Fully automated experiments on cultural transmission through crowdsourcing

Jordan W. Suchow[1,2,3], Thomas J. H. Morgan[1,3], Jessica Hamrick[1], Michael Pacer[1], Stephan C. Meylan[1], and Thomas L. Griffiths[1,2]

The logistics of in-laboratory experiments on cultural evolution, social learning, cooperation, and collective decision-making drive experimental designs towards simple population structures, small groups, and limited interaction between participants. Here we developed Wallace, a software-based tool that automates high-throughput experiments on cultural transmission through crowdsourcing. The tool handles the full experimental pipeline from participant recruitment through data management, enabling experiments that are efficient, reproducible, and unprecedented in their complexity and scale.

The experimental study of cultural evolution, social learning, cooperation, and collective decision-making consider fundamental questions about our capacities to learn, communicate, and decide in a world that is shared with other people. Experiments have revealed, for example, how structured forms of communication emerge from individual learning and decision-making[1, 2], how innovations accumulate in populations to produce technologies that go beyond what any one individual could create[3, 4], and how the format of communication affects the transmission and acquisition of new skills[5, 6]. In-laboratory experiments of this kind are logistically complex and resource intensive, requiring recruitment and coordination of participants to perform tasks sequentially and in concert, with enough space and time to isolate and control their interactions. These requirements drive experimental designs towards simple network structures[7, 8] (Morgan et al., 2014;), small groups, and limited interaction between participants[3].

To address these issues, we created a software-based tool for orchestrating cultural transmission using online crowdsourcing. Our tool, named Wallace, provides high-throughput automation for running behavioral experiments — it recruits participants, obtains their informed consent, arranges them into a network, coordinates their communication, pays them, and then validates and manages the resulting data. Wallace runs on commodity hardware or cloud platforms, communicates by means of its API, uses widely supported languages and markups such as Python, HTML, JavaScript, and CSS, and is released as open-source software under the permissive MIT license (Methods, Supplementary Software).

[1] Department of Psychology, University of California, Berkeley, Berkeley, California, USA. [2] Institute of Cognitive and Brain Sciences, University of California, Berkeley, Berkeley, California, USA. [3] These authors contributed equally to this work. Correspondence should be addressed to J.W.S. (suchow@berkeley.edu), T.J.H.M (thomas.j.h.morgan@gmail.com), or T.L.G. (tom_griffiths@berkeley.edu).

Wallace is modular and includes a library of components that will be useful in the creation of new experiments. Prepackaged network structures include the linear chain[9], scale-free network[10], star and burst formations, micro-society[11], and the discrete generational structure of the Wright–Fisher model from population genetics. Prepackaged behavioral tasks include story recall, category learning, function learning, magnitude estimation, a public goods game, stimulus–response mapping, and numerosity judgment. Custom network structures, processes, and tasks can be built by modifying provided templates.

To validate the efficiency of the tool, we analyzed log data from a large-scale experiment run by it and reported elsewhere. The experiment, which examined genetic encoding of learned behavior via the Baldwin Effect, began with a population of 60 bionic agents — human learners endowed with artificial genes that controlled learning. Each of 39 generations that followed the founding generation was composed of a non-overlapping set of 60 new bionic agents, each of whom inherited (artificial) genetic information from a member of the previous generation, chosen with probability proportional to a fitness measure that tracked performance on the learning task (Methods). This network structure imposes strict dependencies across generations, but permits concurrency within each generation, leading to a time complexity that is linear in the number of generations and constant in the size of each generation (Supplementary Note 5).

Wallace makes efficient use of time, space, and human capital. Using a funnel analysis, we determined that to yield the 2400 participants called for by the design, 3300 needed to be recruited, of whom 90 (2.7%) did not begin the task, 203 (6.2%) began but quit before completion, 87 (2.6%) did not finish within the allotted time, and 460 (13.9%) finished but did not meet the required level of performance, leading to a fractional yield of 77.2% (Supplementary Note). Imperfect yield lengthens an experiment's running time because participants that do not contribute complete, valid data must be replaced. Nested failures, where a replacer needs replacing, are the performance bottleneck in cultural transmission experiments and are particularly troublesome in paradigms such as the Wright–Fisher model, where, because selection depends on relative fitness, recruitment of the next generation is contingent on having completed the parent generation. Wallace uses a combination of techniques to mitigate this issue, including screening for reliable participants, limiting the time allotted to perform the task, and testing for comprehension (Methods).

To validate the extensibility of the tool, we recreated 12 experiments from evolutionary biology, game theory, and psychology (Table 1, Methods).

Table 1. Validating Wallace's extensibility.

