Trading Accuracy for Enjoyment? Data Quality and Player Experience in Data Collection Games

David Gundry
University of York
York, UK
deg500@york.ac.uk

Sebastian Deterding
University of York
York, UK
sebastian@codingconduct.cc

ABSTRACT

Games have become a popular way of collecting human subject data, based on the premise that they are more engaging than surveys or experiments, but generate equally valid data. However, this premise has not been empirically tested. In response, we designed a game for eliciting linguistic data following Intrinsic Elicitation – a design approach aiming to minimise validity threats in data collection games – and compared it to an equivalent linguistics experiment as control. In a preregistered study and replication (n=96 and n=136), using two different ways of operationalising accuracy, the game generated substantially more enjoyment (d=.70, .73) and substantially less accurate data (d=-.68, -.40) – though still more accurate than random responding. We conclude that for certain data types data collection games may present a serious trade-off between participant enjoyment and data quality, identify possible causes of lower data quality for future research, reflect on our design approach, and urge games HCI researchers to use careful controls where appropriate.

CCS CONCEPTS

• **Applied computing** → **Computer games**.

KEYWORDS

Applied Games, Games with a Purpose, Human Computation Games, Crowdsourcing Games, Human Subject Data, Validity

ACM Reference Format:
David Gundry and Sebastian Deterding. 2022. Trading Accuracy for Enjoyment? Data Quality and Player Experience in Data Collection Games. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3502025

1 INTRODUCTION

Across human-computer interaction (HCI), user and market research, human resources, citizen science, and the behavioural sciences, we regularly elicit data from human participants. Unfortunately, standard methods like surveys or experiments are often boring, which can harm participant retention and study completion [5, 39], data quality [18, 42, 57, 60], and can pose ethical concerns [17, 66].

Applied games have become increasingly popular to make data elicitation tasks more enjoyable and less boring [74], a practice that has been variously called game-based methods [62], gamifying research [21], or data collection games [29]. Such *data collection games* are whole games designed with the primary purpose of collecting data from their players, as opposed to gamifying data collection, i.e. adding design elements from games to surveys, online experiments, etc. [29]. Data collection games have become especially popular in citizen science, be it to have human participants collect and analyze data about the natural world (as with *FoldIt* [12] or *Galaxy Zoo* [25]), or to collect data about the human participants themselves, as with *Sea Hero Quest*, which collects a mass online sample of people's spatial navigation ability [14].

Data collection games are grounded in the dual premise that games (a) are more enjoyable than standard methods while (b) producing data of similar if not better quality [32]. Enjoyment here encompasses different conceptualisations of positive user experience (flow, intrinsic motivation, etc.) that are believed to drive behavioural engagement [58] – in our case, participating in and completing more studies, providing more and more complete data, exerting more effort and care in responding. With data quality, we here refer to general "fitness for use" [65], which in human participant research centrally involves validity – the extent to which the data supports the inferences we draw from it [48], impacted by validity threats like confounds, response biases, or low generalizability [11]. Other common aspects of data quality raised in human subject research are accuracy (especially in performance tasks), satisficing and careless versus careful responding, study completion versus dropout, or missing/dropping response items; these can be framed as behavioural engagement or validity threats [32].

1.1 Enjoyment and Data Quality in Gamified Research

Research and practice have treated these two premises — enjoyment and data quality — as largely separable concerns, the standard approach being to use games or game design elements to motivate more participation, and then use separate methods to ascertain and filter out valid data from the resultant larger data pool [28].

Thus, there is ample work on *enjoyment* in data collection games, probing what motivates people to engage, especially in the context of citizen science [15, 16, 36, 55, 67], and on how to design such games in a motivating and enjoyable manner [13, 15, 50, 74]. Work on *data validity* or quality has been sparser (see [29] for a review). Researchers and practitioners have developed proven frameworks for data that can be validated against a ground truth,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502025

such as automatically assessing how well a provided datum satisfies computable constraints [12], or establishing intersubjective consensus [73]. However, these frameworks are limited to data with an external ground truth. The situation is different for *human subject data*, data about particular individuals, especially where such data concerns latent, subjective, not directly intersubjectively accessible or verifiable properties such as personal preferences, attitudes, beliefs, experiences, or dispositions (e.g. [6, 46, 49, 52, 59, 63]). Here, data quality, accuracy, or validity means correspondence with either deliberate, conscious beliefs (such as voting intentions), which requires participants to respond honestly and carefully; or with nondeliberate, non-conscious properties (such as implicit biases), which require participants to respond 'spontaneously', with as little deliberation as possible. For data collection games focusing this kind of human subject data, there are no proven validation frameworks, nor is there good data on their comparative validity [28]. In this study, we intentionally focus on such human subject data as the most difficult case of data collection games.

Empirical research directly testing both premises (game design brings (a) higher enjoyment and (b) equal or better data quality) has been largely limited to *gamified* online surveys and experiments – studies that use presumed-motivating design elements from games, rather than creating full-fledged data collection games [22]. This research shows mixed results. While some found that adding game design elements can increase both data quality and quantity [68], or collect data of equivalent quality and quantity, but with more enjoyment [26, 32, 43], recent reviews of gamified surveys [40] and assessments [45] suggest that gamification tends to improve the user experience (i.e. enjoyment), but not necessarily impact behaviours such as satisficing, omitting items, or abandoning surveys, and with those, data quantity and quality.

1.2 Untested Validity Threats and Enjoyment Claims for Human Subject Data Collection Games

Similar empirical work on full-fledged data collection games has been missing. In our literature review, we found one case study [15] and one comparison between a gamified and full game variant of the same citizen science data classification task [55], both suggesting that the full-fledged game produces similar engagement but lower-accuracy data. But either study features no real experimental control.

This lack of high-quality evidence on the data validity of data collection games matters, as prior work [28, 29] has pointed out, because games *as games* generate *new, systemic validity threats*, especially for human subject data: The social norm and empirically observed reality of entertainment gaming is that players ought to voluntarily participate for the sake of (mutual) enjoyment, and to this end, relegate game-external consequences and concerns [19] and make more or less rational, strategically optimal moves [33] – modulated by other norms like maintaining good relations with the other players [38]. This suggests possible trade-offs between engagement and data quality: If players are playing a game *as a game*, they should not distract themselves with game-external concerns like answering honestly and carefully; and if they are answering honestly and carefully, this may diminish the enjoyment of getting fully engrossed in the game. Put differently, the most strategically optimal or fun in-game action need not be the most honest and considered response out-of-game [28]. To give a practical example for latent subjective properties: Assume we design a charade-style social guessing game to elicit people's preferred ice cream flavours. If people really 'get into the game,' they may claim that they like chocolate even though they actually prefer woodruff – because chocolate is easier to mime and guess, or because the previous person already mimed woodruff and repeating them would be boring. We would not expect similar effects from a gamified survey that was still framed and approached *as a survey*, but e.g. clothed into a nautic theme, juicy feedback, or additional game mechanics that don't connect to data collection [43].

1.3 The Present Study

If data collection games were indeed prone to suffer from lower data validity, especially for human subject data, this would add an important caveat to their current popularity, and suggest that research and practice should look into design strategies for improving data quality and validity. In response, we decided to test *how the enjoyment and data quality of a human subject data collection games compares to a practice as usual control.*

To this end, we designed *Adjective Game*, a browser game to elicit adjective order – the order in which people use adjectives to describe a phenomenon, like "big black cat". Established experimental methods for eliciting people's adjective order offers a good 'practice as usual' control. Further, while individual grammatical intuitions about adjective order are a latent subjective property, they are highly predictable for adult native speakers, which provides a rare opportunity to compare data from game and control conditions to an approximate 'ground truth'. The design of *Adjective Game* followed the Intrinsic Elicitation design approach [28]. To our knowledge, it is the only design approach expressly devised for minimising validity threats in human subject data collection games. While this approach has not been empirically validated, we consider designing a game following it the current best case scenario, akin to "maximal positive controls" [34]: If data collection games can elicit human subject data of equal quality while being more enjoyable, we should best be able to observe this with a game intentionally designed for this purpose, and also get a more informed idea about likely effect sizes of differences.

We conducted two pre-registered studies (n=96 and n=136) that compared our data collection game with an equivalent linguistic experimental setup.[1] Each study operationalised accuracy differently, mirroring two different paradigms in linguistics. In both cases, the game proved significantly more enjoyable than the control, but also produced significantly less accurate data. We discuss ramifications for future research and design and reflect on the Intrinsic Elicitation approach.

The rest of this paper is structured as follows. The Background (2) will introduce our linguistic case study, adjective order and picture description tasks, followed by the Intrinsic Elicitation approach. The Materials section (3) presents the design of our game, how it follows the Intrinsic Elicitation model, and the picture description

[1] All materials, code, data, and pre-registrations for both studies can be found at https://osf.io/jac6s/

control task. Sections (4) and (5) report studies 1 and 2, followed by a general Discussion (6) and Conclusion (7).

