Analysis protocol

# Learning network complexity through gameplay: An experimental study of data game learning outcomes

2021

*Authors:*

Chris Moreno-Stokoe, Claire Haworth & Oliver Davis

Contents

## Background

## Introduction

## Aims & Objectives

## Methods

## Design

## Participants

## Materials

## Procedure

## Data analysis plan

## Inventory of data

## Data cleaning

## Scoring

## Planned analyses

# Background

# Introduction

Games are a good medium for exploring complexity because data can be accurately modelled in games and players can be directed and motivated to engage with the data more than they otherwise would.

The literature on game-based learning demonstrates that games can enhance the learner’s experience and this can translate to improvements in Intended Learning Outcomes. For example, games can provide concrete experiences and active experimentation (Ricardi & DePaulis, 2014) and constructively aligning the learning and assessment modalities to the degree that in-situ stealth assessments are possible (Shute et al., 2013).

Simulation games, that is games designed to reflect reality and teach practical learning outcomes, are often created for professional application. Serious games are in use for medical education for surgeons (Paim et al., 2017) and soldiers in battlefield simulation (Pasquier et al., 2016), as well as for industrial applications such as port management (Bilson, Bekebrede & Mayer, 2010) and policy making (Guerts et al., 2007).

This chapter extends the previous literature on game-based learning to complex networks of the relationships between factors in public health (e.g., BMI, smoking and wellbeing). In addition, previous studies on simulation do not clearly delineate between gameplay and simulation experiences but we will experimentally separate the gameplay from an interactive visualisation simulation. We experimentally investigate whether gamifying an interactive visualisation simulation improves learning outcomes about a public health data network.

Bloom’s Taxonomy of Learning (Anderson & Bloom, 2001) tells us that there are different levels of understanding. These are often visualised as a pyramid, with the lowest and easiest levels being the ability to remember the information; interpret, exemplify, and summarise it, and the higher and harder levels concerning critical evaluation and ability to use the information. It therefore follows that the best test of high-level understanding of complex public health dataset is to ask learners to use this information to design public health interventions. Learners who are capable of this will demonstrate a high degree of understanding.

# Aims

##### Research question

Does gamification improve student’s experience and outcomes of learning the effects between traits in a public health intervention simulator?

##### Objectives

We experimentally compared a gamified and non-gamified interactive visualisation in order to investigate the following hypotheses:

* Gameplay will alter the learning experience in the following ways:
  + It will provide structure for engaging with learning software
  + It will motivate learners to engage with the learning software
  + It will facilitate better learning of public health network data

Additionally, we explored how specific game mechanisms modify the learner experience

# Methods

## Design

Participants were tasked with learning a public health dataset and were provided learning software. This was an online study. Participants were randomly assigned to one of two conditions: 1) experimental condition using gamified software, and 2) control condition using a non-gamified software. Between-subjects comparisons were made to analyse the effects of gamifying learning software on the learning experience and outcomes.

Ethics approval was obtained from the University of Bristol Psychological Science School Research Ethics Committee (ID: 111083).

## Participants

Participants with a background in science were desirable so undergraduate Psychology studies from the University of Bristol were recruited. 210 students signed up to complete this study in return for course credit as part of the Experimental Hours Scheme (of which 197 completed the study). A power analysis was completed estimating that at least 186 individuals were required to detect a minimum meaningful difference at 90% power and an accepted alpha level of 0.01 (Supplementary Table 1 for power analysis).

## Materials

**Software**

*Core (non-gamified) interactive visualisation*

In previous studies, a network dataset for public health traits was created (Moreno-Stokoe et al.), and the MiRANA interactive visualisation tool and JavaScript library was developed (Moreno-Stokoe, Haworth & Davis). In the present study, MiRANA was used to develop an interactive visualisation of this public health data. This formed the control ‘non-gamified’ learning software. Users were given a pre-face containing background information on the dataset as well as instructions on how to interpret and interact with the visualisation.

