' (Lassiter 2005: 15). In the Minecraft case study, participants demonstrated very creative thinking. They had become informants, our guides to knowledge about the phenomena we were researching. This kind of contribution could be acknowledged more openly in language evolution research (indeed, PaperX does so). More generally, this is also an argument for more diverse participant populations in order to elicit more diverse perspectives and solutions.

In this sense, the experimental studies above align with other participatory approaches. For example, the *Language Evolves* project taught science-fiction authors language evolution theories, then ran a short story competition on the theme of language evolution (Roberts et al. 2021). Part of the aim was to harness the speculative skills of science-fiction authors and their ability to embody radically different perspectives in order to reveal gaps in our theories and suggest new avenues for research. For example, when discussing PaperX, one of the authors asked 'what happens when it gets dark?' This point had not occurred to the researchers but was obvious to an author putting themselves into the shoes of our ancestors. In another example, the short story *The Precious Space* by Tim Byrne was inspired by experimental semiotics experiments: explorers trapped in a temporal anomaly invent a new gesture system for diplomatic negotiation. While most theories imagine language evolving in cooperative scenarios, this story offered another perspective—language evolved to help navigate *disagreement*. Exploring this idea further, the researchers found a theory from the 1970s by Soviet scientist Porshnev (1974). This offers an opportunity to revisit this theory using modern methods, creating a productive loop of collaboration between researchers and authors.

5. Conclusion

We have argued that experimental approaches to studying language origins can benefit from an explorative approach that embraces rapid piloting and explicitly reflects on unexpected results in order to uncover gaps in the theory. We suggest that methods from video game development (playtesting) and ethnography (thick description) can provide concrete methods for

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

conducting this approach. Here we make some final observations about our position.

We feel that our own research practice is closer to video game design than to traditional psychology methods, and not just because we were using a video game for data collection: we were using the experiments as a way of evaluating the robustness of theories about language evolution and for finding weaknesses in the assumptions of the theories. For example, the insight that pointing is a *barrier* to the emergence of symbolic referential signals did not come from the theory or hypothesis stages, but from reflecting on the unexpected results and how they related to the theory.

We suspect that many researchers in language evolution and many researchers in other fields doing exploratory research already implicitly follow this kind of process. However, few studies discuss this explicitly. Therefore, we suggest some modification to the 'linear' process described above into a 'hybrid' model that combines cyclic and linear phases:

- Identify the relevant theories about the target phenomenon
- Identify the causal claims in the theories and generate a testable hypothesis (using tools from causal inference)
- Design an experiment to test the hypothesis using principles from experimental semiotics and a common task framework
- Run a small number of trials
- Use thick description to capture participant behaviour
- Identify failures and unexpected results, and relate these back to the theory/hypothesis in terms of unidentified assumptions
- Revise the theory/hypothesis and iterate this process until no more unexpected results are found
- Run a new set of final experiments to test the robustness of patterns

To visualize this, we draw again on user interface design theory, and adapt the 'waterfall' model into a hybrid 'waterwheel' model which combines both the cyclic, iterative piloting phase and the final linear experiment phase (Fig. 4). This is similar to some game development life cycle models, with the final phase aimed at the final release of the game (Ramadan and Widyani 2013). Given the discussion of participatory research above, if we see participants as a central resource for external evaluation on all aspects of a project, we might be able to go even further and adopt a 'star-shaped' approach, based on maximally iterative software development, where external evaluation is the nexus that connects all the other phases (Hix and Hartson 1993).

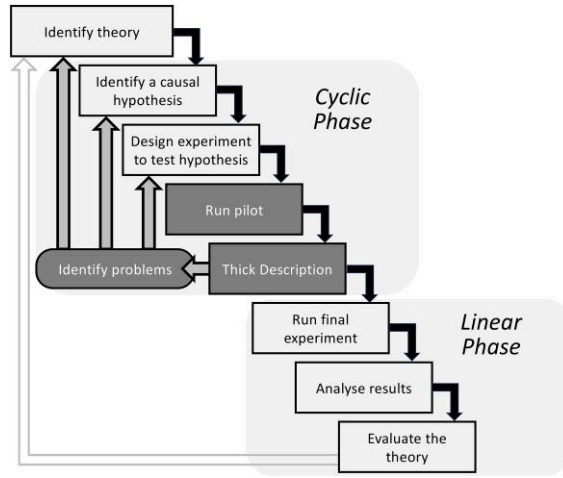

Figure 4. A hybrid 'waterwheel' approach to empirical research, with cyclic and linear phases.