2 BACKGROUND

2.1 Eliciting Adjective Order with Picture Description Tasks

2.1.1 Adjective Order. The order of words we use expresses shared rules of grammar. For example, in reference to a sea-going vessel that is large and scarlet, an English speaker would say it is a "big red boat", but not (in a neutral context) a "red big boat". This is an example of adjective order. People's intuitive judgments about adjective order are strong and reliable [4, 10, 64]. Different paradigms in linguistics offer different explanations for this fact: A *generativist* would take it as evidence of shared innate universal rules [8, 9]. In contrast, a *constructivist* would argue that people's individual grammatical intuitions are indeed individual, but became coordinated with those of other speakers in the course of socialisation.

Either way, linguists studying a language like English (and native speakers speaking it) can predict other speakers' intuitions about adjective order with high accuracy. This makes adjective order an ideal case study for our purpose, because it gives us a rare approximate 'ground truth' for a latent subjective property (individual adjective order intuitions) against which we can compare data elicited by both a game and a standard data elicitation task. Mirroring the two linguistic paradigms mentioned above, we will do so using alternatively an assumed generativist universal grammar (study 1) and a person's separately elicited individual grammaticality judgments (study 2).

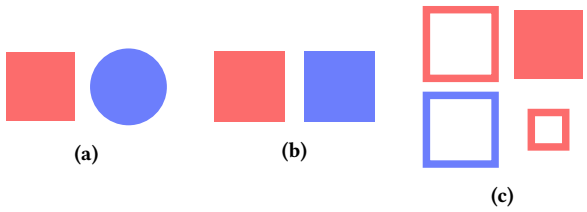


Figure 1: Simplified examples to illustrate contrasts in a picture description task

2.1.2 Picture Description Tasks. One common linguistic method for getting participants to produce language data revealing adjective order is the picture description task [2, 24]. Here a picture (or video, object, etc.) is shown to participants and they are asked to describe it. This can happen in various degrees of pre-structuring. If we present a shape and just ask "What is this?", we have a relatively unstructured elicited production task. By adding structure, we can make it more likely that certain forms of language (like sequences of multiple adjectives) are used. One effective way to do this is to use contrasts [23]. To identify one shape in contrast to another, e.g. figure 1b, a participant might simply say, for example, "square". However, to distinguish between two squares that differ only in colour, as in figure 1a, the participant might say "red square", or "blue square". In this way, a context can be constructed such that a phrase of arbitrary length might be elicited, such as figure 1c, which might elicit "big empty red square". Such a string of adjectives followed by a noun is called a modified noun phrase. In our case study, we will use this language elicitation paradigm as our control condition, since it is an easily implemented and replicated paradigm that is well-established and thus ecologically valid in linguistics.

2.2 Design Approach: Intrinsic Elicitation

To design a data collection game that elicits adjective order, we followed the Intrinsic Elicitation approach [28], grounded in an extended rational choice model of player action, building on prior work by Jonas Heide-Smith [33]. Following this Rational Game User Model (figure 2), a game user chooses in-game actions that maximise their expected total intrinsic and extrinsic utility, where *intrinsic utility* captures payoffs gained from gameplay (such as enjoyment or meaning), usually studied in games research, and *extrinsic utility* payoffs and costs outside of gameplay, usually studied in survey and experimental design research, such as social desirability, being paid to participate, or the effort involved in taking a particular action. The major factor driving intrinsic utility in this model is the enjoyment afforded by experiences of in-game progress, achievement, and competence, which arises from making rationally optimal in-game moves that maximise expected in-game *virtual utility* – or put plainly, chances of winning.[2]

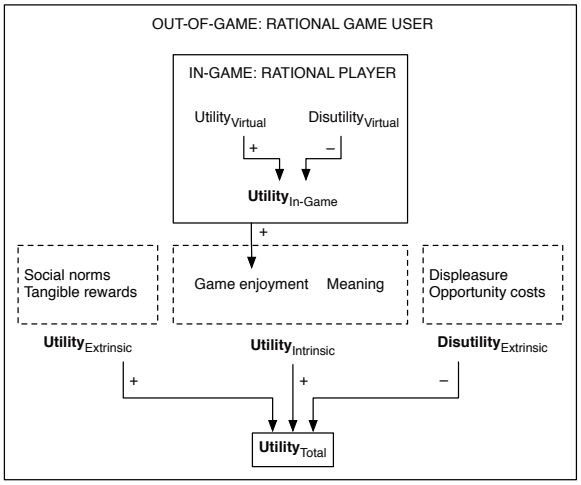


Figure 2: The Rational Game User Model [28]

For design, this model implies that to elicit a certain type and quality of data, providing such data should be the most rational, utility-maximising way to play the game, while different, equally valid responses should not differ in overall utility. Put differently, as with the educational game design principle of intrinsic integration [30], eliciting valid data should be integrated into the core game mechanic or loop, which the Intrinsic Elicitation approach expands into three design principles: necessity, centrality, and veracity.

[2]There are of course many other psychological, social, and material factors influencing what action a player takes at any given moment. This approach embraces a rational choice model as a *useful abstraction*, one that is notably already in use in game design in practices like game balancing, and has some empirical support [33]. See [28] for an extended discussion.

Necessity. Players will only engage in data provision if this effects a change in the game state. The designer must therefore ensure that providing the desired type of data is *necessary* or inherent to actuating one or more game mechanics – the verbs or methods used by game players to change the game state [61]. For example, imagine a *Space Invaders* [51] clone designed to elicit the vocal pitch of the player's speech; one way of making such vocal pitch data *necessary* would be to make the primary move mechanic in the game voice-controlled.

Centrality. Baking data provision into mechanics would be no use if players rarely invoke these mechanics. While multiple factors affect this, one is easy to design for: the virtual utility of that mechanic. If a mechanic is regularly the only and/or strategically optimal choice, it is *central* to the game. E.g. voice-controlled movement of the laser cannon in our *Space Invaders* clone is both necessary and central: at most given moments, moving your laser cannon is the best and only option. If you instead controlled movement by joystick and added a voice-controlled "rebuild bunkers" mechanic where once a game, players could say "rebuild bunkers" to reinstate bunkers shot down by aliens, this data-eliciting mechanic would satisfy Necessity, but not be central.

Veracity. There are generally multiple ways of actuating any given mechanic, which can differ in virtual utility (e.g. taking 2 not 3 cards is a better move), intrinsic utility (e.g. saying "dare" in a round of *Truth or Dare* may be more fun for you), and extrinsic utility (e.g. making one move might require greater physical and cognitive effort than another). Players will actuate mechanics to maximise their combined total utility. Therefore, actuating a mechanic in a way that provides valid, honest, or *veracious* data should be the option the player perceives as utility-maximising – or there should at least be no other way of actuating it that would be of higher total utility. E.g., if we want to elicit spoken words with a game where you voice-control an aeroplane, if the aeroplane is controlled by pitch, not spoken words, players are likely to fall back on just humming at different pitches than continually producing words, since just humming is less cognitively effortful. Or if we want to elicit preferences about people's faces with a game where people choose between two faces presented at a time, and the game rewards speed with a competitive race (first to go through all faces wins), this would give fast-but-careless responding a higher virtual and possibly intrinsic utility than desired slow-but-careful responding, also violating the veracity principle.

The veracity principle is predicated on the premise that *all else being equal*, people are biased to respond accurately or honestly. While cross-cultural evidence shows that adults often behave dishonestly *if there is an incentive to do so*, it also shows that adults across cultures consider honesty an important personal and social value [35]. Importantly, data collection games are subject to the same general response biases and reactivity issues as any human subject research, such as acquiescence (yes-saying), social desirability (responding in a way that presents oneself in a more socially desirable light), or demand characteristics [11, 27]. The veracity principle acknowledges and incorporates these as extrinsic utilities, and expands on them by highlighting that data collection games *as games* introduce a new systemic response bias toward responding that maximises virtual (in-game strategic value) and intrinsic utility (enjoyment). Thus, even if we design a data collection game that equals out the virtual and intrinsic utility of available responding options, we cannot expect perfect honesty or accuracy, as all the other known response biases and wider validity threats and sources of measurement error will still impact the data collected, pointing to known general methods for reducing these [11, 71].

3 MATERIALS

3.1 The Data Collection Game

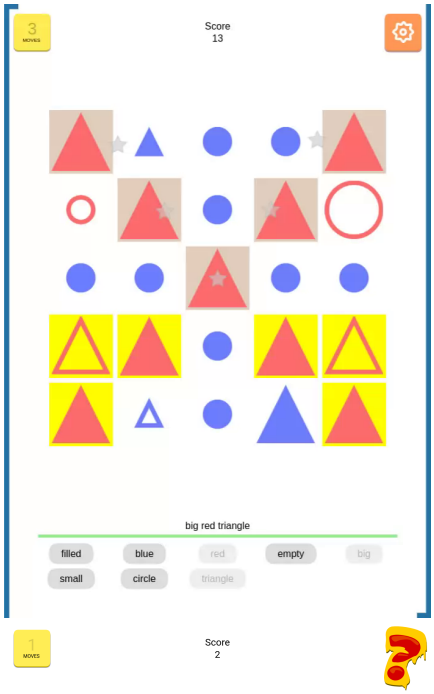
To maximise participant reach, we built our *Adjective Game* as an HTML5/JavaScript game, playable in-browser both on desktop PCs and smartphones. Following Intrinsic Elicitation, we started designing the game by identifying a core mechanic that would necessarily generate modified noun phrases while staying as close as possible to the picture description task. This resulted in a casual puzzle game similar to tile-matching games such as *Two Dots* [37] or *Candy Crush Saga* [41], albeit with a novel data-eliciting input mechanic (see figure 3).

