Enjoyment and Data Quality in a Human-Subject Data Collection Game

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 as originally designed. 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 (defined as corresponding to an assumed standard English word order) 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.

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. When players actuate the data-providing mechanic in the game condition, they are **biased towards providing valid data**, in comparison to other potential valid (game-equivalent) mechanic actuations.

Methods

Design

Dependent Variables

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

DV2(a). Proportion of valid data. Operationalised as proportion of the 20 complete (3-word) inputs that correspond to standard English grammar (i.e. exactly 1 noun and standard word order), regardless of their in-game effect. There is broad agreement on what counts as standard English word order, and all relevant cases can be supplied in an appendix or published with the dataset.

DV2(b). Proportion of valid data-providing mechanic actuations. Operationalised as the proportion of inputs demonstrating standard English word-order in all complete (3-word) inputs containing exactly 1 noun that a user provides, regardless of their in-game effect. This is a subset of DV2(a)

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”

Sample Size

100 participants recruited online through Prolific (https://www.prolific.co/) Participants will be pre-filtered to select adults (18+) whose first language is English. 50 per condition.

A power analysis has been conducted (see attached file `power-analysis.Rout`), suggesting that 50 participants would be required per condition, making 100 overall.

Exclusion Criteria

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

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 in a grid. Their order 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 made exactly 20 inputs (excluding the tutorial), 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 made exactly 20 inputs (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(a)) will be lower in the game condition than the task condition. A two-tailed two-sample t-test will be used to test whether the mean scores of DV2(a) is less in the game condition than the task condition. $\alpha = 0.05$

Hypothesis 3: Proportion of valid data-providing mechanic actuations (DV2(b)) will be greater in the game condition than what would be expected if ordering was random. A two-tailed one-sample t-test will be used to compare DV2 for the game condition against the theoretical mean expected if players provided word orderings that were completely random. This is given below. $\alpha = 0.05$

Theoretical random proportion = (valid permutations / total permutations of 3 words) = 1/6 = 16.67%

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.