There are several challenges to adopting the hybrid approach. First, we acknowledge that there are good reasons that psychological experiments avoid drawing conclusions based on single trials of an experiment. We are also not advocating p-hacking: p-hacking involves running the same experiment until the critical differences become statistically significant. Instead, we are advocating changing the experiment until all the hidden assumptions in the design are revealed. We also note that this method is particularly suited to finding problems with theories rather than positive support, for which standard psychological methods are more suitable.

Second, video game design is aimed at engineering an effective commercial product so there are questions about its legitimacy for conducting scientific research. However, it has much in common with exploratory experimentation as discussed in the philosophy of science (Steinle 2002; Waters 2007; García and Velasco 2013). In particular, we see a parallel with the idea that researchers work with 'local theories' which inform 'auxiliary hypotheses' (temporary, working hypotheses) that they use to make choices about their research design (Colaço 2018) or, indeed, with the ethnographic distinction between 'big-T' and 'little-T' theories (Snow et al. 2003; Lichterman and Reed 2015; Wherry 2017). There are related concepts in the design theory of 'design experiments' as an important step in exploratory phases (Gibbons and Bunderson 2005), as well as the study of 'ideation' (Shah et al. 2003; Stremersch 2024) and creativity (Lubart 2001). The value of using video game playtesting will likely apply to exploratory research rather than confirmatory research, and especially where the target system under investigation cannot be accessed directly (as in language evolution).

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Third, while playtesting uses participants who are directly representative of their target population and prototypes that are in the same medium and mode as their target product, some experimental semiotics experiments assume that modern humans are suitable for evaluating theories about ancient humans, raising questions about parallels with cognitive and social abilities, and with virtual environments (Irvine et al. 2013; Müller and Raviv 2025). This could mean that the method is more suitable for investigating general conditions for language emergence than what actually happened in the case of humans (Roberts 2018). There are related questions about how naive participants should be, and in line with playtesting, we suggest a range of experience and informedness may be useful in different situations.

Finally, evolutionary linguistics has used psychology methodology and practice to enhance its position as an empirical science. Researchers may resist radical changes in procedures. However, we are only advocating that playtesting is used as one part of research practice. That is, papers may adopt playtesting to refine the assumptions of the theory and finalize the experimental design, then use that design in a more traditional psychology process to obtain the final results.

We look forward to reinvigorating many theoretical debates using new empirical methods.

Acknowledgements

The authors thank the reviewers and the participants for their creative insights.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

Funding

This project was supported by an AHRC grant AH/T006927/1.

Data availability

Videos of the experiments are available online: https://doi.org/10.17605/OSF.IO/SMRB9, https://doi.org/10.17605/OSF.IO/85AZS, https://doi.org/10.17605/OSF.IO/A8Q3N

References

Allen, K. et al. (2024) 'Using games to understand the mind', *Nature Human Behaviour*, 8: 1035–1043. https://doi.org/10.1038/s41562-024-01878-9

Ambinder, M. (2009) 'Valve's Approach to Playtesting: the Application of Empiricism', Game Developer's Conference 2009. https://www.gdcvault.com/play/1566/Valve-s-Approach-to-Playtesting, accessed 1 Oct. 2024.

Bednarik, R. G. (1998) 'An Experiment in Pleistocene Seafaring', *International Journal of Nautical Archaeology*, 27: 139–49. https://doi.org/10.1111/j.1095-9270.1998.tb00797.x

Bernard, H. R. (2017) *Research Methods in Anthropology: Qualitative and Quantitative Approaches*. London: Bloomsbury Publishing PLC.

