



# THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice

**Citation for published version:**

Aeberhard, A, Gschwind, L, Kossowsky, J, Luksys, G, Papassotiropoulos, A, De Quervain, D & Vogler, C  
2018, 'Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice', *Journal of Technology in Behavioral Science*.  
<https://doi.org/10.1007/s41347-018-0071-5>

**Digital Object Identifier (DOI):**

[10.1007/s41347-018-0071-5](https://doi.org/10.1007/s41347-018-0071-5)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

Journal of Technology in Behavioral Science

**Publisher Rights Statement:**

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.


**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.





# Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice

Andreas Aeberhard<sup>1</sup> · Leo Gschwind<sup>1</sup> · Joe Kossowsky<sup>2,3,4</sup> · Gediminas Luksys<sup>1,5,6</sup> · Andreas Papassotiropoulos<sup>1,7</sup> · Dominique de Quervain<sup>7,8</sup> · Christian Vogler<sup>1,7</sup> 

© The Author(s) 2018

## Abstract

We have established the *CO*gnitive Science Metrics Online Survey (COSMOS) platform that contains a digital psychometrics toolset in the guise of applied games measuring a wide range of cognitive functions. Here, we are outlining this online research endeavor designed for automatized psychometric data collection and scalable assessment: once set up, the low costs and expenditure associated with individual psychometric testing allow substantially increased study cohorts and thus contribute to enhancing study outcome reliability. We are leveraging gamification of the data acquisition method to make the tests suitable for online administration. By putting a strong focus on entertainment and individually tailored feedback, we aim to maximize subjects' incentives for repeated and continued participation. The objective of measuring repeatedly is obtaining more revealing multitrial average scores and measures from various operationalizations of the same psychological construct instead of relying on single-shot measurements. COSMOS is set up to acquire an automatically and continuously growing dataset that can be used to answer a wide variety of research questions. Following the principles of the open science movement, this data set will also be made accessible to other publicly funded researchers, given that all precautions for individual data protection are fulfilled. We have developed a secure hosting platform and a series of digital gamified testing instruments that can measure theory of mind, attention, working memory, episodic long- and short-term memory, spatial memory, reaction times, eye-hand coordination, impulsivity, humor appreciation, altruism, fairness, strategic thinking, decision-making, and risk-taking behavior. Furthermore, some of the game-based testing instruments also offer the possibility of using classical questionnaire items. A subset of these gamified tests is already implemented in the COSMOS platform, publicly accessible and currently undergoing evaluation and calibration as normative data is being collected. In summary, our approach can be used to accomplish a detailed and reliable psychometric characterization of thousands of individuals to supply various studies with large-scale neurocognitive phenotypes. Our game-based online testing strategy can also guide recruitment for studies as they allow very efficient screening and sample composition. Finally, this setup also allows to evaluate potential cognitive training effects and whether improvements are merely task specific or if generalization effects occur in or even across cognitive domains.

**Keywords** Gamification · Smart data acquisition · Online phenotyping · Next-generation high-throughput psychometrics

✉ Christian Vogler  
christian.vogler@unibas.ch

<sup>1</sup> Department of Psychology, Division of Molecular Neuroscience, University of Basel, Basel, Switzerland

<sup>2</sup> Program in Placebo Studies and the Therapeutic Encounter, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, USA

<sup>3</sup> Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital/Harvard Medical School, Boston, MA, USA

<sup>4</sup> Department of Clinical Psychology & Psychotherapy, University of Basel, Basel, Switzerland

<sup>5</sup> Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

<sup>6</sup> ZJU-UoE Institute, Zhejiang University School of Medicine, Haining, Zhejiang, China

<sup>7</sup> Psychiatric University Clinics, Basel, Switzerland

<sup>8</sup> Department of Psychology, Division of Cognitive Neuroscience, University of Basel, Basel, Switzerland

## Introduction

Objectively measuring inter- and intra-individual differences in human behavior is a fundamental core mission in psychology as it provides the solid fundament on which the entirety of research endeavors in psychology and related fields depend upon (Jenkins and Lykken 1957). The availability of accurate, reliable, and comprehensive phenotypic measures is not only essential for psychological hypothesis testing per se, but is also crucial for the successful elucidation of biological underpinnings of neurocognitive traits that are amenable to for instance imaging or genetic studies (Congdon et al. 2010). While computers have already been used to assist in test evaluations for more than half a century (Kleinmuntz 1963), advances in computer technology now allow for the development of completely digitalized assessment strategies with automated scoring and evaluation procedures (Luciana 2003). Automatization of psychometric assessment is a highly valuable approach for meeting for example the demands that are put forward by the recent revolutions in biotechnology: while high-throughput cost- and time-efficient individual whole genome scans in large cohorts have become a matter of course, phenotypic assessments typically still rely on laborious testing batteries, often requiring trained administrators and stationary attendance time of study participants.