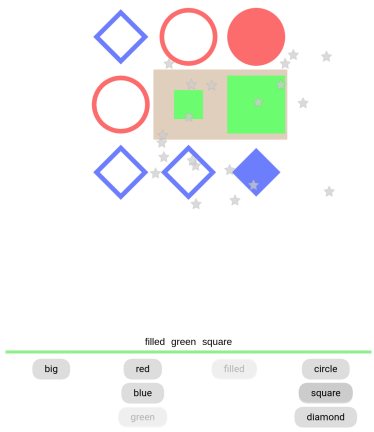
3.1.1 Gameplay. The game presents a series of levels of increasing difficulty, each consisting of a grid filled with blocks that have various shapes, colours, sizes, and filling. The goal of each level is to clear all blocks in a given number of moves. To clear blocks, players enter a string of exactly three words that must contain exactly one noun. Players enter strings by tapping/clicking the labelled buttons at the bottom screen, which show the permitted words for the level – these are nouns describing possible shapes (circle, square, triangle), and adjectives for the possible colours (green, blue, red), sizes (small, large), and filling (empty, filled).

The string entered so far is displayed on the screen. Once a player has constructed a three word string, all blocks currently visible on screen that match this string are cleared simultaneously. It makes not difference in which order words are entered, e.g. players could input "triangle empty small" to clear small empty triangles. Only strings that don't contain a noun or don't identify any blocks visible on the screen are rejected by the game. The player can undo partial inputs. Entering a valid string expends one of a limited number of moves.

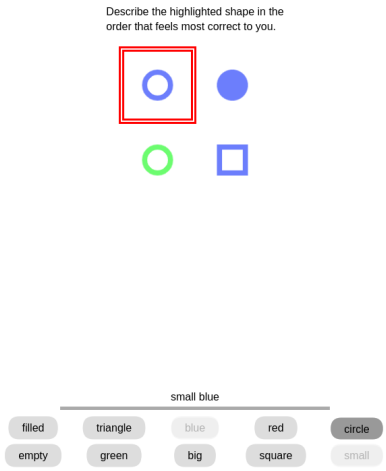
When blocks are cleared, the remaining visible blocks fall down to fill spaces in the screen grid, and new blocks fall in from above to refill the grid. This continues until the total number of blocks for a level is hit. Clearing groups of orthogonally adjacent blocks is worth more points than clearing the same number of isolated ones. If a cleared group contains three blocks or more, the player earns a bonus move. The bigger the size of the cleared group, the higher the score bonus and number of bonus moves.

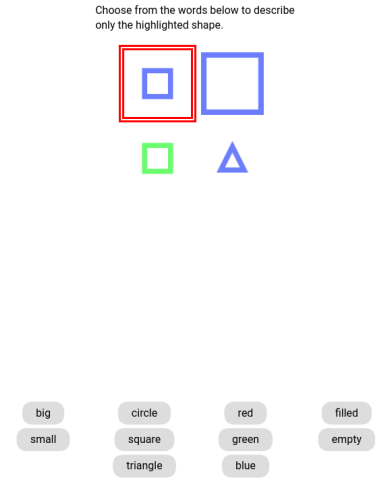
It is possible to perform better or worse at clearing a level (the score achieved upon clearing a level is translated into a three-star rating), and most levels are impossible to clear without earning bonus moves. This invites and requires strategic planning ahead to identify sequences of groups of blocks that can be efficiently created and cleared to maximise score and bonus moves. On each turn, the player thus has to balance simply maximising the number of blocks cleared, manipulating the board to create and clear groups, and tracking how many moves they have left.





(a) Game interfaces from the first (top) and second (bottom) study during play, shown upon completing an input. Players assemble a three-word string by tapping/clicking word bubbles at the bottom of the screen. Shapes identified by this string are then cleared from the board above, with bonuses for contiguous groups of shapes. This triggers multiple forms of feedback: a bar filling and turning green when completing a word string; matching shapes are highlighted; the score at the top is increasing; the number of moves left is increasing (top) and decreasing (bottom) depending on the size of groups cleared; a bar around the edge of the screen fills corresponding to progression through the level; and stars have burst out of the cleared group. Highlighted shapes will disappear and the remaining shapes will drop down into empty spaces, with new shapes falling from above.





(b) Experimental control interfaces from the first (top) and second (bottom) study. A target shape, always the top left, is indicated by a double red box in a 2x2 grid of shapes. Each other shape differs from the first in one dimension: shape, size, colour, or filling. Participants again tap/click on word bubbles in the bottom screen to assemble a word string identifying the target shape. In study 1 (top), word bubbles are arranged in a grid, in study 2 (bottom), word bubbles are arranged in columns by type.

Figure 3: Screenshots from the game and control conditions.

3.1.2 Realising Intrinsic Elicitation. How did *Adjective Game* realise the three design principles of Intrinsic Elicitation? In terms of *necessity*, our input mechanic involves selecting a sequential order of two adjectives and one noun – the building blocks of a modified noun phrase. Players cannot enter their own words, nor can the mechanic be triggered by any other input. This makes producing a modified noun phrase a necessary part of the game's mechanics.

Inputting three-word adjective and noun strings is also *central* as there are no other mechanics available to the players, making it the one with the highest virtual utility by default. Had we allowed players to clear blocks with just one or two-word strings ("square", "green circle"), these would on average have cleared far more blocks and thus would have had a higher virtual utility and centrality.

Veracity required the most consideration and iteration by far. To do so, we worked systematically through the different virtual, intrinsic, and external utilities proposed by the Intrinsic Elicitation model (meaning, social norms, opportunity costs, etc.).

First off, we chose not to prescribe, hard-code, display, or reward a particular word order (e.g. that the noun comes last). Had we done so, this would have rendered the mechanic inexpressive of the players' own grammatical intuitions of adjective order. This would be comparable to running a multiple-choice questionnaire about voting preferences and telling participants which party to select.

Second, we quickly realised that we couldn't change the board state or give any dynamic feedback in response to any 'partial' input of just the first or second word, such as highlighting presently selected words. This would have created an informational virtual utility in starting with particular inputs to probe and explore the game state for potential combinations.

Third, while we decided that each valid input must have a noun, we ensured that in the levels we designed, no single type of adjective (colour, shape, filling) is always required. This not only adds an interesting asymmetry to the game, but also makes it more likely that in the course of play, players will produce a richer range of different adjective orders, as different orders will be more strategically optimal with changing board states.

Finally, we spent significant time and effort to find ways to control actuation effort as a potential disutility. We particularly tried to avoid any order effects of button arrangements that make certain word sequences easier or harder to input, as eye, mouse, and/or finger have to travel different distances from one word to the next. We thus chose to randomise the order in which buttons appeared in every level. This left some differences in actuation effort, but ones that would be randomised out in the final data. (In study 1, the order of buttons is entirely random. In study 2, buttons are grouped into columns by type (noun, colour, size, etc.), but the order of columns and the order within columns is random.) We also considered whether and how to allow players to reset or undo partial or incorrect inputs. Allowing partial clearing of an input string back to front, as with a 'backspace' key, might have biased players to retain words entered first. The game therefore only allows you to clear the entire input at once.

3.2 The Experimental Control Task

We designed the experimental control task to (a) be as close and ecologically valid to the standard linguistic contrastive picture description task [23], while also (b) minimally deviating from the data collection game in all non-game aspects of the interface and interaction. The task can be seen in figure 3 in the right column.

To this end, we reused the HTML5/JavaScript framework of the game, retaining styling, layout, and core interaction, while removing all immediate 'gamy' interface features (e.g. juicy feedback like exploding stars), core structural game features (goals, levels, progress feedback), and emergent game dynamics and aesthetics (core challenge, increasing difficulty, uncertainty). We also reworded all text speaking of "game" and "play" with equivalent phrases like "experiment" or "interact". We summarise all differences between game and control in table 1.

The control task presents screens that each require a single input. Each screen shows four blocks, one of which is indicated by a double-lined red box (always the top left). The three remaining blocks all differ from the first in only one aspect (colour, shape, etc.) to ensure a three-word description is appropriate and needed to contrastingly describe the target. Users still input the three-word string by clicking/tapping on word buttons, as in the game. Only an input selecting the target block is accepted; incorrect inputs trigger a prompt to the user to select the highlighted block. When the target is selected, the next task starts automatically, without the blocks moving or being cleared.