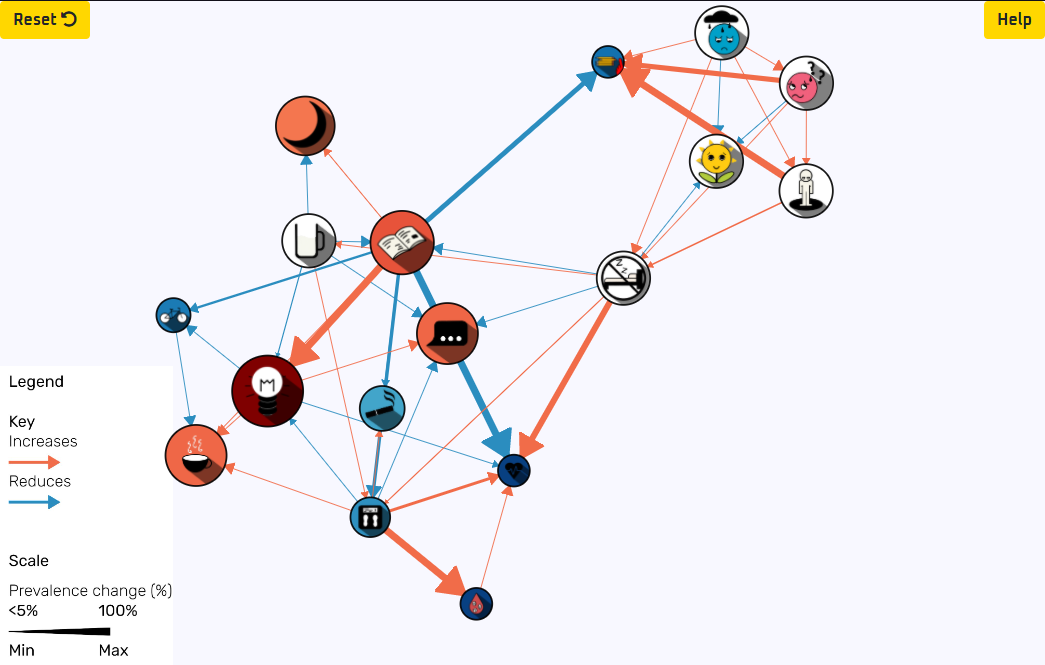


Figure 1. Interactive visualisation learning software available from: <https://www.morenostok.io/mendel/interactiveVisualisation.html>

*Gamified learning software*

The interactive visualisation software was developed into a game with minimal modifications so that it could serve as a control in a ‘gamified’ vs ‘non-gamified’ experimental comparison. Full details of the gameplay will be detailed in Moreno-Stokoe, Haworth and Davis.

In order to control differences between control and experimental conditions, the gamified and non-gamified software contained the exact same available information. The core interactive visualisation, the traits and data, were the same in both the gamified and non-gamified software, as was the pre-face information. Rather than presenting new information, gameplay provided structure to one’s interaction with the core interactive visualisation.