We argue that bringing down the effort for both researchers and testees involved in collecting repeated phenotypic measurements of healthy large cohorts is feasible through online-based test administration. Yet, this requires a substantial redesign and redevelopment of psychometric assessment procedures and instruments. Conceptualizing the novel strategies for large-scale assessments should be led by the idea that participant compensation is essential and constituted by existing ethical guidelines yet does not necessarily need to be monetary. Entertainment that can be achieved through gamification and task design is not only a highly valued benefit itself, it is also key to nurse the participant's motivation required for repeated measurements (Lumsden et al. 2016). Designing the data collection process as a rewarding experience itself is a valuable strategy, as previous studies have found that the demanding nature of data entry is one of the primary reasons respondents stopped using health apps (Krebs and Duncan 2015). Additionally, automation of data collection and evaluation can be used to provide test persons with graphically illustrated feedback on their own performance as this also serves as an incentive for repeated and continuous participation. Finally, computer game-based tests and experiments provide scientists with a novel technique to test ecological validity of laboratory-based procedures, which is always assumed, but rarely tested (Krakauer et al. 2017).

A large online-based research platform that collects sensitive personal data requires continuous attention and efforts to ensure the best possible standard of security for safeguarding participants' personal data from security gaps and potential misuse. It is a question of respect towards the study participants to view and

treat the gathered data as a good that the scientist is only entrusted with for conducting research, but that ultimately still belongs to the testees. The fact that the data might be used in currently undefined future research projects or may yield to potential monetization of research outcomes calls for more control options through participants during the data life cycle than a single "open ended" consent form (Lipworth et al. 2017). Yet, despite these concerns, using a single platform framework to simultaneously obtain a wide variety of different psychometric data comes with a set of very appealing options: based on the concepts of "open-science," "open-data," and collaboration, we outline our prototype for automatized and smart phenotypic data acquisition, which holds the potential for reshaping standard procedures in psychological research practice and for facilitating productivity and study outcome reliability. Specifically, we plan to implement a pre-registration system that grants publicly funded scientists' script-based access to the collected data through the COSMOS platform. Scientists can develop their scripts on a dummy database system that mimics the database system of the COSMOS backend. Relying on script-based analyses, which will be run in a secure environment and only return the result of the analysis, is a safety precaution which eliminates the need to grant access to raw data. Only revealing combined and summarized data still allows making highly flexible and efficient use of the existing data pool, while maximizing the security of the dataset against identifying individual test participants.

Conveniently, the ongoing automatic data acquisition continuously generates novel samples that can be used for effortlessly replicating the obtained findings as soon as a large enough additional batch of data has been collected. Additionally, the comparably low maintenance and personnel costs of data gathering can contribute to alleviate the time-consuming competition over limited funding resources. At the same time, the centralization of longitudinal data gathering enables a higher phenotypic resolution per individual than single studies could achieve. The large N high-resolution data allows building models of higher complexity that are better suited to account for confounding factors, which typically would be out of scope for small N single-hypothesis testing study designs. Depending on the respective research question and the hypothesis tested, the available detailed assessment of a large number of individuals allows the application of sampling strategies that either are currently not taken into consideration at all or are only feasible at large expenses of cost and time: preselecting subgroups as homogenous as possible, closely matching experimental groups on potentially confounding factors, evaluating whether a detected correlation can be found in a set of different subgroups, whether it is largely stable or may be even reversed along the continuum of the normal distribution of a given trait.

Finally, platforms like COSMOS can facilitate settling the question, whether so-called brain training (i.e., repeatedly engaging in cognitively demanding tasks) can actually have generalizing beneficial effects: based on a large N, without taking

money from the participants and thus without the inherent conflict of interest the brain training industry-affiliated scientists are faced with.

## Game Tests

The COSMOS platform (<https://cosmos.psycho.unibas.ch/>) is now in its pilot phase, hosting five prototypes of games that currently undergo refinement and calibration as psychometric testing instruments, which are described in more detail below. Table 1 gives an overview of all developed instruments together with the phenotypic constructs they have been designed to measure.