The control task thus matches a standard contrastive picture description task for eliciting adjective orders. It does so with the same core interaction (and actuation method) as the game. However, as the player does not need to choose the shape, and each trial is separate from the others, all strategic considerations are removed. Similarly, uncertainty is reduced as no unseen new shapes fall down to replace those removed. While new shapes appear each trial, the arrangement is always similar and does not express any properties of strategic interest. Challenge is much reduced as the visual search of shapes is much reduced, as is the strategic challenge of selecting the optimal description to use.

The control condition has a single tutorial screen which shows the participant a single shape (a filled red circle) and presents three words in a random order: 'filled', 'red', and 'circle'. It gives brief instructions to the participant what to do. After selecting all three words in any order, the tutorial is complete.

4 STUDY 1

Our first study compared *Adjective Game* and control task in terms of participants' self-reported enjoyment and accuracy (operationalising data validity). Based on prior work and the Rational Game User Model, we hypothesised that:

- **H1** Players experience more enjoyment in the game condition than the control.
- **H2** Accuracy is lower in the game condition than the control.
- **H3** Accuracy in the game condition will be higher than expected by random chance.

A preregistration can be found at https://osf.io/hab82/, along with a repository containing all study materials, code, and data at https://osf.io/u2nze/. The study received ethical approval from our departmental ethics board at the University of York.

Feature	Game Condition	Control Condition
Tutorial	3-level tutorial introducing 1) single word input, 2) block falling mechanic 3) strategic choice in description with three-word inputs	1 level tutorial introducing three-word input for a single shape
Tutorial Instructions	"Matching blocks disappear / Select a word below", "You've got to clear each level in a limited number of moves. / Clear 3 adjacent blocks for a bonus move", "Only blocks that match every word are cleared"	(ex. 1) "Choose from the words below to describe the shape / Think carefully and use the order that feels most grammatically correct to you" (ex. 2) "Choose from the words below to describe the shape"
Instructions Per Trial	(ex. 1) None (ex. 2) Help button ("Choose a 3 word phrase to clear blocks", "Only blocks matching every word are cleared", "Clear groups of 3 to get an extra move (group of 4 = 2 extra moves, etc.)")	(ex. 1) "Describe the highlighted shape in the order that feels most correct to you" (ex. 2): "Choose from the words below to describe only the highlighted shape."
Gameplay		
Mechanic	Enter 3 words to identify one or more blocks	Enter 3 words to identify a single block
Game-specific Goal	Clear screen of blocks	None
Failure Condition	Run out of moves	None
Strategy	Identifying larger groups of blocks. Clearing blocks to create groups of blocks next turn. Gaining bonus moves.	None
Scoring	Current score displayed. Score increases for clearing blocks. Larger groups cleared increases the score more	None
Challenge	Visual search of blocks and buttons. Planning/predicting multiple turns	Blocks always in same arrangement. Visual search of buttons.
Uncertainty	Initial arrangement of level. New blocks from above per move	Arrangement of next trial per move
Interface		
Input	(ex. 1) Randomised grid of buttons (ex. 2) Randomised columns of buttons	As game
Graphics	Simple coloured shapes.	As game
Input feedback	Move successful. Illegal move.	As game
Game-board Feedback	Particle effects on blocks cleared. Level progress indicator.	None
Interstitials	Level start and end dialogues with numerical and three-star scores	None

Table 1: Comparison between game and control conditions

4.1 Method

4.1.1 Materials. We used the *Adjective Game* and Control Task described above as materials.

4.1.2 Dependent Variables. In this first study, we operationalised accuracy in adjective order following a generativist linguistic paradigm, according to which there is a single shared true English grammar that native speakers have access to. Thus, we counted participant inputs as accurate when they conformed with English grammar and inaccurate otherwise, proposing a correct adjective order for our task of *size, filling, colour, noun*. For each participant we calculated accuracy as the number of accurate inputs as a proportion of the number of recorded inputs. It is this proportion that was used in comparing accuracy between conditions. An example set of resulting judgements is provided in (1), with ungrammatical forms prefaced by an asterisk. We welcome readers to compare their own judgements. When doing so, consider each phrase in a neutral context and without special intonation.

(1) a. big red circle
b. big empty circle
c. big filled circle
d. *red empty circle
e. *red filled circle
f. *red big circle
g. filled red circle
h. *filled big circle
i. empty red circle
j. *empty big circle

We operationalised enjoyment using the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory[3], a well-established

[3]https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

5-point likert scale that is frequently used in games HCI to assess player enjoyment [47].

4.1.3 Sample. Powered to detect a difference in enjoyment with an effect size of d=.5, using an alpha of 0.05 and power of 0.8, our power analysis suggested a minimum sample of 50 participants per condition. We recruited 100 adults with the first language of English via Prolific, using the demographic filters provided by that platform. The study offered £1.00 for completing a 10 minute task (£6 per hour) entitled "A study where you describe shapes". After exclusions, 96 participants completed the study, 47 in the game, 49 in the control condition[4].

4.1.4 Procedure. Participants completed a short demographics questionnaire which included their age, gender, whether their first language was English, and gaming experience. They were then randomly assigned to either a game or a control condition. In each case, the participants continued until they had supplied 20 complete inputs, excluding inputs made in the tutorial section of each condition. At the end of the experiment (once players had made 20 successful inputs), participants completed the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory.

4.1.5 Analysis. While the pre-registered analysis used t-tests, following reviewer requests, we report non-parametric equivalents below. This change to a more conservative statistical test fitting our non-parametric data patterns made no difference to the direction of the results; the original t-test results can be found at https://osf.io/u2nze/. Analysis was conducted in Python 3.9.7 [69] using SciPy [72], Pandas [54], Numpy [31] and raincloud plots [1]. Power analyses were performed in R 4.1.2 [56] using the pwr package [7].

4.2 Results

4.2.1 Demographics. Out of 96 participants, 60 reported their gender as female, 36 as male. The median age was 27, with ages spanning from 18 to 58. One participant who had presumably mistakenly entered their age as 130 was also included in the analysis. Most participants reported playing digital games frequently, with 65 (68%) playing at least several times a week. Only 19 (20%) participants reported playing once a month or less frequently.

4.2.2 Enjoyment. We used a one-tailed Mann-Whitney U test to see if enjoyment is greater in the game than control. Enjoyment was significantly greater in the game (M=3.70, SD=0.83) than control condition (M=3.09, SD=0.93), U=1590.5, p<.001, d=.70, see figure 4.

4.2.3 Accuracy. A two-tailed Mann-Whitney U test found that the game elicited significantly less accurate inputs (M=0.45, SD=0.35) than the control (M=0.67, SD=0.30); U=734.5, p=.001, d=-.68, see figure 5.

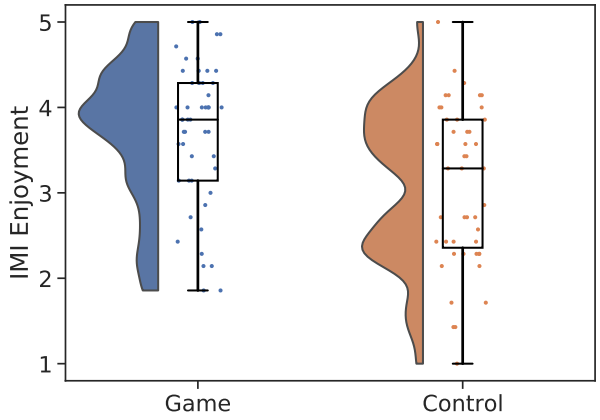


Figure 4: Enjoyment as measured by the IMI Interest/Enjoyment subscale is higher for the game than control.

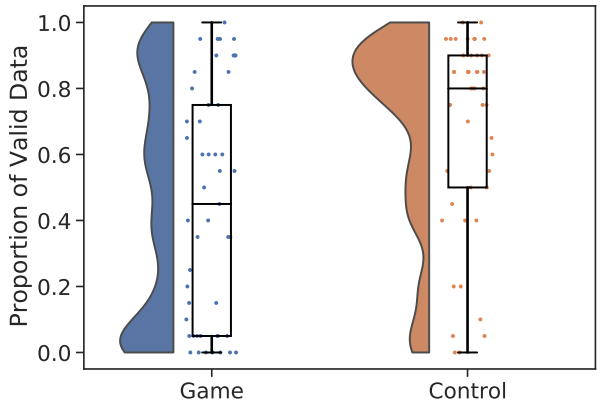


Figure 5: Accuracy as share of standard English grammar-conforming word-orders in total word-orders collected is higher in the control than game condition

To determine whether accuracy in the game condition was higher than expected by random chance (H3), we used a two-tailed Wilcoxon test to compare data quality for the game condition against a theoretical random player mean. The proportion of word orderings that would be correct based on completely random answering can be calculated as the proportion of the correct word orders for any given input out of the total possible word orders. For this, we are looking at the proportion of grammatical inputs out of "potentially mechanic actuating inputs", a subset of recorded inputs. This is necessary for comparison with our theoretical player. The game can only ever be triggered by inputs of the correct form. Firstly, a hard requirement is that they contain a noun. Secondly, because the adjectives are mutually exclusive, only a single adjective of a given category can be used in a single description. Therefore, our theoretical player who gives functional

[4] 2 participants were excluded and their data deleted because they reported an age of under 18 (as their reported ages were -129 and -131, this may have been an input error). 2 participants were excluded because they reported their first language as other than English during the study. Both of these exclusions were in line with the experiment's preregistration. The above sample size excludes the following: Data was not correctly recorded for 2 participants, suggesting they did not complete the study. 2 data records were not associated with a Prolific ID so these were deleted in line with our ethics application. Finally, due to a mistake configuring the study, we over-sampled participants. As, following anonymisation, we would not be able to identify the 'extra' participants (to support re-analysis in the case that significance was extremely close), data from the over-sampled participants was deleted during anonymisation. Anonymisation was performed immediately upon study completion.

inputs, but behaves purely randomly where it does not have a mechanical impact would always follow this constraint. Note that this considers inputs possible with the game and not the inputs likely to be necessary in the first 20 moves.[5] Taking all this into account, there is a single correct word order for an input and six possible permutations of 3 words, making $\frac{1}{6}$ or 16.67%.