Birdwell, K. (1999) 'The Cabal: Valve's Design Process For Creating Half-Life. Game Developer'. https://www.gamedeveloper.com/design/the-cabal-valve-s-design-process-for-creating-i-half-life-i-, accessed 10 Dec. 1999.

Blomberg, J. and Karasti, H. (2012) 'Ethnography: Positioning Ethnography Within Participatory Design', in J. Simonsen, T. Robertson (eds.) *Routledge International Handbook of Participatory Design*, pp. 86–116. Abingdon: Routledge.

Braitenberg, V. (1984) *Vehicles: Experiments in synthetic psychology*. Cambridge: MIT Press.

Brinck, I. (2001) 'Attention and the Evolution of Intentional Communication', *Pragmatics & Cognition*, 9: 259–77. https://doi.org/10.1075/pc.9.2.05bri

Bromley, S. (2011) 'Valve's philosophy with User Research in Games' Steve Bromely's Blog. https://www.stevebromley.com/blog/2011/09/01/valves-philosophy-with-user-research-in-games-habe-newell-and-mike-ambinder/, accessed 1 Oct. 2024.

Brown, M. (2018) '*How Overcooked's Kitchens Force You to Communicate*'. https://www.youtube.com/watch?v=C3M8BvWcJQY, accessed 1 Oct. 2024.

Brown, M. (2023) '*Valve's "Secret Weapon"*'. https://www.youtube.com/watch?v=9Yomqk0C6kE, accessed 01/10/2024.

Byun, K. -S. et al. (2022) 'Distinguishing selection pressures in an evolving communication system: Evidence from color-naming in "cross signing"', *Frontiers in Communication*, 7. https://doi.org/10.3389/fcomm.2022.1024340

Choi, J. O. et al. (2016) 'Playtesting with a Purpose, *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, pp. 254–65. New York: Association for Computing Machinery https://doi.org/10.1145/2967934.2968103

Colaço, D. (2018) 'Rethinking the role of theory in exploratory experimentation', *Biology & Philosophy*, 33. https://doi.org/10.1007/s10539-018-9648-9

Corballis, M. C. (2019) 'Language, Memory, and Mental Time Travel: An Evolutionary Perspective', *Frontiers in Human Neuroscience*, 13: 455779. https://doi.org/10.3389/fnhum.2019.00217

Deffner, D. et al. (2024) 'Collective Incentives Reduce Over-Exploitation of Social Information in Unconstrained Human Groups', *Nature Communications*, 15: 2683. https://doi.org/10.1038/s41467-024-47010-3

Dessalles, J. L. (1998) 'Altruism, status, and the Origin of Relevance', in J. R. Hurford, M. Studdert-Kennedy, C. Knight (eds.) *Approaches to the Evolution of Language*, pp. 130–47. Cambridge: Cambridge University Press.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Donnellan, M. B. and Lucas, R. E. (2013) 'Secondary Data Analysis', in T. D. Little (ed.) *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*. Oxford: Oxford University Press https://doi.org/10.1093/oxfordhb/9780199934898.013.0028

Dunbar, R. I. (2004) 'Gossip in Evolutionary Perspective', *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, 8: 100. https://doi.org/10.1037/1089-2680.8.2.100

Duncan, P. (2016) 'Building Truly Cooperative Play in Overcooked'. *Game Developer*. https://www.gamedeveloper.com/design/game-design-deep-dive-building-truly-cooperative-play-in-i-overcooked-i-, accessed 1 Oct. 2024.

Ensslin, A. (2011) *The Language of Gaming*. London: Macmillan International Higher Education.

Escobar-Lamanna, J. (2019) 'Why Speed Matters: Collective Action and Participation in Speedrunning Groups', MA thesis, University of Toronto.

Fessenden, T. (2024) Design Processes for High Usability: Iterative Design, Parallel Design, and Competitive Testing. Nielsen Norman Group. https://www.nngroup.com/articles/parallel-and-iterative-design/, accessed 1 Oct. 2024.

García, P (2013) 'Exploratory strategies: experiments and simulations', in E. Arnold, J. M. Durán (eds.) *Computer Simulations and the Changing Face of Scientific Experimentation*. Cambridge: Cambridge Scholars Publishing.