Topic of origin	Task	Structure, process, size	Iters.	Citation
Memory & culture	Story recall	10-person transmission chain	1	Bartlett, 1932
Inductive biases in learning	Function learning	10-person transmission chain	5	Kalish et al., 2008
Wisdom of the Crowds	Magnitude estimation	100 people, unconnected		Galton, 1907
Game theory	El Farol Bar Problem	20 people, 10 rounds		
Organizational behavior	Delphi method	5-person panel with 1 overseer		
Social learning	Replacement method	10 people, 4 active at a time		
Language	Telephone game	10-person transmission chain		???
Herding in humans	Numerosity judgment	10-person forward-linking chain		
Baldwinian evolution	Category learning	60 × 40 Wright–Fisher process	125	Morgan et al., 2016
Evolution of social learning	Numerosity judgment	40 × 40 Wright–Fisher process	125	Morgan et al., 2016
Cooperation	Public goods game	40 × 40 Wright–Fisher process	125	Morgan et al., 2016
Design	Stimulus–response mapping	10-person transmission chain		

A growing concern amongst behavioral scientists is reproducibility, which is weakened by unscripted interactions with participants, small sample sizes, vaguely reported methods, unavailable source code, and undocumented data. Guided by the principle that the adoption of best practices can be promoted by the default behavior of technology, Wallace implements best practices from the XXXX organization's guidelines for data handling and management (Methods). The tool's automation and efficiency lessen the burdens that ordinarily hinder reproducibility. For example, because the experiments are run entirely via code, they can be self-documenting, creating as a byproduct shareable packages containing the original source code, a register of all events that took place during the experiment, and the data (Methods). Wallace also offers automated preregistration (Methods).

In conclusion, Wallace is an extensible platform for automating experimentation on cultural evolution, social learning, cooperation, and collective decision-making. The tool makes efficient use of time, space, and human capital, while promoting reproducibility in the behavioral sciences. We anticipate that its greatest potential will be found in its facilitation of paradigms that go beyond small linear transmission chains, leading to the proliferation and mainstreaming of experimental paradigms such as simulating evolution with bionic agents and other forms of human-in-the-loop computation.

[Insert Figure 1 here.]

Methods

Methods and any associated references are available in the online version of the paper.

Acknowledgments

We thank Sally Kleinfeldt, Alec Mitchell, and Cris Ewing for discussions and assistance. This

work was funded by the National Science Foundation (grants BCS-1456709 (to T.L.G) and SPRF-IBSS-1408652 (to T.L.G and J.W.S)).

Author Contributions

All authors conceived the research. J.W.S, T.J.H.M, and J.H. wrote the software. J.W.S performed the analyses. J.W.S wrote the paper with input from T.J.H.M. All authors reviewed the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Verhoef, T., Kirby, S. & Padden, C. in Proceedings of the 33rd annual conference of the cognitive science society 483-488 (2011).
2. Claidière, N., Smith, K., Kirby, S. & Fagot, J. Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society of London B: Biological Sciences* **281**, 20141541 (2014).
3. Caldwell, C.A. & Millen, A.E. Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior* **29**, 165-171 (2008).
4. Dean, L.G., Kendal, R.L., Schapiro, S.J., Thierry, B. & Laland, K.N. Identification of the social and cognitive processes underlying human cumulative culture. *Science* **335**, 1114-1118 (2012).
5. Morgan, T. et al. Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature communications* **6** (2015).
6. Hill, K.R., Wood, B.M., Baggio, J., Hurtado, A.M. & Boyd, R.T. Hunter-gatherer inter-band interaction rates: Implications for cumulative culture. (2014).
7. Flynn, E. & Whiten, A. Cultural transmission of tool use in young children: A diffusion chain study. *Social Development* **17**, 699-718 (2008).
8. Horner, V., Whiten, A., Flynn, E. & de Waal, F.B. Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences* **103**, 13878-13883 (2006).
9. Bartlett, F.C. Remembering: An experimental and social study. *Cambridge: Cambridge University* (1932).
10. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).
11. Jacobs, R.C. & Campbell, D.T. The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *The Journal of Abnormal and Social Psychology* **62**, 649 (1961).

Online Methods

Summary. Details of Wallace and of the reported example experiments are provided here. Additional material, including documentation for the software, is available at http://cocosci.berkeley.edu/wallace.

Code availability and licensing. The code base for Wallace version 1.0.0 is provided as Supplementary Software. It is open source and available under the MIT (X11) license, a permissive free software license. Ongoing development of Wallace and new releases of the source code are hosted on GitHub at https://github.com/berkeley-cocosci/wallace.

Command-line utility. Experiments are managed through a command-line utility, `wallace`, which includes commands to launch new experiments and monitor existing ones (Supplementary Note 1).