### HoNk-Back

The HoNk-Back task is a gamified redesign of one of the most widely used working memory tasks in neuroscience, the N-Back paradigm (Owen et al. 2005). This gamification of the task goes beyond simply adding game-like reinforcement mechanics such as a score or a progress bar. We put special attention on developing a setting that lets the actual task of monitoring a sequence of stimuli appear as natural and plausible as possible, aiming at increasing ecological validity. The task setting makes the test subject to assume the role of a truck driver who gets overtaken by a constant stream of cars. Cars appearing in the review mirror trigger the required response signal by the truck driver which consists of either flashing the headlights at cars that also gave a light signal or waving at the cars that overtook the truck without emitting a headlight signal. Tilting of the rearview mirror controls the N condition as this allows regulating the number of cars disappearing into the blind spot.

### Drag Race

This test in form of a drag race game is designed to measure reaction times to unpredictable and predictable cues and variation in response time accuracy. A light signal sequence of two yellow lights indicates that the driver has to get ready. The green light that indicates the take-off signal then is given after a variable random time interval allowing the measurement of spontaneous reaction time (SRT). The process of shifting gears requires a defined motor response pattern: releasing the accelerator button (spacebar), hitting the gear-shifting button (return) and releasing it again, and pushing the accelerator button again. The gear-shifting procedure is used to record the response times to predictable signals: the revmeter continuously moves towards the optimal switching moment, when the response pattern has to be executed. This allows measuring several reaction times of simple motor responses in the form of foreseeable reaction times (FRT). Evaluating repeated runs allows assessing variation in response time accuracy. We are aware

of software and/or hardware-related issues concerning reaction time measurements such as monitor response time, operating system design, and input device-related delay such as key debouncing time (Garaizar et al. 2014; Salmon et al. 2017) that impact the accuracy of the response time measurements. Nevertheless, the provided test should yield rough estimates of individual response times and allow group comparisons under the assumption of equally distributed noise. Also the argument has been brought forward that the error introduced by response devices is bound to be small relative to human variability and will only exert potential effects in experiments that lack statistical power in the first place (Damian 2010). Given that the game will be made freely available as a standalone application, it can serve as test instrument in a controlled lab-based environment with identical hard- and software setups allowing unbiased inter-individual comparisons.

### Frog Life

Frog Life is a combinatorial task with increasing difficulty levels consisting of a go/no-go paradigm to measure sustained attention and additionally assesses visual vigilance. The task setting lets participants control a virtual frog in a pond that feeds on dragonflies (go-condition) while avoiding devouring hornets (no-go condition). Simultaneously, the testee needs to escape predators, which are announced through changes in coloring of three different display details, namely the color of the water in the pond, the clouds, or a depicted bush (Fig. 1). Insects only become catchable after they entered the proximity range outlined by a spherical contour around the frog. Snatching of the insects is achieved by pressing the corresponding left or right cursor buttons of the keyboard depending on which side of the screen the insects emerged from. Correct responses of the go-task (eating dragonflies) are rewarded with increasing of the score, while incorrect responses to the no-go condition (eating hornets) decreases the score. Color changes of one of the three display details announce an upcoming predator and require the player to trigger an escape jump by pressing the spacebar. Faster reaction times to the color changes are rewarded with more points. Yet, pressing the spacebar while no actual color change is taking place causes the player to lose one of three health points indicated by hearts. If all health points are lost, the player character is granted “game-over.” After every successful escape, the game mechanics difficulty level is increased. In case the player fails to detect a color change of the display details, appearance of a predator terminates the game. The color change thus constitutes an additional go/no-go task based on signal detection.

