Enjoyment and Data Quality in a Human-Subject Data Collection Game, a conceptual replication

This experiment is a conceptual replication of an earlier study that differs by ending the experimental task after a set time rather than after a set number of inputs. It also operationalises what counts as 'valid data' differently – by asking participants to perform a short grammaticality judgement task. Finally, it arranges the words that the players use for input in columns rather than a random grid, and removes access to the menu, which was previously available in the game condition. Some of the wording in the tool condition is changed to remove explicit instructions to use grammatical word order.

This experiment investigates an applied game for a (human subject) data collection that has been developed for the elicitation of English adjective word order. Two conditions will be tested. In one condition, the game will be played. In a second condition, the game interface will be used to approximate a standard experimental design for the same data.

- The enjoyment of the two conditions will be compared using a post-test questionnaire. The questionnaire used will be the Intrinsic Motivation Inveotry (IMI) Interest/enjoyment subscale.
- The proportion of "naturalistic" data elicited from players (in comparison to grammaticality judgements from a post-test questionnaire) will be compared.

Background

Broadly studies of human behaviour are concerned with eliciting natural inputs. Many experimental measures assume that players respond naturally, e.g. studies of language generally assume that participants speech reflects their everyday linguistic performance and isn't affected by the experiment, for example. Biases, such as social desirability bias, can affect this.

Applied games can be designed to collect human-subject data from players. Games are supposed to be intrinsically motivating. This motivation can support or replace other forms of motivation for participation (e.g. payment) in a data collection task.

As traditional (non-game) controlled experiments using established methodologies are considered the gold-standard for data quality, use of games might be assumed to provide poorer quality data, especially as they potentially introduce many poorly understood biases. Data collected from a game may therefore be of lower quality, particularly behavioural data.

We have designed an applied game to collect data. This game is designed to collect adjective orderings (i.e. what is the way you naturally order the words

“big”, “red” and “square” in a noun phrase). The game bears some similarity to a picture description task, a standard experimental paradigm in linguistics. whereby participants are asked to describe a picture.

A previous, pre-registered study was run that was similar to the one described here. The first three hypotheses listed below were confirmed by the experiment. The foruth is a new hypothesis for this experiment based on exploratory analysis of the data from the previous experiment.

Hypotheses

1. Players experience **more enjoyment** from the game condition than the task condition.
2. Players provide **poorer quality data** in the game condition than the task condition.
3. **Time per input is higher** in the game condition than the task condition.

Methods

Design

Dependent Variables

DV1. Enjoyment. Operationalised using an IMI Interest/Enjoyment subscale questionnaire administered at the end of the experiment.

For the purposes of the following two variables, *correct word order* is considered to be 1) noun-final, and 2) correspond to a word order judged as grammatical in the post-test grammaticality judgement task described below.

DV2. Proportion of valid data. Operationalised as proportion of the last 16 complete (3-word) inputs with correct word order that the participant provides, regardless of their in-game effect.

DV3. Mean time per input. Operationalised as total time spent performing the game/task (8 minutes) divided by the number of inputs entered in that time.

Demographic Variables

Age.

> “What is your age?”

Gender.

> “What is your gender?
>
> - Female
> - Male
> - Other
> - Prefer not to say”

English as first language.

> "What is your first language?
>
> - English
> - Other"

Gaming Experience. Operationalised as frequency of game play using the following question:

> "How often do you play digital games?
>
> - Every day
> - Several times a week
> - About once a week
> - About once a month
> - (Almost) never"

Grammaticality Judgement Task

Grammaticality Judgement. Correct word order will be operationalised as being both noun final and matching an order judged as grammatical in the following post-test multiple answer question:

> "Select the phrases which are grammatical.
>
> - red big square
> - big red square
> - big filled square
> - filled red square
> - red filled square
> - filled big square"

Sample Size

140 participants recruited online through Prolific (https://www.prolific.co/) Participants will be pre-filtered to select adults (18+) whose first language is English. The participants will be randomly assigned to a condition on an individual basis. This means the number of participants in each condition may vary due to chance. If either of the conditions has fewer than 68 participants (after the exclusion criteria have been applied), the sample size will be increased by 5 until both conditions have at least 68 participants. This will be done before the data is analysed.

A power analysis was performed (see `power-analysis.r.Rout`), which determined that each condition would require 68 participants.

Exclusion Criteria

- Incomplete data records not included
- Users reporting an age of under 18.
- Users reporting that English is not their first language
- Users submitting fewer than 16 inputs

Procedure

Participants will be recruited online. They will be randomly assigned to an experimental condition. This is thus a double-blind design. After providing informed consent, they will fill in a short demographics questionnaire. Then they begin either playing the game or performing the task. At the end of this, they are asked to complete an on-screen questionnaire. They are then thanked for their time.

Game Condition

The game is played in a desktop web browser. It is a casual puzzle game. The aim of the game is to clear all the blocks (shapes of various colours, sizes, etc.). There are a succession of levels. A ‘group’ of blocks are orthogonally adjacent blocks that are all cleared at the same time. Clearing groups of blocks provides bonus moves, which are required to complete the level. When blocks are cleared, blocks above fall down to replace them. Players therefore think ahead to form groups.

In order to clear blocks, players select 3 words from a set provided, all either adjectives or nouns, which all describe blocks in the level. These descriptive words are found at the bottom of the screen. Words of the same type (colour adjective, size adjective, noun, etc.) are always shown in columns. The order of columns is randomised per level. Once 3 words are selected, all of the blocks matching the intersection of those descriptions are cleared. The words can be provided in any order. If no blocks are described by the words (or the words do not include a noun), the attempt fails.

The game starts with a tutorial introducing the interface and mechanics. Following this, the players play a series of levels. Once they have played for 8 minutes, the game ends.

Task Condition

The task condition is similar in interface to the game condition. Players are shown a grid of blocks and with one indicated. They are asked to describe the indicated block. When they do this, they move on to the next level.

Game-specific interface elements such as moves, scores, and level indicators are removed from this condition.

The task starts with a tutorial introducing the interface and the task. Following this, players complete a series of levels. Once they have both performed this task for 8 minutes (excluding the tutorial), the task ends.

Analysis Plan

Main Test

Hypothesis 1: Enjoyment (DV1) will be greater in the game condition than the task condition. A one-tailed two-sample t-test will be used to test whether the mean scores of DV1 is greater in the game condition than the task condition. $\alpha = 0.05$.

Hypothesis 2: Proportion of valid data (DV2) will be lower in the game condition than the task condition. A two-tailed Mann-Whitney U test will be used to test whether the distribution of DV2 differs significantly between the game condition than the task condition. $\alpha = 0.05$

Hypothesis 3: Time per input (DV3) will be higher in the game condition than the task condition. A two-tailed two-sample t-test will be used to test whether scores of DV3 are greater in the game condition than the task condition. $\alpha = 0.05$

Further Exploration

We will look for further insights in the data without being guided by strong hypotheses. This will inform future studies that may empirically test any such findings. For example, we will explore for systematic biases within the data that may be the result of e.g. the game interface, and whether prior game experience moderates observed effects on dependent variables.