There was a significant difference in the scores for the game condition (M=0.48, SD=0.37) compared to the theoretical expected value; w=154.5, p<.001, d=.84. More grammatical inputs were elicited than would be expected from our theoretical random player.

4.3 Discussion

In line with our hypotheses, the game showed higher enjoyment (H1), but lower accuracy (H2), though higher accuracy than expected from a random player (H3). The effect size for accuracy differences between game and control was surprisingly large (d=-.68). We therefore reviewed our study to find potential confounds and alternative explanations for this large effect size that we could control for in a conceptual replication. First, while native speakers of the same language are widely considered to be consistent in adjective order, this claim may not be convincing to all readers. Some may take a constructivist stance that participants' individual intuitions of adjective order did not align with the ideal grammar we used. Hence, we decided to compare adjective order in the second study with a separate elicitation of participants' own grammatical intuitions. Second, we found that time played differed significantly between the game (M=348.72, SD=96.72) and control (M=270.17, SD=209.84) conditions; U=469.0, p<.001, d=.48. Overall, participants in the game condition had spent longer playing. This might have had an effect on reported enjoyment, as engaging longer with the interface might have made the task less novel and more boring. We therefore decided to delimit the second study by time frame, not number of inputs. Third, in the control condition, participants were expressly instructed to provide words in a 'correct' order: "Describe the highlighted shape in the order that feels most correct to you". A similar instruction was missing in the game condition. This may have increased the observed difference in accuracy. Hence, we decided to replace this instruction with a more neutral one in study 2.

5 STUDY 2

Study 2 was a conceptual replication of study 1 with several changes to test the robustness of our results: using a constructivist operationalisation of accuracy, holding usage time constant, and removing a prompt that could have induced demand characteristics. A preregistration can be found at https://osf.io/sg3uk/, along with a repository containing all materials, code, and data at https://osf.io/4g9fh/. As in our first study, we hypothesized that:

- **H1** Players experience more enjoyment in the game condition than the control.
- **H2** Accuracy is lower in the game condition than the control.

Additionally, given our observation that participants in the game condition took longer to play, we hypothesized that:

- **H3** Participants in the game condition will take more time per game input than in the control.

5.1 Method

5.1.1 Materials. We used the *Adjective Game* and control task described above, with the following updates: In the control condition, to reduce possible demand characteristics, we replaced the instruction "Describe the highlighted shape in the order that feels most correct to you" with "Choose from the words below to describe the shape." In the game condition, we fixed various bugs. Importantly, we removed a menu button at the top-right of the screen that was accidentally included in the first study, which opened a sliding menu that gave the option to return to the main menu, repeat the tutorial, restart the level, and showed a line of debug information revealing the condition, labelled either 'Game' or 'Tool'. It is unlikely that this would have affected participants' responses, as it did not reveal the nature of the other condition to the participant, and telemetry showed that participants did not use the menu. We also altered the levels and order of levels to improve the game's learning curve and game balance, improved particle effects, and added a help button that opened a modal dialogue with brief instructions about how to play. Finally, we changed how we randomized the order of word buttons: instead of randomly positioning buttons on a grid, buttons were positioned in columns by type (colour, size, shape, etc.), and the order of columns and of buttons in each column was randomized per level. This solution was slightly more usable for players while still retaining randomisation.

5.1.2 Sample. Power analyses was performed for each of our hypotheses based on the effect sizes observed in study 1[6]. The largest of these suggested we needed a sample size of 140 to detect an effect of d=.48 (for H3) with a statistical power of 0.8 with an alpha of 0.05. We recruited 185 adults with the first language of English via Prolific.[7]. The study offered £1.20 for completing a 12 minute task (£6 per hour) entitled "A study where you describe shapes". After excluding 4 incorrectly submitted records, 9 participants who reported their first language as other than English, and 1 participant who withdrew their submission, 171 participants were included in the published data set. Of these, a further 32 were excluded from our statistical tests: 7 were excluded because they had submitted fewer than 16 moves, and 25 were excluded because they reported a bug that, in their judgement, may have influenced how they played the game. Their records are still included in the participant demographics. All exclusions followed the process specified in the preregistration. Thus, a total of 139 participants were included in the statistical analysis. Of these, 66 were in the game condition, and 73 in the control condition.

For each participant, we used the last 16 valid inputs before their task time ran out to determine accuracy. This value was selected to ensure we included as broad a range of players as possible in the analysis. The value 16 corresponds to two standard deviations below the mean for inputs per user in study 1, meaning we expected to include approximately 95% of participants.

[5] This is subtly different from the preregistration which overlooked that adjectives of a category in the game were mutually exclusive. This change makes no difference to the direction of the results.

[6] An exploratory test corresponding to H3 was reported in the discussion of Study 1

[7] Sampling stopped with 67 in the game condition rather than 68 as was preregistered. It was not anticipated that so many exclusions would be required, so the study ran out of money to continue.

5.1.3 Procedure. The procedure was the same as reported for study 1 with the following differences. Rather than each participant providing 20 inputs to the game (excluding the tutorial), once the participants had finished the tutorial, the participants played the game or engaged with the control task until 8 minutes had passed, regardless of how many inputs they provided.

5.1.4 Dependent Variables. As before, we operationalised enjoyment using the Interest/Enjoyment subscale of the Intrinsic Motivation Inventory.

To operationalise accuracy, instead of comparing participants' word strings to ideal English grammar, we compared each participants' strings to that participant's own grammaticality judgments. To this end, after the completion of the Intrinsic Motivation Inventory at the end of the study, participants were presented with a list of modified noun phrases and asked to judge each as either grammatical or ungrammatical. The phrases corresponded to the different ways the types of adjectives (colour, size, etc.) could be ordered in the game/control task. Only noun-final phrases were included, as English has a strong requirement for noun-finality in these contexts. We elicited a total of six judgements from each participant on the phrases given in list (2).

(2) a. red big square
b. big red square
c. big filled square
d. filled red square
e. red filled square
f. filled big square

Accuracy was determined for each participant as the proportion of their recorded game/control inputs for which they had a positive corresponding grammaticality judgement. An input and a judgement correspond if both phrases are similarly ordered with regards to the adjectives they contain. For example, if a participant entered 'small blue circle', this would be compared against their grammaticality judgement for (2b), as both phrases are similarly ordered for colour and size adjectives. If (2b) was judged grammatical, this input would be considered accurate, and inaccurate otherwise. Inputs that were not noun-final were judged inaccurate. For each participant, accuracy was calculated as the number of inputs determined to be accurate in this way as a proportion of recorded inputs. It is this proportion that was used in comparing accuracy between conditions.

5.1.5 Analysis. As with study 1, where there were parametric tests in our pre-registration we have performed non-parametric equivalents. This change makes no difference to the direction of the results. Analysis software was the same as study 1.

5.2 Results

5.2.1 Demographics. Out of the 171 participants included in the initial data set, 106 reported their gender as female, 62 as male, and 1 as other. They ranged in age from 18 to 70. The median age was 31. Of the 139 participants included in the statistical tests, 57% reported playing at least several times a week or more, where 29% played once a month or less.

5.2.2 Enjoyment. A one-tailed Mann-Whitney U test found that enjoyment was significantly higher in the game (M=3.77, SD=1.07) than control (M=2.99, SD=1.06) condition; U=3387.0, p<.001, d=.73, see figure 6.

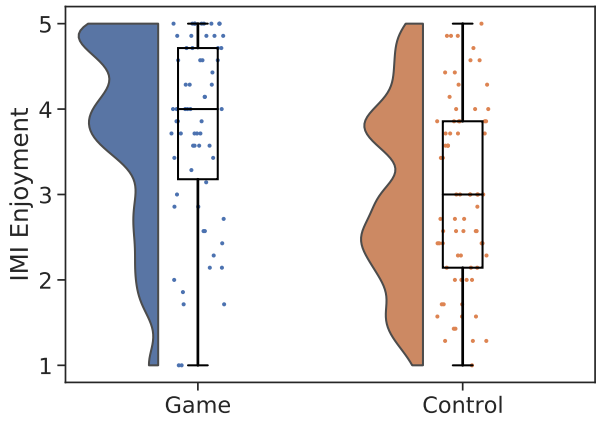


Figure 6: Enjoyment as measured by the IMI Interest/Enjoyment subscale is higher in the game than control condition.