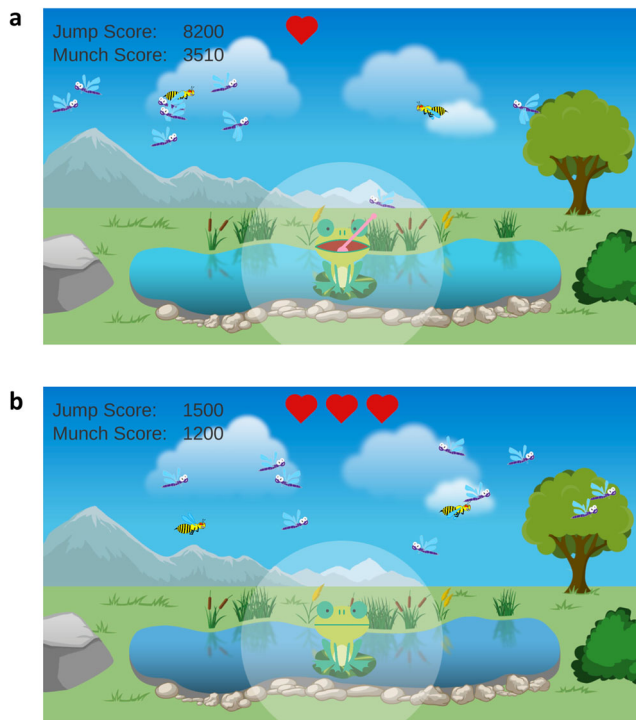
### Shortcuts

This game is designed as a two-tier short-term memory performance assessment consisting of an episodic picture recognition

**Table 1** Overview of all games that have been developed for the COSMOS platform to date. The column “Features” shortly describes game setup and content. Also listed are approximate durations to complete one testing unit (depending on the game type, the test unit refers to a single level or to a complete game). “Technological specifications” lists the IT technologies/programming languages that were used to create the game and its backend

Game name	Target phenotypes	Features	Implemented	Approx. duration per test unit	Technological specifications
BANKWORLD	Risk-taking behavior, decision-making, strategic thinking	Text-based; managing financial projects that requires social investing to minimize risk of adverse events	Yes	120–360 s	HTML5, CSS3, RestAPI, MySQL
BRAINLAB	Spatial memory	3D, first-person perspective, free roaming; solve a series of tasks while staying oriented in a multileveled maze	No	120–600 s	Unity, MySQL
CAKE	Altruism, fairness, trust, suspiciousness	Text-based; complex social interaction game based on advanced ultimatum game scenarios	No	60–120 s	HTML5, CSS3, MySQL
DRAGRACE	Reaction times, eye-hand coordination	2D; cartoon-styled drag race game	Yes	15–80 s	Unity, MySQL
FROGLIFE	Attention, reaction times, impulsivity, eye-hand coordination	2D; keep a frog alive by escaping predators while feeding on dragonflies and avoiding swallowing hornets	Yes	60–180 s	Unity, MySQL
HONK-BACK	Attention, working memory	3D, first-person perspective; react differently (honk or wave) based on whether an overtaking car flashed its headlight	Yes	30–180 s	HTML5, CSS3, JavaScript, MySQL
INVESTMENT BANKING	Decision-making, risk-taking behavior, multitasking	2D; optimize revenue achieved by a team of bankers through individually rewarding while avoiding reward over-saturation	No	300–600 s	HTML5, CSS3, JavaScript, MySQL
JOYRATE	Theory of mind, humor appreciation	Text-based; guess humor appreciation of others based on how they rated statements, covering different topics like politics, personality, etc.	Yes	120–300 s	HTML5, CSS3, xAPI, MySQL
MEMORY RACER	Episodic long- and short-term memory, eye-hand coordination	3D, racing game; rate and memorize pictorial or text-based stimuli presented during gameplay	No	60–120 s	Unity, MySQL
SHORTCUTS	Episodic long- and short-term memory	3D, panda bear climbing mountains; player needs to use recognition memory for time-saving shortcuts in the climbing route	Yes	90–300 s	Unity, MySQL





**Fig. 1** Scenery examples from the game “Froglife.” Dragonflies and hornets constitute a go/no-go paradigm. Eating dragonflies increases the “Munch Score,” eating hornets decreases it: upon entering the white circle that is surrounding the frog, the player can make the frog eat the insects using the left and right cursor button, depending on the side from which the insects are entering the white circle. **a** The dragonfly entered from the right side is captured by pressing the right arrow cursor button. **b** Color changes of the pond, the bush on the right side or the clouds indicate an approaching predator requiring the player to escape the current scenery by jumping to a different pond using the spacebar. In the depicted scene, the hue of the pond is changed