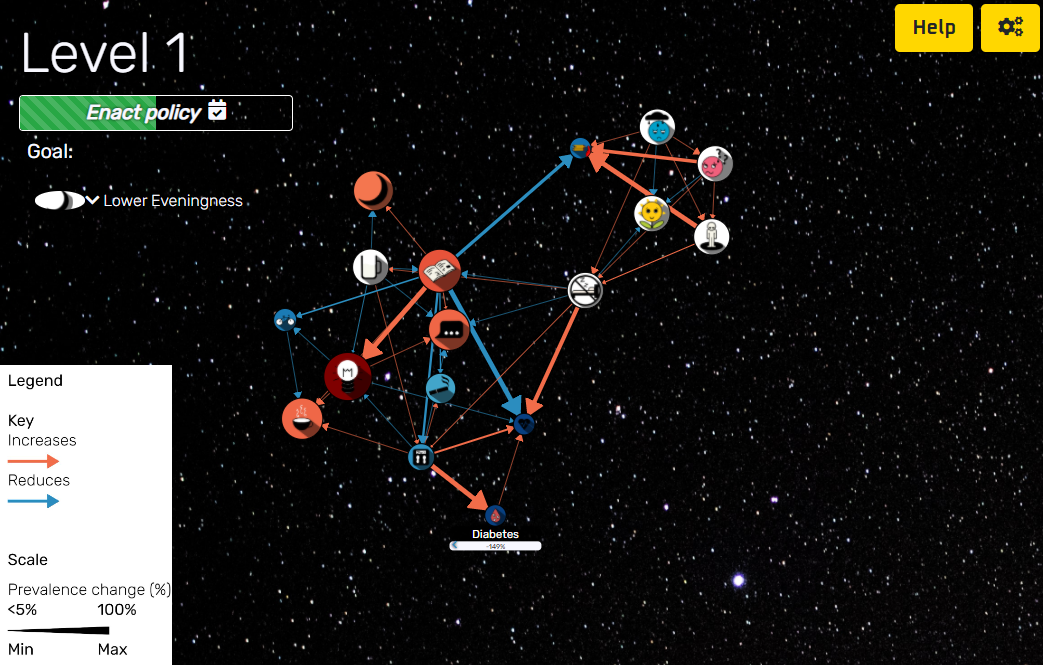


Figure 2. Game learning software available from: <https://www.morenostok.io/mendel/game.html>

**Questionnaires**

*Self-report user experience questionnaire*

Four questions were asked:

* Please indicate your agreement with the following statements using 7-point Likert scales (strongly agree - strongly disagree):
  + I felt motivated to make interventions
  + I felt guided to interact with the simulation in a specific way
* Please drag the experiences below to the appropriate box depending on whether you experienced them or not (based on Playful Experiences Framework: Lucero et al., 2013):
  + Captivation - Forgetting one’s surroundings
  + Challenge - Testing abilities in a demanding task
  + Competition - Contest with oneself or an opponent
  + Completion - Finishing a major task, closure
  + Discovery - Finding something new or unknown
  + Progression - Earning momentum and achievement
  + Exploration - Investigating an object or situation
  + Fantasy - An imagined experience
  + Humor - Fun, joy, amusement, jokes, gags
  + Nurture - Taking care of oneself or others
  + Relaxation - Relief from bodily or mental work
  + Sensation - Excitement by stimulating senses such as sights or sounds
* A free-form answer question:
  + In your own words please describe your experience with the software. For example, did you have a strategy? Did anything prevent you from achieving what you wanted? Did you find any effects memorable? Did you have any opinions about the presentation?

*Learning assessment*

The learning assessment was comprised of 29 questions presented in five progressively more difficult sections. Each section tested whether participants had learned intended learning outcomes. These intended learning outcomes mapped onto areas of competency:

1. Ability to read information about nodes and edges in the network visualisation
2. Understanding of direct effects: Infer the direct effects of interventions which increase the prevalence of a trait
3. Understanding of network properties: Ability to make general inferences about the network
4. Understanding of interactions: Critically analyse interaction effects between multiple interventions which increase the prevalence of different traits
5. Ability to negate effects: Infer the direct effects of interventions which decrease the prevalence of a trait
6. Understanding of indirect effects: Infer the indirect effects of interventions which increase the prevalence of a trait
7. Working understanding of complex network effects: Ability to design solutions to public health problems involving multiple interventions which have a mix of direct and indirect effects, which both increase and decrease traits

The assessment had two modalities; the first 25 questions are delivered as a multiple-choice questionnaire, and the last 4 questions are a free-form intervention design exercise. For a full list of questions and scoring see supplementary table 2. Each area of competency was matched with specific intended learning outcomes and corresponding assessable competencies which individuals should be able to perform. For a full list of intended learning outcomes and competencies see supplementary table 3.

## Procedure

An online form was used to conduct the study and collect participant’s responses (hosted by Qualtrics). Participants completed this online during December 2020.

This form contained five sections:

1. Participant information and consent

Participants were given information relating to the study and required to give their consent before continuing with the study.

1. Assignment to experimental condition

Participants were randomly assigned to either the experimental or control condition. The gamified and non-gamified software were hosted on separate web pages and participants were provided the corresponding URL links.

Participants were prompted to access these links and to “Spend as long as you like until you feel like you have a good enough understanding of the relationships between public health traits and then return to this questionnaire for an open-book learning assessment”.