Deployment options. The software can be deployed on a local Unix-based server; on Heroku, a cloud platform-as-a-service that makes use of a managed container system (recommended); or on Amazon's EC2 cloud-computing services.

Architecture. Deployed experiments have at their center a web application built on the Flask microframework, which responds according to a versioned RESTful API that returns JSON response messages (Supplementary Note 4). The web application is responsible for responding to requests for content from the participants' front-end clients as well as notifications from participant-handling services such as Amazon's Mechanical Turk.

File formats for storing data. During an experiment, data are stored in a PostgreSQL database, an ACID-compliant object-relational database system. Once complete, data are exported to plain-text files, one for each table in the database. Each file is formatted as comma-separated values (CSV) in accordance with a schema defined in the CSV Schema Language of The National Archives (UK), provided as Supplementary Files 1–8. These files, alongside a readme and unique identifier, are compressed into the ZIP archive file format. An example of a data archive is provided as Supplementary File 9.

File formats for storing code. The `wallace` command-line utility is launched from within a directory of code that defines the experiment to be run. At a minimum, the directory must include a configuration file in INI file format, a plain text or Markdown readme, and a Python file that defines the experiment. Most experiment directories will contain additional files, including a set of HTML templates for consent forms, task instructions, etc., as well as all front-end assets (Supplementary Note 2).

Preregistration. To achieve preregistration, Wallace first verifies that the archived code contains a statement declaring any planned analyses. It then uses the SHA512 cryptographic hash function to compute a hexadecimal digest of the code archive at the time the experiment was run. Finally, the digest is uploaded to a publicly viewable webpage hosted by the Open Science Framework, where it is time stamped (Supplementary Note 3).

Objects. Wallace manages eight kinds of objects, described in Table 2.

Table 2. Kinds of objects managed by Wallace.

Kind of object	Role
Network	A set of nodes and vectors between them
Node	An individual
Vector	A connection from one individual to another
Info	Information that is transmitted
Transmission	
Transformation	
Participant	
Notification	

Measures to increase efficiency and data quality. Wallace achieves a speedup through a kind of apoptosis that replaces participants who do not complete the task within the allotted time, typically set at 2–3× the expected completion time.

Sourcing participants. Participants are recruited through Amazon's Mechanical Turk (MTurk), an online labor platform where people perform short tasks for pay[12-14]. With MTurk, it is possible to limit recruitment to participants from a particular geographic region. When recruitment is limited to the United States, the demographics of workers are fairly representative of the population of US internet users, though on average they are younger, have lower income, are more educated, and include more females. [Fill out description of MTurk and describe anything relevant to running experiments on it.] Experiments were approved by the Committee for Protection of Human Subjects at the University of California, Berkeley and carried out in accordance with the approved protocols.

Ensuring high-quality data. A concern regarding data quality often arises when conducting experiments by means of online crowdsourcing. With a new medium comes a new set of best practices for ensuring quality data. Unreliable participants are excluded from recruitment by requiring a minimum reputation of 95% approval on previous MTurk tasks[15].

Recruiting participants. The logic of participant recruitment is determined in part by Wallace and in part by the experimental design. The number of participants that are initially recruited is determined by the experimental design — whereas a chain starts with a single individual, a Moran process starts with an entire generation. Occasionally, a participant must be recruited to replace an existing participant. This may occur because the original participant did not begin the task, quit before completing it, did not complete it in the allotted time, finished but did not meet the required level of performance, or experienced some kind of technical error during the course of the experiment. In these cases, Wallace automatically recruits a new participant and updates its database to exclude the replaced participant from the ongoing experiment. Eventually, all the participants needed for that stage of the experiment are done, at which point a new batch of participants is recruited, as defined by the experimental design.

Compensating participants. Participants are compensated immediately following their completion of the task. Wallace allows custom logic defining the amount of compensation, making possible performance-based bonuses that depend on a participant's behavior, as is necessary for example in many experiments on cooperation.

Time complexity of Wallace-based and in-lab experiments. When running an experiment in the lab, the... An experiment's *best-case execution time* is the minimum length of time needed to complete the experiment given the dependency structure of the experiment and the time needed by a participant to complete the task. Analyze the time complexity for an in-lab vs. Wallace experiment.

Describe the README linter.

12. Paolacci, G., Chandler, J. & Ipeirotis, P.G. Running experiments on amazon mechanical turk. *Judgment and Decision making* **5**, 411-419 (2010).
13. Buhrmester, M., Kwang, T. & Gosling, S.D. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* **6**, 3-5 (2011).
14. Paolacci, G. & Chandler, J. Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science* **23**, 184-188 (2014).
15. Peer, E., Vosgerau, J. & Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* **46**, 1023-1031 (2014).