Geertz, C. (2005 [1973]) '*Deep play: Notes on the Balinese cockfight*', *Daedalus*, 134: 56–86. https://doi.org/10.1162/001152605774431563

Ghost Town Games. (2016) '*Overcooked*' [Nintendo Switch], Team17.

Gibbons, A. S. and Bunderson, C. V. (2005) 'Explore, Explain, Design', in K. Kempf-Leonard (ed.) *Encyclopedia of Social Measurement*, pp. 927–38. Amsterdam: Elsevier.

Hales, A. H., Wesselmann, E. D. and Hilgard, J. (2019) 'Improving Psychological Science Through Transparency and Openness: An Overview', *Perspectives on Behavior Science*, 42: 13–31. https://doi.org/10.1007/s40614-018-00186-8

Harkin, S. (2020) 'The Only Thing You've Managed to Break So Far is My Heart": An Analysis of Portal's Monstrous Mother GLaDOS', *Games and Culture*, 15: 529–43. https://doi.org/10.1177/1555412018819663

Haslam, S. A. and McGarty, C. (2019) *Research Methods and Statistics in Psychology*. London: SAGE Publications Limited.

Hegel, G. W. F., Miller, A. V. and Findlay, J. N. (1977) *Phenomenology of Spirit* (Vol. 632). Oxford: Oxford University Press. (Original work published 1807).

Hix, D. and Hartson, H. R. (1993) *Developing User Interfaces*. New Jersey: Wiley.

Hoonhout, J. (2022) *Let the Game Tester do the Talking: Think Aloud and Interviewing to Learn About the Game Experience*, *Game Usability*, pp. 115–25. Boca Raton: CRC Press.

Hurford, J. R. (2007) *Language in the Light of Evolution*. Oxford: Oxford University Press.

Irvine, E. and Roberts, S. (2016) 'Deictic Tools Can Limit the Emergence of Referential Symbol Systems', in S. G. Roberts (ed.) *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. New Orleans: EvoLang Scientific Committee. http://evolang.org/neworleans/papers/99.html, accessed 1 Oct. 2024.

Irvine, E., Roberts, S. and Kirby, S. (2013) 'A Robustness Approach to Theory Building: A Case Study of Language Evolution', *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35: 2614–2619. https://escholarship.org/uc/item/1rw611dp

Johnson, B. R. (2010) 'Division of Labor in Honeybees: Form, Function, and Proximate Mechanisms', *Behavioral Ecology and Sociobiology*, 64: 305–16. https://doi.org/10.1007/s00265-009-0874-7

Kantowitz, B. H., Roediger, H. L. and Elmes, D. G. (2009) *Experimental Psychology*. Boston: Wadsworth Cengage Learning.

Kazakov, D. and Bartlett, M. (2004) 'Cooperative Navigation and the Faculty of Language', *Applied Artificial Intelligence: AAI*, 18: 885–901. https://doi.org/10.1080/08839510490509072

Kekecs, Z. et al. (2023) 'Raising the Value of Research Studies in Psychological Science by Increasing the Credibility of Research Reports: The Transparent Psi Project', *Royal Society Open Science*, 10: 191375. https://doi.org/10.1098/rsos.191375

Kingston, N. M. and Kramer, L. B. (2013) 'High-Stakes Test Construction and Test Use', in T. D. Little (ed.) *The Oxford Handbook of Quantitative Methods in Psychology* (Vol. 1). Oxford: Oxford University Press https://doi.org/10.1093/oxfordhb/9780199934874.013.0010

Knight, C. and Lewis, J. (2017) 'Wild Voices: Mimicry, Reversal, Metaphor, and the Emergence of Language', *Current Anthropology*, 58: 435–453. https://doi.org/10.1086/692905

Knight, C., Power, C. and Watts, I. (1995) 'The Human Symbolic Revolution: A Darwinian Account', *Archaeological Journal*, 5: 75–114. https://doi.org/10.1017/S0959774300001190

Langton, C. G. (1989) 'Artificial Life, *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*, pp. 1–47. Boston: Addison-Wesley.