Once finished, participants were asked to confirm that they had accessed the links provided. The number of mouse clicks and time duration individuals spent on this page of the online form page were measured.

1. Self-report feedback

Participants were then asked about their experience with the software.

1. Learning assessment

For the final section involved a learning assessment intended to measure learning of the relationship between traits in a public health dataset.

First, participants were given an introduction to an essential term which would be used throughout the assessment to refer to relationships between public health traits (direct and indirect relationships).

Second, participants were prompted to open, and keep open for reference, a visualisation of the data via URL link (<https://www.morenostok.io/mendel/visualisation.html>). The assessment was an open book test in order to test better player’s implicit learning rather than rote learning, or memorisation of the traits in analysis.

Third, participants completed the learning assessment comprising of 29 questions which took participants about half an hour. Participants’ time duration was recorded.

1. Debrief

Participants were debriefed in this section and on completion were reimbursed with course credits.

# Analysis plan

# Inventory of data

Participant information

* Experimental condition (control / experimental)
* Time duration spent on section 2 (assignment to condition)
* Clicks made on section 2

Self-report user experience

* Feeling of direction to interact with software
* Feeling of motivation to use software
* Categories of playful experiences that they experienced
* Free-form feedback

Assessment

* Time duration to complete assessment
* Multiple-choice section (Q n=25) score
* Intervention design scores (Q n=4)
  + Primary score (achieving the given objective with intervention)
  + Side-effect score

# Data cleaning

Output data from the online form was inspected. Data contained responses from pilot and pre-flight testing were manually identified and removed (n = 21). 3 individuals had made more than one response; their first responses (n=3) were kept and their additional responses (n=6) were removed.

It is expected that many individuals will not spend a lot of time engaging with either learning software and these individuals are not intended to be excluded, but individuals who do not even try the learning software will be excluded. External links were injected as custom code since they are not natively supported to Qualtrics, and as such it was not possible to concretely verify whether individuals had clicked these links. However, time duration and click data from the condition assignment page (section 2 of the online form) can be used to estimate this. To this end individuals who make fewer than the 2 click minimum to progress page (without opening the link) AND spend fewer than 5 seconds on the condition assignment page will be excluded.

It is also expected that many individuals will have found the assessment very difficult and will show fatigue. However, it is not expected that individuals will fatigue after the first section. Therefore, individuals who finished either the first section, or every section thereafter, in under 5 seconds will be excluded on the basis of not engaging with the assessment as expected. The sections vary in length but it is not possible to complete any section in less than 5 seconds, therefore this is the chosen timeframe to indicate inattention.

These exclusion criteria will be evaluated against the distribution of these factors among participants and adjusted if they appear inaccurate in their application.

# Scoring

The two modalities (multiple-choice and free-form) are scored differently.

If it appears that both tests load onto each other (i.e., strongly correlate with each other), these scores will be merged into a single learning assessment score (weighting: 25 + 8). Conversely, if there is evidence that they measure different constructs (i.e., high scorers for one test are not high scorers in the other), then each participant will receive multiple separate scores.

**Multiple-choice assessment**

The first 25 (multiple-choice) questions were scored simply according to the number of correct answers (for scoring per question see list of questions). Scores will be adjusted by subtracting the score obtainable by random guessing.

**Free-form assessment**

The last 4 (free-form) questions were intervention design scenarios. For these questions, individuals were scored by how much their intervention achieved the optimal results in two dimensions: 1) how well their intervention achieved the objective (e.g., lowering smoking) and 2) how well their intervention improved health generally (e.g., having beneficial side-effects such as improving well being).

Network MR is performed for this. The network MR method (detailed in Moreno-Stokoe and colleagues) is employed to calculate scores using the MiRANA JavaScript library (Moreno-Stokoe, Haworth, Davis) . This will be modified from an existing intervention scoring system used to score players in the game (Moreno-Stokoe, Haworth, Davis). The process of scoring and weighting is extensive and is fully documented in supplementary text 1.