task and a sequence-learning test. At the beginning of every game round, testees need to memorize a set of picture stimuli (3 to 10 items). The test participants can choose between different categories, such as “food,” “animals,” and “sport,” and will be presented a set of pictures to memorize for the current game round. Additionally (apart from the “easy” condition featuring only three pictorial items to be remembered), a sequence of differently colored and shaped symbols is presented at the beginning of the game round. The accurate encoding of sequential information is a key cognitive element in human cognition, setting humans apart from other species, but also shows large variation in performance within species (Ghirlanda et al. 2017). During the actual game, the player controls a panda bear climbing rock walls that gets rewarded for correctly solving recognition tasks: at given intervals, the player is presented with a selection of pictures and required to identify the picture shown in the beginning. If she clicks on the correct picture, a bird lifts the panda bear to a higher position in the climbing wall thus rewarding the player with “shortcuts” in the climbing route. Also, at predefined intervals, buttons appear on the screen that require the player to reproduce the symbol sequence that was

shown at the beginning of the level. It must be entered correctly in order for the climbing to continue. If the sequence is not entered correctly by the player, the entire sequence will be displayed, so that the player can continue. The player is awarded with points for correct recognition of the pictures, reproducing the symbol sequence correctly and for speed.

### Joyrate

This task is primarily designed to measure a subtype of theory of mind (ToM) by employing entertaining stimulus material. At the beginning of the game, the player is asked to rate the jocularity of 10 items consisting of cartoons, memes, and written jokes on a scale from 0 to 10. Additionally, the participants also rate how strongly they agree with 18 statements touching topics such as politics, religion, society, sports, education, and personality. This initial phase has to be completed only once. The actual game then consists in guessing as how amusing a given item has been rated by another person whose ratings the player is randomly assigned to. Whenever an item appears from the joke or the statement pool that has not yet been answered by the player himself, he is asked to make his own rating prior to estimating/learning the estimation of his/her counterpart. This way, the pool of rated jokes and statements for all individuals is constantly increased. The goal of the game is to estimate as accurately as possible how entertaining a given stimuli was perceived by the other person. Apart from the demographic info on the other player that is always provided (gender, age, education), the participant can unlock further information on how the statements were rated using an in-game-generated currency (JokeCoins). The accuracy of the estimation process is rewarded with points and JokeCoins.

## COSMOS Environment

### Individual Data Visualization

All gamified testing instruments developed in the scope of the COSMOS project feature an application-specific relational SQL database that records the user’s input. This makes it easy to set up any application as a standalone implementation and to integrate the applications into a specific laboratory test setting, for example, as a subtest in a given brain mapping experiment. In the scope of the COSMOS web platform, all user data is assigned to unique identifier codes (UIC) and thus to a certain person, by means of a separate central authentication system implemented in the secure software framework used to host the website. All single SQL databases are linked via experience APIs to a Learning Record System employing a mongoDB that serves the purpose of graphically representing the obtained data. We have developed a data visualization application that allows platform administrators fast and easy creation and configuration

of interactive plots. COSMOS participants can choose from a variety of preconfigured plots to learn about their individual performance over time, compare their scores to all participants or to specific subgroups only, e.g., a given age range or gender (examples of plots are depicted in Fig. 2). Visualization of achieved high scores and selected performance measures like for instance average reaction times also allows COSMOS participants to monitor their performance over the course of the day to identify peak performance time periods when they usually achieve the best concentration and attention levels.

### Automated Data Processing Pipeline

The independent SQL databases that all games running on the COSMOS platform are equipped with facilitate a streamlined and automated data analysis pipeline. While there may be specific deviations for single games, the general rule is that data will be marked as an unfinished

run or simply not stored in the database, if the level was aborted due to player inactivity, closing of the browser, or loss of internet connection. All user responses and summary statistics generated by the games are recorded and stored along with a timestamp and linked to a specific UIC in the games' databases. The UIC is generated when an account is registered and thus pertains to specific login credentials. This procedure allows data to be uniquely assigned to a specific person and therefore enables data collection over multiple trials, time-points, levels, and different tasks. Since the exact timestamp of each reaction is always stored in the database, it is easy to calculate for example the average reaction time per game round: large intervals between stimulus presentation and the reaction of the player or a large variance in task performance indicators can be used to detect a lack of concentration or distraction and thus can be used to create QC filters. Of course, those statistical filters

**Fig. 2** Example of a typical visualization of test results generated by the mongoDB-based visualization feature of the COSMOS web platform. The generated graphical representations are partially configurable and allow the user to customize which data is displayed. Any data fed into the mongoDB can be visualized in either bar charts, pie charts, or progress charts. Database schema of the game Frog Life



themselves can be evaluated if participants are asked to rate for example the attention or level of concentration they were exhibiting during gameplay after a level is completed.