Lassiter, L. E. (2005) *The Chicago guide to collaborative ethnography*. Chicago: University of Chicago press.

Latour, B. and Woolgar, S. (1979) *Laboratory Life: The Construction of Scientific Facts*. Beverly Hills: Sage Publications.

Lau, J. Y. (2011) *An introduction to Critical Thinking and Creativity: Think More, Think Better*. New Jersey: John Wiley & Sons.

Lichterman, P. and Reed, I. A. (2015) 'Theory and Contrastive Explanation in Ethnography', *Sociological Methods & Research*, 44: 585–635. https://doi.org/10.1177/0049124114554458

Lubart, T. I. (2001) 'Models of the Creative Process: Past, Present and Future', *Creativity Research Journal*, 13: 295–308. https://doi.org/10.1207/S15326934CRJ1334_07

Mangan, M. et al. (2023) 'A Virtuous Cycle Between Invertebrate and Robotics Research: Perspective on a

Downloaded from https://academic.oup.com/jole/article/10/1/laf006/8362096 by guest on 10 December 2025

Decade of Living Machines Research', *Bioinspiration & Biomimetics*, 18: 035005. https://doi.org/10.1088/1748-3190/acc223

Mühlenbernd, R., Wacewicz, S. and Żywiczyński, P. (2021) 'Politeness and Reputation in Cultural Evolution', *Linguistics and Philosophy*, 44: 1181–213. https://doi.org/10.1007/s10988-020-09315-6

Müller, T. F. and Raviv, L. (2025) 'Communication Games: Social Interaction in the Formation of Novel Communication Systems', in C. Boeckx, L. Raviv (eds.) *Oxford Handbook of Approaches to Language Evolution*, pp. 41–62. Oxford: Oxford University Press.

Namboodiripad, S., Lenzen, D., Lepic, R. and Verhoef, T. (2016) 'Measuring conventionalization in the manual modality', *Journal of Language Evolution*, 1: 109–118. https://doi.org/10.1093/jole/lzw005

Nielsen, J. (1993) 'Iterative user-interface design', *Computer*, 26: 32–41. https://doi.org/10.1109/2.241424

Nölle, J. (2021) *How language adapts to the environment: An evolutionary, experimental approach (Doctoral Dissertation)*. Edinburgh: University of Edinburgh Press. https://doi.org/10.7488/era/2144

Nölle, J. and Galantucci, B. (2022) 'Experimental Semiotics: Past, Present and Future, *The Routledge Handbook of Semiosis and the Brain*, pp. 1–12. Abingdon: Routledge.

Nölle, J. and Peeters, D. (2025), in L. Raviv, C. Boeckx (eds.) *The Oxford Handbook of Approaches to Language Evolution*, pp. 63–82. Oxford, UK: Oxford University Press https://doi.org/10.1093/oxfordhb/9780192886491.013.5

Ohnuma, K., Aoki, K. and Akazawa, T. (1997) 'Transmission of Tool-making through Verbal and Non-verbal Communication: Preliminary Experiments in Levallois Flake Production.', *Anthropological Science*, 105: 159–168. https://doi.org/10.1537/ase.105.159

Ostrov, J. M. and Hart, E. J. (2013) 'Observational Methods', in T. D. Little (ed.) *The Oxford Handbook of Quantitative Methods in Psychology, Vol. 1, Oxford Library of Psychology*, pp. 286–304. Oxford: Oxford Academic.

Outram, A. K. (2008) 'Introduction to experimental archaeology', *World Archaeology*, 40: 1–6. https://doi.org/10.1080/00438240801889456

Peña, J. and Hancock, J. T. (2006) 'An Analysis of Socioemotional and Task Communication in Online Multiplayer Videogames', *Communication Research*, 33: 92–109. https://doi.org/10.1177/0093650205283103

Pereyra, N. A. (2018) *Logic for Physicists*. San Rafael: Morgan & Claypool Publishers.

Pinker, S. and Bloom, P. (1990) 'Natural language and natural selection', *Behavioral and Brain Sciences*, 13: 707–727. https://doi.org/10.1017/S0140525X00081061

Porshnev, B. F. (1974) *On the Beginning of Human History* (Problems of Paleopsychology). Moscow: Mysl.