5.2.3 Accuracy. A two-tailed Mann-Whitney U test was used to compare accuracy in the game and control conditions. Accuracy was calculated as the proportion of the last 16 inputs whose order matched the grammaticality judgement separately elicited for that participant. Accuracy was significantly lower in the game (M=0.32, SD=0.26) than control condition (M=0.43, SD=0.29); U=1872.0, p=.01, d =-.40, see figure 7.

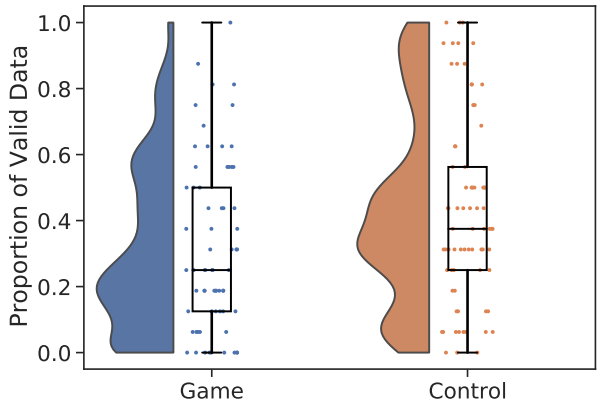


Figure 7: Accuracy as proportion of word orders that correspond to the participant's own grammaticality judgments is higher for the control than game condition

5.2.4 Time per input. A two-tailed Mann-Whitney U test showed that time per input was significantly higher in the game condition (M=0.26, SD=0.08) compared to the task condition (M=0.16, SD=0.06); U=751.0, p<.001, d=1.30.

5.3 Discussion

We again found that enjoyment was higher and accuracy was lower in the game condition than the control, in line with our hypotheses (H1, H2). Also in line with H3, time per input was greater in the game condition compared to the control. This makes the game less efficient as a data collection method than the control in inputs-per-minute. As the interfaces for inputting data were largely identical, this extra time cannot come from inputting the data itself. The difference probably lies in increased cognitive effort from increased visual search and planning involved in strategising during gameplay. However, while the effect size might appear very large (d=1.30), this only amounted to about 6 seconds per input in real terms. For the amount of data required for the experiments reported here, this would work out at a difference of around two minutes between conditions. Such numbers suggest time efficiency is a lesser concern.

6 GENERAL DISCUSSION

We opened this paper with the worry that data collection games have increased in popularity, although we lack good quality evidence that they are in fact more enjoyable and provide equally valid data as comparable standard surveys or experiments. Specifically, we pointed out that data collection games *as games* may present new systemic response biases that threaten validity and data quality – namely that participants choose responses that are more in-game strategically optimal (virtual utility) or more fun (intrinsic utility) than a more careful and honest alternative response. Our results give the worry reason: Across two studies, the game condition proved more enjoyable, but also produced less accurate data. Data collection games, at least for the kind of human subject data we elicited in our case study, present a trade-off between enjoyment and data quality. That said, since we carefully designed our test game to minimise game-germane response biases, our results also suggest other, presently unaccounted factors at work not accounted for in the Intrinsic Elicitation approach. We will work through the ramifications of these findings in order, addressing accuracy, enjoyment, and the Intrinsic Elicitation approach.

6.1 Accuracy

In our two studies, accuracy as a form of data validity was significantly lower in the data collection game than the comparable experiment. This was the case no matter whether accuracy was measured relative to English grammar or to the participants' own judgments. This stands in contrast to the majority of current research on *gamified* surveys and experiments, which finds that gamification at least does not negatively impact data quality [40, 45]. We cannot say whether whether this reduced accuracy holds across all kinds of (latent) human subject properties – Gundry and Deterding [28] for instance suggest that eliciting competencies may be less prone to inaccurate responding, where the game can simply encourage and reward people to do their possible best.

Thus, we encourage future research to try to replicate our findings for different elicited properties to establish their generalisability. Methodologically, we urge that such work use careful "practice as usual" controls, akin to gold-standard randomised controlled trials in medical research. Good controls have been largely amiss in past research. While studies without controls (such as [12, 15, 36, 74]) are necessary first steps, they can also easily provide false comfort that applied games are 'quite' enjoyable and produce 'lots' of data with 'above-random' quality. But this doesn't answer the hard, practically important question whether the extra work of turning a survey or experiment into a game pays off with better enjoyment, engagement, and better-or-equal data quality.

While we expected *some* differences in accuracy, following prior evidence that data collection games are less accurate than gamified equivalents [55] and the suggestion that games as complex stimuli will necessarily increase variance [29], the substantial effect sizes we observed surprised us, not the least since we followed the Intrinsic Elicitation approach [28] to give each possible input the same utility as much as possible. This suggests that there were relevant factors affecting player input choice outside of this model.

What, then, differed between game and experimental task that may not similarly manifest in gamified surveys? Perhaps whether or not the game is played *as a game*. Framing a task as a game, in contrast to an experiment, may affect participants' inclination towards careful answering. When Orne [53] reflected on the nature and origin of *demand characteristics* – the social cues a research participant uses to make sense of what kind of situation they find themselves in and what behaviour therefore is expected of them –, he expressly linked this to situational frames as understood in sociology and recent game studies [19]. He suggested that participants would recognise that their role in the frame of an experimental study was to be a good participant. In contrast, in the frame of gaming, participants might take on the role of players and act accordingly, thus disregarding implicit or imputed normative expectations to respond carefully entailed in the good participant role. This would fit findings by Lieberoth [44] that merely framing an activity as a game changes people's experience and behaviour. While we took take great care to label the overall study and each condition as neutrally as possible and avoid any direct instructions to respond grammatically 'correct' in either condition, especially in study 2, participants in the control condition might still have imputed from the overall format of the task that they are engaged in a linguistic experiment where 'correct' word order is expected, while participants in the game condition did no such thing. After all, the framing of "game" or "experimental task" was not just afforded by explicit verbal labeling, but by the whole structure and characteristics of the task itself. The game not only presented superficial characteristics of a game (like a gamified survey), but played like a game. We therefore suggest future research looking to directly induce and assess different framings and see whether this affects behaviours like careful responding and with it, data quality.

Another possible explanation for the observed lower accuracy in the game condition is that the particular mental operations involved in assembling puzzle-solving moves in the game differ from those involved in assembling natural language structures. That is, game participants approached the task as a puzzle whereas control condition participants approached it analogously to spoken language production. Relatedly, the cognitive load involved in strategising optimal moves in the game might have led participants to adopt tactics like 'offloading' a first likely choice as a first input. One way of probing this explanation would be to further separate the puzzle-solving part of the game from data entry, e.g. ask participants to first choose a combination of words and then say them aloud.

Another alternative partial explanation may be that the operationalisation of accuracy we chose and the direct instruction to provide grammatically correct inputs in the control condition of study 1 inflated differences. Both may indeed have had an effect – after all, the effect size for accuracy shrank from $d = -0.68$ to $d = -0.40$ when we changed these two aspects of the study design in study 2. This invites future research into the effects of directly asking participants to provide accurate data (akin to common survey design strategies asking participants to respond honestly [71]), and future conceptual replications using different operationalisations of accuracy and data validity.

Inspecting our data, we also observed that accuracy seemed to markedly decrease overall between studies, for both conditions. To test this, we looked at the first 20 inputs in both conditions for study 1 and 2, evaluating accuracy in comparison to the idealised grammar. An exploratory two-tailed Mann-Whitney test shows that accuracy indeed is significantly higher in study 1 (M=.58, SD=.35) than study 2 (M=.37, SD=.27) value; U=4161, p<.001, d =.70.

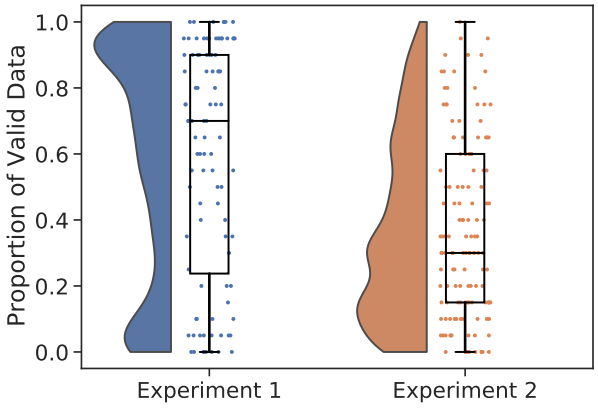


Figure 8: Combining both conditions, overall accuracy between experiments decreases, measured as agreement with ideal English grammar

One possible explanation for this overall decrease is simple regression to the mean. Here, direct replications can help. Another possible explanation are history effects: our first study was run before the COVID-19 pandemic, our second study in the middle of it. A recent analysis of Amazon Mechanical Turk studies identified a surge of new participants, an increase in diversity of participants, a reduction in participant reflectiveness, and an increase in failed attention checks during the pandemic [3]. Similar history effects may also apply between the studies using Prolific reported here. Finally, study 2 was time-limited rather than input-limited, which may have induced a sense of time pressure in some participants, resulting in less careful responding. Though again, neither time pressure nor the impact of COVID-19 in study 2 would explain the differences in accuracy observed in study 1.