The use of standard SQL databases allows accessing the data with all common statistical analysis tools/languages like R, python, matlab, octave (Eaton et al. 2014; MATLAB Optimization toolbox 2017; R Core Team 2018), etc. This allows the creation of standard query scripts that are customizable to retrieve the data best suited to answer a given research question: e.g., retrieve all data for game x, y, and z for all individuals meeting a given age range, gender, or educational level that finished at least 10 trials per game within a specified time period. The exact procedure of reading, processing, summarizing, and blending data may of course depend on the specifics of the research question to be answered.

Figure 3 depicts a description of the SQL database schema for the game Frog Life. This description together with the information on the different response types (as shown in Table 2) helps understanding how simple database queries can be used to sum up different correct answers and/or errors depending on the difficulty level of the task in order to serve as a data basis for modeling.

## Modeling Phenotypes

In order to understand and analyze complex behaviors, a promising approach has been computational models (Corrado and Doya 2007; Luksys and Sandi 2011; Mars et al. 2012; Nassar and Frank 2016). Most widely popularized in the field of reinforcement learning (Tanaka et al. 2004; Daw



**Fig. 3** The SQL database schema for Frog Life. The games-table records all the games (numbered incrementally starting from 1) a given user (user\_id defined by the UIC) has played, along with the scores s/he achieved and the timestamp the game was started (creation\_time) and finished (modification\_time). The finished field contains the info whether the game was normally finished or prematurely terminated. All lines between the tables are dotted, since the UIC serves as foreign key for all other tables. The rounds-table contains information about every single round played as indicated by the “one to one or many” relationship (since many rounds per game are possible). A round starts either directly at the beginning of the game or after the player escaped an upcoming predator

and jumped to a new scenery. After every round, the difficulty level is increased (and the current difficulty level gets stored in the difficulty field), i.e., the speed of the insects accelerates, the color change time decreases, and the hue intensity change gets less pronounced. The action table stores all the actions that are exhibited by the player during a given round. The action\_types-table comprises all the possible response types a player can display (see Table 1 for action type definitions). The levels-table holds the information about the background sceneries, which is recorded in the rounds-table (level\_id). Currently, three different sceneries are available



**Table 2** Description of possible action types a user can display during playing Frog Life. Only if a given response as defined by an action type is exhibited, one of the described database entries in the Description column is triggered. Thus, e.g., if no entry for

action type HORNET\_EATEN exists in a given round table, the player did not make this type of mistake during that round. Type describes the psychometric characteristics attached to the potential user responses

Action	Description	Type
DRAGONFLY_EATEN	Time difference between required action trigger and correct response	Correct response
DRAGONFLY_NOT_EATEN	Time span of omission error	Omission error/go-error
HORNET_EATEN	Time difference between no-go trigger and incorrect response	No-go error
HORNET_NOT_EATEN	Time span of correct omission	Correct response
WRONG_LEFT_PRESS	Timestamp of pressing the opposite of the required arrow key	Motor control error
WRONG_RIGHT_PRESS	Timestamp of pressing the opposite of the required arrow key	Motor control error
COLOR_REACTION	Time difference between upcoming predator warning and pressing the spacebar (finish current round)	Correct response
COLOR_NO_REACTION	Time difference between upcoming predator and game over	Visual attention error
CAUSELESS_JUMP	Timestamp of pressing spacebar without predator approaching	Visual threshold error
ARROW_NO_INSECT	Timestamp of pressing an arrow button in vain	Motor control error

et al. 2006; Behrens et al. 2007; Frank et al. 2007; Luksys et al. 2009), they have also been applied to study working (Collins and Frank 2012; Collins et al. 2014) and episodic (Luksys et al. 2014, 2015) memory as well as decision-making (Forstmann et al. 2008), including strategic reasoning (Zhu et al. 2012; Seo et al. 2014). The main principle is that a computational model is fitted to experimental data (based on how well model-produced behaviors match experimentally observed ones), and then the best-fitting model parameters and/or variables are used as correlates for neurobiological data such as neuron recordings (Samejima et al. 2005), fMRI activations (Tanaka et al. 2016; Daw et al. 2006, Behrens et al. 2007), genetic differences (Frank et al. 2007; Set et al. 2014; Luksys et al. 2014, 2015), levels of stress (Luksys et al. 2009), and neuropsychiatric disorders (Collins et al. 2014). The main advantage of model-based analysis is that it can test neurocomputational mechanisms of behavior, which different candidate models aim to represent, and reduce a variety of behavioral measures, which can strongly depend on the specific task, to fewer model parameters that are directly comparable between the tasks. For example, reinforcement learning can model behavior in a number of tasks where rewards or punishments of some kind (sometimes implicit) are involved, and despite different formalizations, most of these models have common parameters such as the learning rate, exploration-exploitation tradeoff, and future discounting (Tanaka et al. 2004; Frank et al. 2007; Schweighofer et al. 2008; Luksys et al. 2009). Due to unusual richness of the acquired data, gamification provides a special opportunity to convincingly show usefulness of computational models compared to traditional analyses of behavior, and most importantly link platform-derived behaviors to laboratory-based tasks, which can be analyzed using more simple models that share parameters with more complex models of games. Where