Putt, S. S., Woods, A. D. and Franciscus, R. G. (2014) 'The role of verbal interaction during experimental bifacial stone tool manufacture', *Lithic Technology*, 39: 96–112. https://doi.org/10.1179/0197726114Z.00000000036

Ramadan, R. and Widyani, Y. (2013) 'Game Development Life Cycle Guidelines, *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 95–100. New York: IEEE.

Ravignani, A. et al. (2013) 'Primate Drum Kit: A System for Studying Acoustic Pattern Production by Non-Human Primates Using Acceleration and Strain Sensors', *Sensors*, 13: 9790–9820. https://doi.org/10.3390/s130809790

Rescher, N. (2022) *Philosophical Fallacies: Ways of Erring in Philosophical Exposition*. London: Springer Nature.

Roberts, G. (2017) 'The Linguist's Drosophila: Experiments in Language Change', *Linguistics Vanguard: Multimodal Online Journal*, 3: 20160086. https://doi.org/10.1515/lingvan-2016-0086

Roberts, S. G. (2018) 'What are the Social, Economic and Ecological Conditions for the Evolution of complex Communication Systems? Comment on "Rethinking Foundations of Language from a Multidisciplinary Perspective" by T. Gong et al', *Physics of Life Reviews*, 26: 152–4. https://doi.org/10.1016/j.plrev.2018.06.016

Roberts, S. G., Little, H. and Silvey, C. (2021) 'Language Evolves on Scientists and Authors Writing Stories Together', *New Welsh Review*, 126: 26–29.

Rouder, J. N., Haaf, J. M. and Snyder, H. K. (2019) 'Minimizing Mistakes in Psychological Science', *Advances in Methods and Practices in Psychological Science*, 2: 3–11. https://doi.org/10.1177/2515245918801915

Ruck, L. (2014) 'Manual Praxis in Stone Tool Manufacture: Implications for Language Evolution', *Brain and Language*, 139: 68–83. https://doi.org/10.1016/j.bandl.2014.10.003

Ryle, G. (1968) 'Thinking and Reflecting', *Royal Institute of philosophy supplements*, 1: 210–26. https://doi.org/10.1017/S0080443600011511

Schell, J. (2015) *The Art of Game Design: A Book of Lenses*. Boca Raton: CRC Press.

Schnell, S. (2015) 'Ten Simple Rules for a Computational Biologist's Laboratory Notebook', *PLoS Computational Biology*, 11: e1004385. https://doi.org/10.1371/journal.pcbi.1004385

Schott, E., Rhemtulla, M. and Byers-Heinlein, K. (2019) 'Should I Test More Babies? Solutions for Transparent Data Peeking', *Infant Behavior & Development*, 54: 166–76. https://doi.org/10.1016/j.infbeh.2018.09.010

Shah, J. J., Smith, S. M. and Vargas-Hernandez, N. (2003) 'Metrics for Measuring Ideation Effectiveness', *Design studies*, 24: 111–34. https://doi.org/10.1016/S0142-694X(02)00034-0

Shilton, D. (2019) Is Language Necessary for the Social Transmission of Lithic Technology?', *Journal of Language Evolution*, 4: 124–133. https://doi.org/10.1093/jole/lzz004

Smith, A. D. M. (2014) 'Models of Language Evolution and Change', *Wiley Interdisciplinary Reviews: Cognitive Science*, 5: 281–93. https://doi.org/10.1002/wcs.1285

Snow, D. A., Morrill, C. and Anderson, L. (2003) 'Elaborating Analytic Ethnography: Linking Fieldwork and Theory', *ethnography*, 4: 181–200. https://doi.org/10.1177/14661381030042002

Sommerville, I. (1995) *Software Engineering*, 5th edn. Harlowe: Addison-Wesley.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025

Sommerville, I. (2011) *Software Engineering*, 9th edn. London: Pearson Education Inc.