6.2 Enjoyment

By and large, prior work found that data collection games are enjoyable [55], and that gamified data collection (i.e. the incorporation of only game design *elements*) features a more positive user experience than comparable controls [32, 40, 45]. Our findings align with this: players did experience more enjoyment in the game than control, with effect sizes in line with a recent meta-analysis on the engagement/experience effects of gamification in cognitive assessments [70]. This supports prior claims that data collection games can make participation in data provision more enjoyable. We hasten to add that this does not answer whether such higher enjoyment corresponds or leads to higher behavioural engagement – prior work on gamified data collection suggests it may not [32, 40, 45]. Future work is needed that would track correlations between experiential enjoyment and behavioural engagement in data collection *games* (not gamification), especially in non-paid, volunteer contexts.

We suggested above that framing might be responsible for how participants engaged with the game compared to the control. This might have affected enjoyment as well: if people perceived their activity to be voluntarily play rather than paid labour, this in and of itself might have e.g. satisfied people's need for autonomy experiences, generating higher enjoyment [20]. This opens the broader question whether the same qualities that make data collection games less accurate also make them more enjoyable (e.g. framing), or if the two are separable. The fact that *gamified* data collection seems to improve enjoyment without a loss of data quality suggests the latter. Contrary to our original intuitions, one possible upshot of this is that *gamified* data collection may prove to be a better option for practitioners than data collection *games*. If future research is able to identify and dissociate design aspects that drive one but not the other, this would hold great practical value, as it would help designers avoid the observed trade-off between accuracy and enjoyment.

6.3 Intrinsic Elicitation

While a detailed reflection on our design process is beyond the scope of this paper, we would like to close with some observations on the Intrinsic Elicitation approach we used [28]. We found the first two of its three design principles (necessity and centrality) straightforward to understand and implement. The veracity principle proved more difficult to realise for several reasons. First, while the model provided some suggestions for different extrinsic, intrinsic, and virtual utilities that might make one choice more utility-maximising than another, outside game-internal virtual utilities, there are no ready ways of identifying likely relevant factors, let alone estimating their likely impact. Here, the approach provided very little practical guidance. This feeds into a second issue: the approach presently offers no pragmatic stopping point for when different actuation options for the data-providing mechanic can be considered reasonably balanced – how utility-balanced is *enough* for the mechanic actuation to be sensitive to the latent property we wish to elicit? Playtesting may be a partial pragmatic solution to this, but to work, we would need some 'ground truth' idea of the likely tendencies a given player should reveal in their responses, to see whether the current iteration of the game produces comparable response patterns. The final obvious issue is that our studies found sizeably lower accuracy in the game condition *despite* our usage of the Intrinsic Elicitation approach, suggesting that in its current form, it doesn't account for all likely important factors.

We think many of these issues stem from the fact that the approach is really an articulation of sensible principles or design goals, but lacks underlying methods that would walk a practitioner towards accomplishing these goals. Thus, future work might expand on the approach with such methods and processes, and incorporate to-be-determined factors causing lower accuracy. Still, we feel that the *Adjective Game* we designed benefited from the approach in avoiding a larger number of potential validity threats that we might have otherwise overlooked.

7 CONCLUSION

Games are not an easy path to engaging participants in providing valid data. They are more challenging and effortful to design than "practice as usual" experiments or surveys. Thus, we need to justify their additional cost with commensurate benefits. The standard rationale is that data collection games are more enjoyable, but provide equal if not better data quality. Yet to our knowledge, this rationale had never been put to a rigorous test, comparing a data collection game to an equivalent control. In two studies using linguistic data elicitation as case material, we found that data collection games are indeed substantially more enjoyable than a "practice as usual" study design. Yet we also observed a significant trade-off, in that these games also provided less accurate data. Since we expressly followed a design approach dedicated to minimising validity threats and eliminated other likely confounds in a conceptual replication study, we have reason to believe that this trade-off is real, but have no ready explanation for its existence, apart from the possibility that framing a task as a game versus as an experiment may induce different demand characteristics for careful responding.

It is not yet clear how widely this observed trade-off applies. In particular, it may not extend to data about individual competencies (e.g. typing speed), as in such cases a game could provide performance-contingent rewards to further maximise accuracy. However, spontaneous, non-deliberate behaviours or preferences allow no clear ways to further mandate or encourage higher-quality data and thus are likely to be subject to this trade-off. We thus urge future research to both test the generalisability of our findings, using study designs with practice as usual controls, and to explore the potential impact of framing on data quality and enjoyment in game-based data collection.

ACKNOWLEDGMENTS

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/L015846/1] and the Digital Creativity Labs (digitalcreativity.ac.uk), jointly funded by EPSRC/ AHRC/Innovate UK under grant no. EP/M023265/1.

REFERENCES

[1] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, Jordy van Langen, and Rogier A. Kievit. 2021. Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Res* 63 (2021). https://doi.org/10.12688/wellcomeopenres.15191.2

[2] Ben Ambridge and Caroline F. Rowland. 2013. Experimental methods in studying child language acquisition. *WIREs Cognitive Science* 4, 2 (2013), 149–168. https://doi.org/10.1002/wcs.1215

[3] Antonio A Arechar and David G Rand. 2021. Turking in the time of COVID. *Behavior Research Methods* 53 (December 2021), 2591–2595.

[4] Carl Bache and Niels Davidsen-Nielsen. 2010. *Mastering English: An advanced grammar for non-native and native speakers*. Vol. 22. Walter de Gruyter, Berlin.

[5] Melanie L Bell, Michael G Kenward, Diane L Fairclough, and Nicholas J Horton. 2013. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ* 346 (2013), e8668.

[6] Harriet R Brown, Peter Zeidman, Peter Smittenaar, Rick A Adams, Fiona McNab, Robb B Rutledge, and Raymond J Dolan. 2014. Crowdsourcing for cognitive science–the utility of smartphones. *PloS one* 9, 7 (2014), e100662.

[7] Stephane Champely. 2020. pwr: Basic functions for power analysis. https://CRAN.R-project.org/package=pwr Version 1.3-0.

[8] Noam Chomsky, Ángel J Gallego, and Dennis Ott. 2019. Generative grammar and the faculty of language: Insights, questions, and challenges. *Catalan Journal of Linguistics* 2019 (2019), 229–261.

[9] Guglielmo Cinque. 2002. *Functional structure in DP and IP*. Vol. 1. Oxford University Press, Oxford.

[10] Guglielmo Cinque. 2010. *The syntax of adjectives: A comparative study*. The MIT Press, Cambridge, Massachusetts.

[11] Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston, MA.

[12] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–60.

[13] Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver-Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, et al. 2010. The challenge of designing scientific discovery games. In *Proceedings of the fifth international conference on the foundations of digital games*. ACM, New York, NY, 40–47.

[14] Antoine Coutrot, Sophie Schmidt, Lena Coutrot, Jessica Pittman, Lynn Hong, Jan M Wiener, Christoph Hölscher, Ruth C Dalton, Michael Hornberger, and Hugo J Spiers. 2019. Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PloS one* 14, 3 (2019), e0213272.

[15] Kevin Crowston and Nathan R Prestopnik. 2013. Motivation and data quality in a citizen science game: A design science evaluation. In *Proceedings of the 2013 46th Hawaii international conference on system sciences*. IEEE, Washington, DC, 450–459.

[16] Vickie Curtis. 2015. Motivation to participate in an online citizen science game: A study of Foldit. *Science Communication* 37, 6 (2015), 723–746.

[17] Amedeo D'Angiulli and Lavonia Smith LeBeau. 2002. On Boredom and Experimentation in Humans. *Ethics & Behavior* 12, 2 (2002), 167–176. https://doi.org/10.1207/S15327019EB1202_4

[18] Jonathan DeRight and Randall S. Jorgensen. 2015. I just want my research credit: Frequency of suboptimal effort in a non-clinical healthy undergraduate sample. *The Clinical Neuropsychologist* 29, 1 (2015), 101–117. https://doi.org/10.1080/13854046.2014.989267

[19] Sebastian Deterding. 2013. *Modes of play: A frame analytic account of video game play*. Ph.D. Dissertation. Hamburg University, Hamburg, Germany. https://ediss.sub.uni-hamburg.de/handle/ediss/5508

[20] Sebastian Deterding. 2016. Contextual autonomy support in video game play: a grounded theory. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, New York, NY, 3931–3943.

[21] Sebastian Deterding, Alessandro Canossa, Casper Harteveld, Seth Cooper, Lennart E Nacke, and Jennifer R Whitson. 2015. Gamifying research: Strategies, opportunities, challenges, ethics. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*. ACM, New York, NY, 2421–2424.