After inspecting the distribution of scores (before main analysis), participants will be scored using ranks or percentages, depending on the distribution. For example, ranks could help differentiate individuals where there is a bimodal distribution of ‘good’ and ‘bad’ interventions.

# Planned analyses

###### Hypothesis testing

Hypotheses will be tested using T-tests. While the accepted alpha level will be 0.05, this will not be used as a strict requirement, and instead this will be evaluated in the context of the magnitude of effect size as well (this following IEU guidelines on significance thresholds). A hypothesis will be rejected if all its tests return non-significant results by this criteria. The three hypotheses will be tested as follows by comparing between experimental and control condition groups:

**Gameplay will alter the learning experience…**

**It will provide structure for engaging with learning software**

* Compare Likert responses to the item ‘I felt directed to use the software…’

**It will motivate learners to engage with the learning software**

* + - Compare Likert responses to the item ‘I felt motivated to use the software…’
    - Compare time duration using learning software

**It will facilitate better learning of public health network data**

* Compare multiple-choice assessment total scores
* Compare free-form primary scores
* Compare free-form side effect scores
* Compare total time to complete assessment

###### Planned exploratory analysis

Exploration of the data is planned to help understand “How do specific game mechanisms modify the learner experience?”. This will be explored by:

* + - Compare participant experiences (PLEX) to see if the game led to feelings of ‘competition’ or ‘thrill’ for example
    - Exploring responses given in the open-form text box about their experience

# References

Anderson, L. W., Bloom, B. S., & others. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Longman,.

Biggs, J. (2003). Aligning teaching for constructing learning John Biggs Keywords What is constructive alignment ? Defining the ILOs. *Education*, *94*(11), 112106. Retrieved from http://egusdsecondaryed.pbworks.com/f/aligning\_teaching\_for\_constructing\_learning.pdf

Geurts, J. L. A., Duke, R. D., & Vermeulen, P. A. M. (2007). Policy Gaming for Strategy and Change. *Long Range Planning*, *40*(6), 535–558. https://doi.org/10.1016/j.lrp.2007.07.004

Giessen, H. W. (2015). Serious Games Effects: An Overview. *Procedia - Social and Behavioral Sciences*, *174*(November), 2240–2244. https://doi.org/10.1016/j.sbspro.2015.01.881

Kamath, H. S., Rao, V., Santhosh, & Kamath, R. (2017). Simulation of long term evolution (LTE) based communication system with different protocols. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)* (Vol. 2018-Janua, pp. 488–491). IEEE. https://doi.org/10.1109/UEMCON.2017.8249059

Lucero, A., Holopainen, J., Ollila, E., Suomela, R., & Karapanos, E. (2013). The Playful Experiences (PLEX) framework as a guide for expert evaluation. *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces, DPPI 2013*, 221–230. https://doi.org/10.1145/2513506.2513530

Paim, C. P. P., & Goldmeier, S. (2017). Development of an educational game to set up surgical instruments on the mayo stand or back table: Applied research in production technology. *JMIR Serious Games*, *5*(1). https://doi.org/10.2196/games.6048

Pasquier, P., Mérat, S., Malgras, B., Petit, L., Queran, X., Bay, C., … Mignon, A. (2016). A serious game for massive training and assessment of french soldiers involved in forward combat casualty care (3d-sc1): Development and deployment. *JMIR Serious Games*, *4*(1), 1–10. https://doi.org/10.2196/games.5340

Ricciardi, F., & De Paolis, L. T. (2014). A Comprehensive Review of Serious Games in Health Professions. *International Journal of Computer Games Technology*, *2014*. https://doi.org/10.1155/2014/787968

Shute, V., Ventura, M., Small, M., & Goldberg, B. (2013). *Design Recommendations for Intelligent Tutoring Systems: Volume 1. Chapter 13*. (R. A. Sottilare, A. Graesser, X. Hu, & H. Holden, Eds.). US Army Research Laboratory.

van Bilsen, A., Bekebrede, G., & Mayer, I. (2010). Understanding complex adaptive systems by playing games. *Informatics in Education*, *9*(1), 1–18. https://doi.org/10.15388/infedu.2010.01