explicit modeling of games is not practical (e.g., due to their complexity), the recorded patterns of game-derived data could be linked to laboratory-based behaviors (or their model parameters) using machine learning tools. Finally, a game-based psychometric assessment platform such as COSMOS provides a unique chance to test and compare different candidate models using a much wider variety of tasks and populations than used in most model-based analysis studies, where usually a narrow range of models are tested against each other based on one or few tasks selected by authors (which may benefit their favorite models compared to alternatives).

## Discussion

The pervasive problem of low-powered studies in the behavioral and social sciences leading to non-replicable and spurious findings has already been identified more than 60 years ago. Yet today, it does not only still persist, but is even being exacerbated by system design faults (Ioannidis 2015; Smaldino and McElreath 2016; Szucs and Ioannidis 2017): using the amount of published original research as a quality criterion for awarding funding or tenured positions incentivizes increasing the number of publications. This creates a conflict of interest with the researcher's intrinsic goal to maximize study outcome reliability. In addition, the novelty of findings based on a small number of observations is often valued higher than replication in large cohorts (Higginson and Munafa 2016; Nosek et al. 2015; Vinkers et al. 2015). In combination with short-term contracts for the junior scientific staff (Kreeger 2004; Langenberg 2001) that render planning and implementing of larger-scale projects almost impossible as they require substantially more time than single-hypothesis small N studies, the scientific community has formed an

optimal hotbed for keeping the well-known problems alive and prospering.

The ongoing replication crisis of psychological research presses us to figure out how the above-mentioned systemic shortcomings can be overcome. Luckily, computerization and automatization in combination with interdisciplinary cooperation and an open-data philosophy offer a solution to a very basic but crucial problem: in our eyes, sample size is the elephant in the room for improvement of psychological research that needs to be addressed promptly. Towards this end, we have initiated the COSMOS platform: we are striving to facilitate recruitment of study participants through automatization, i.e., creating experimental setups that no longer require staff to implement them and to observe and record behavior. Although our platform is still in its pilot phase, we argue that digitally oriented research endeavors like COSMOS will eventually serve the scientific community in several ways. Online screening platforms can be used to either carefully preselect individuals or to simply increase sample size without skyrocketing costs. Being able to substantially increase the number of study participants is arguably a compelling strategy to counteract the overestimation of effect sizes and the non-replicability of study findings.

For the scaling of psychometric assessment, especially for the online-based test setting, our overall philosophy is that the testing instruments need to be as fun and absorbing for the participants as possible to increase the intrinsic motivation to engage. At the same time, tests should require minimal effort with regard to manual data entry in order to prevent significant issues with subject adherence. Finally, the novel assessment tools should provide investigators in the psychological and biomedical sciences with research-grade cognitive and psychological metrics. Technological advances along with a strongly grown computer literacy in the general population and widespread familiarity with computer games (Granello and Wheaton 2004; Palaus et al. 2017) open up a plethora of possibilities for the operationalization of psychological research questions. Leveraging gamification to repeatedly obtain behavioral samples paves the way for a next-generation high-throughput psychometric toolset. Hence, the COSMOS platform is conceptualized to collect a vast array of psychometric and cognitive data from a large pool of study participants in a highly automated and thus very cost- and time-efficient way.

It is obvious that the goal of gathering in-depth phenotypic data by employing web-based administration of psychometric tests in the guise of entertaining serious games chaperoned by individual automatic performance feedback requires a highly interdisciplinary skill set: social, computer, and data scientists need to work closely together to design, develop, refine, and validate the tools and put them to work. Yet, this aggregate competence is often readily available in university settings and easily accessible through close collaborations between disciplines.