[22] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. ACM, New York, NY, 9–15.

[23] Sonja Eisenbeiss. 2009. Contrast is the name of the game: contrast-based semi-structured elicitation techniques for studies on children's language acquisition. *Essex Research Reports in Linguistics* 57, 7 (2009), 1–27.

[24] Sonja Eisenbeiss. 2010. Production methods in language acquisition research. In *Experimental methods in language acquisition research*, E. Blom and S. Unsworth (Eds.). John Benjamins, Amsterdam, 11–34.

[25] Lucy Fortson, Karen Masters, Robert Nichol, EM Edmondson, C Lintott, J Raddick, and J Wallin. 2012. Galaxy zoo. *Advances in machine learning and data mining for astronomy* 2012 (2012), 213–236.

[26] Maximilian Achim Friehs, Martin Dechant, Sarah Vedress, Christian Frings, and Regan Lee Mandryk. 2020. Effective gamification of the stop-signal task: Two controlled laboratory experiments. *JMIR Serious Games* 8, 3 (8 Sep 2020), e17810. https://doi.org/10.2196/17810

[27] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and Individual Differences* 7, 3 (1986), 385–400.

[28] David Gundry and Sebastian Deterding. 2018. Intrinsic Elicitation: A model and design approach for games collecting human subject data. In *Proceedings of the 13th international conference on the foundations of digital games* (Malmö, Sweden) *(FDG '18)*. ACM, New York, NY, Article 38, 10 pages. https://doi.org/10.1145/3235765.3235803

[29] David Gundry and Sebastian Deterding. 2019. Validity threats in quantitative data collection with games: A narrative survey. *Simulation & Gaming* 50, 3 (2019), 302–328. https://doi.org/10.1177/1046878118805515

[30] MP Jacob Habgood and Shaaron E Ainsworth. 2011. Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences* 20, 2 (2011), 169–206.

[31] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[32] Guy E Hawkins, Babette Rae, Keith V Nesbitt, and Scott D Brown. 2013. Gamelike features might not improve data. *Behavior Research Methods* 45, 2 (2013), 301–318.

[33] Jonas Heide-Smith. 2006. *Plans and purposes how videogame goals shape player behaviour.* Ph.D. Dissertation. IT University of Copenhagen.

[34] Joseph Hilgard. 2021. Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology* 93 (2021), 104082.

[35] David Hugh-Jones. 2016. Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization* 127 (2016), 99–114.

[36] Ioanna Iacovides, Charlene Jennett, Cassandra Cornish-Trestrail, and Anna L Cox. 2013. Do games attract or sustain engagement in citizen science? A study of volunteer motivations. In *CHI'13 extended abstracts on human factors in computing systems*. ACM, New York, NY, 1101–1106.

[37] Playdots Inc. 2014. *Two Dots.* Tencent Games.

[38] Jesper Juul. 2008. The magic circle and the puzzle piece. In *Conference Proceedings of the Philosophy of Computer Games*. University Press, Potsdam, Germany, 55–67.

[39] Saskia M Kelders, Robin N Kok, Hans C Ossebaard, and Julia EWC Van Gemert-Pijnen. 2012. Persuasive system design does matter: A systematic review of adherence to web-based interventions. *Journal of Medical Internet Research* 14, 6 (14 Nov 2012), e152. https://doi.org/10.2196/jmir.2104

[40] Florian Keusch and Chan Zhang. 2017. A review of issues in gamified surveys. *Social Science Computer Review* 35, 2 (2017), 147–166.

[41] King. 2012. *Candy Crush Saga.* King, St. Julian's, Malta.

[42] Michael W. Kirkwood, John W. Kirk, Robert Z. Blaha, and Pamela Wilson. 2010. Noncredible effort during pediatric neuropsychological exam: A case series and literature review. *Child Neuropsychology* 16, 6 (2010), 604–618. https://doi.org/10.1080/09297049.2010.495059

[43] Laura Levy, Rob Solomon, Jeremy Johnson, Jeff Wilson, Amy J Lambeth, Maribeth Gandy, Joann Moore, Jason Way, and Ruitao Liu. 2016. Grouches, extraverts, and jellyfish: Assessment validity and game mechanics in a gamified assessment. In *DiGRA/FDG: Proceedings of the first international joint conference of DiGRA and FDG*. Digital Games Research Association and Society for the Advancement of the Science of Digital Games, Dundee, Scotland, 1–16.

[44] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture* 10, 3 (2015), 229–248.

[45] Jim Lumsden, Elizabeth A Edwards, Natalia S Lawrence, David Coyle, and Marcus R Munafò. 2016. Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games* 4, 2 (15 Jul 2016), e11. https://doi.org/10.2196/games.5888

[46] Heikki Lyytinen, Miia Ronimus, Anne Alanko, Anna-Maija Poikkeus, and Maria Taanila. 2007. Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. *Nordic Psychology* 59, 2 (2007), 109–126.

[47] Elisa D Mekler, Julia Ayumi Bopp, Alexandre N Tuch, and Klaus Opwis. 2014. A systematic review of quantitative studies on the enjoyment of digital entertainment games. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, 927–936.

[48] Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist* 50, 9 (1995), 741.

[49] Janet Metcalfe, Nate Kornell, and Bridgid Finn. 2009. Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & cognition* 37, 8 (2009), 1077–1087.

[50] Josh Aaron Miller, Uttkarsh Narayan, Matthew Hantsbarger, Seth Cooper, and Magy Seif El-Nasr. 2019. Expertise and engagement: re-designing citizen science games with players' minds in mind. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*. ACM, New York, NY, 1–11.

[51] Tomohiro Nishikado. 1987. *Space Invaders.* Atari, Sunnyvale, CA.

[52] Patrick Oladimeji, Harold Thimbleby, Paul Curzon, Ioanna Iacovides, and Anna Cox. 2012. Exploring unlikely errors using video games : An example in number entry research. In Workshop on Safety-Critical Systems and Video Games: Contradictions and Commonalities, Held at Fun and Games. , 5 pages. http://oro.open.ac.uk/47237/

[53] Martin T Orne. 1981. The significance of unwitting cues for experimental outcomes: Toward a pragmatic approach. *Annals of the New York Academy of Sciences* 364, 1 (1981), 152–159.

[54] The pandas development team. 2020. pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134 Version 1.2.1.

[55] Nathan Prestopnik, Kevin Crowston, and Jun Wang. 2017. Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior* 68 (2017), 254–268.

[56] R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[57] Roger Ratcliff. 1993. Methods for dealing with reaction time outliers. *Psychological Bulletin* 114, 3 (1993), 510.

[58] Johnmarshall Reeve. 2014. *Understanding motivation and emotion.* John Wiley & Sons, Hoboken, NJ.

[59] Dmitry Rudchenko, Tim Paek, and Eric Badger. 2011. Text text revolution: A game that improves text entry on mobile touchscreen keyboards. In *Pervasive Computing*. Springer, Berlin, 206–213.

[60] David A Sawin and Mark W Scerbo. 1995. Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human factors* 37, 4 (1995), 752–765.

[61] Miguel Sicart. 2008. Defining game mechanics. *Game Studies* 8, 2 (2008), 1–14.

[62] Karin Slegers, Bernhard Maurer, Lizzy Bleumers, Alina Krischkowsky, Pieter Duysburgh, and Mark Blythe. 2016. Game-based HCI methods: Workshop on playfully engaging users in design. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. ACM, New York, NY, 3484–3491.

[63] John P Spencer and Alycia M Hund. 2002. Prototypes and particulars: geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General* 131, 1 (2002), 16.

[64] Richard Sproat and Chilin Shih. 1991. The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language*. Springer, Dordrecht, 565–593.

[65] Giri Kumar Tayi and Donald P Ballou. 1998. Examining data quality. *Commun. ACM* 41, 2 (1998), 54–57.

[66] Richard I Thackray. 1981. The stress of boredom and monotony: a consideration of the evidence. *Psychosomatic Medicine* 2 (April 1981), 165–176.

[67] Ramine Tinati, Markus Luczak-Roesch, Elena Simperl, and Wendy Hall. 2017. An investigation of player motivations in Eyewire, a gamified citizen science project. *Computers in Human Behavior* 73 (2017), 527–540.

[68] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[69] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 reference manual.* CreateSpace, Scotts Valley, CA.

[70] Julie F Vermeir, Melanie J White, Daniel Johnson, Geert Crombez, and Dimitri ML Van Ryckeghem. 2020. The effects of gamification on computerized cognitive training: Systematic review and meta-Analysis. *JMIR serious games* 8, 3 (2020), e18644.

[71] Vaka Vésteinsdóttir, Adam Joinson, Ulf-Dietrich Reips, Hilda Bjork Danielsdottir, Elin Astros Thorarinsdottir, and Fanney Thorsdottir. 2019. Questions on honest responding. *Behavior Research Methods* 51, 2 (2019), 811–825.

[72] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[73] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *CHI '14: Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, 319–326.

[74] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.