Spector, P. E. (2013) 'Survey Design and Measure Development', in T. D. Little (ed.) *The Oxford Handbook of Quantitative Methods in Psychology* (Vol. 1). Oxford: Oxford University Press https://doi.org/10.1093/oxfordhb/9780199934874.013.0009

Stade, CM. (2017) *Lithic morphological variability as a proxy for Palaeolithic linguistic ability: A knapping training study exploring cultural transmission, theory of mind and language* (Doctoral dissertation). Southampton: University of Southampton. http://eprints.soton.ac.uk/id/eprint/416633

Steels, L. (1997) 'The Synthetic Modeling of Language Origins', *Evolution of communication*, 1: 1–34. https://doi.org/10.1075/eoc.1.1.02ste

Steels, L. (2003) 'Evolving Grounded Communication for Robots', *Trends in cognitive sciences*, 7: 308–12. https://doi.org/10.1016/S1364-6613(03)00129-3

Steinle, F. (2002) 'Experiments in History and Philosophy of Science', *Perspectives on Science*, 10: 408–432. https://doi.org/10.1162/106361402322288048

Sterelny, K. (2012) *The Evolved Apprentice*. Cambridge: MIT press.

Stone, D. et al. (2005) *User interface design and evaluation*. San Francisco: Morgan Kaufman.

Stremersch, S. (2024) 'How can Academics Generate Great Research Ideas? Inspiration from Ideation Practice', *International Journal of Research in Marketing*, 41: 1–17. https://doi.org/10.1016/j.ijresmar.2023.10.002

Strimling, P. and Frey, S. (2020) 'Emergent Cultural Differences in Online Communities' Norms of Fairness', *Games and Culture*, 15: 394–410. https://doi.org/10.1177/1555412018800650

Sulik, J. and Lupyan, G. (2018) 'Perspective taking in a novel signaling task: Effects of world knowledge and contextual constraint.', *Journal of Experimental Psychology: General*, 147: 1619–1640. https://doi.org/10.1037/xge0000475

Tamariz, M. and Kirby, S. (2016) 'The Cultural Evolution of Language', *Current Opinion in Psychology*, 8: 37–43. https://doi.org/10.1016/j.copsyc.2015.09.003

Thabane, L. et al. (2010) 'A Tutorial on Pilot Studies: The What, why and how', *BMC medical research methodology*, 10: 1–10. https://doi.org/10.1186/1471-2288-10-1

Thomas, J. and Kirby, S. (2018) 'Self Domestication and the Evolution of Language', *Biology & philosophy*, 33: 1–30. https://doi.org/10.1007/s10539-018-9612-8

Twomey, T. (2013) 'The Cognitive Implications of Controlled Fire use by Early Humans', *Archaeological Journal*, 23: 113–28. https://doi.org/10.1017/S0959774313000085

Wadley, L. (2023) in T. Wynn, K. Overmann, F. Coolidge (eds.) *Oxford Handbook of Cognitive Archaeology*, pp. 391-410. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780192895950.013.15

Waters, CK. (2007) 'The Nature and Context of Exploratory Experimentation: An Introduction to Three Case Studies of Exploratory Research', *History and Philosophy of the Life Sciences*, 29: 275–284. http://www.jstor.org/stable/23334262

Webb, B. (2000) 'What Does Robotics Offer Animal Behaviour?', *Animal behaviour*, 60: 545–58. https://doi.org/10.1006/anbe.2000.1514

Webb, B. (2020) 'Robots with Insect Brains', *Science*, 368: 244–5. https://doi.org/10.1126/science.aaz6869

Wherry, F. F. (2017) 'Fragments from an Ethnographer's Field Guide: Skepticism, Thick Minimalism, and big Theory', *Ethnography*, 18: 46–56. https://doi.org/10.1177/1466138115592422

Wichmann, F. A. and Jäkel, F. (2018) 'Methods in Psychophysics, *Stevens Handbook of Experimental Psychology and Cognitive Neuroscience, Methodology*, pp. 265–306. New Jersey: John Wiley & Son.

Downloaded from https://academic.oup.com/jole/article/10/1/lzaf006/8362096 by guest on 10 December 2025