The possibilities of the outlined web platform go way beyond the scope of only gathering data, if additional opportunities offered by the digital era are harnessed: it could also provide a framework to present, discuss, and continuously update scientific findings. We think that eventually such approaches will not only help online participants better understand their own behavior and detect patterns that may be early signs of neuropsychiatric disorders; they could also open up venues for the development of efficient, individualized, and most importantly scientifically sound methods of cognitive enhancement.

**Acknowledgments** COSMOS (<https://cosmos.psych.unibas.ch/>) was started by the Divisions of Molecular and Cognitive Neuroscience of the University of Basel in close cooperation with the Informatics Department of the University of Applied Sciences Northwest-Switzerland.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.
- Corrado, G., & Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(31), 8178–8180. <https://doi.org/10.1523/JNEUROSCI.1590-07.2007>.
- Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European Journal of Neuroscience*, *35*, 1024–1035.
- Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *The Journal of Neuroscience*, *34*, 13747–13756.
- Congdon, E., Poldrack, R. A., & Freimer, N. B. (2010). Neurocognitive phenotypes and genetic dissection of disorders of brain and behavior. *Neuron*, *68*, 218–230. <https://doi.org/10.1016/j.neuron.2010.10.007>.
- Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods*, *42*(1), 205–211. <https://doi.org/10.3758/BRM.42.1.205>.
- Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. <https://doi.org/10.1038/nature04766>.
- Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007a). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. <https://doi.org/10.1073/pnas.0706111104>.

- Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, 105(45), 17538–17542. <https://doi.org/10.1073/pnas.0805903105>.
- John W. Eaton David Bateman, S. H., & Wehbring, R. (2014). {GNU Octave} version 3.8.1 manual: A high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. Retrieved from <http://www.gnu.org/software/octave/doc/interpreter>.
- Garaizar, P., Vadillo, M. A., López-De-Ipiña, D., & Matute, H. (2014). Measuring software timing errors in the presentation of visual stimuli in cognitive neuroscience experiments. *PLoS One*, 9(1), e85108. <https://doi.org/10.1371/journal.pone.0085108>.
- Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science*, 4(6). <https://doi.org/10.1098/rsos.161011>.
- Granello, D., & Wheaton, J. (2004). Online data collection: Strategies for research. *Journal of Counseling & Development*, 82(4), 387–393. <https://doi.org/10.1002/j.1556-6678.2004.tb00325.x>.
- Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, 14(11), e2000995. <https://doi.org/10.1371/journal.pbio.2000995>.
- Ioannidis, B. J. P. A., & Sc, D. (2015). Failure to replicate: Sound the alarm. *Cerebrum*, 2015 (November), 1–12. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/27420921>. Accessed 21 June 2017.
- Jenkins, J. J., & Lykken, D. T. (1957). Individual differences. *Annual Review of Psychology*, 8(1), 79–112. <https://doi.org/10.1146/annurev.ps.08.020157.000455>.
- Kleinmuntz, B. (1963). Personality test interpretation by digital computer. *Science*, 139(3553), 416–418. <https://doi.org/10.1126/science.139.3553.416>.
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marín, A., Maciver, M. A., & Poeppel, D. (2017). Neuron perspective neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93, 480–490. <https://doi.org/10.1016/j.neuron.2016.12.041>.
- Krebs, P., & Duncan, D. T. (2015). Health App Use Among US Mobile Phone Owners: A National Survey. *JMIR MHealth and UHealth*, 3(4), e101. <https://doi.org/10.2196/mhealth.4924>.
- Kreeger, K. (2004). Short-term limbo. *Nature*, 427(6976), 760–761.
- Langenber, H. (2001). Uncertainty of short-term contracts is turning talent away from science. *Nature*, 410(6830), 849–850.
- Lipworth, W., Mason, P. H., & Kerridge, I. (2017). Ethics and epistemology of big data. *Journal of Bioethical Inquiry*, 14(4), 485–488. <https://doi.org/10.1007/s11673-017-9815-8>.
- Luciana, M. (2003). Practitioner review: Computerized assessment of neuropsychological function in children: Clinical and research applications of the Cambridge Neuropsychological Testing Automated Battery (CANTAB). *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 44(5), 649–663.
- Luksys, G., Gerstner, W., & Sandi, C. (2009). Stress, genotype and nor-epinephrine in the prediction of mouse behavior using reinforcement learning. *Nature Neuroscience*, 12(9), 1180–1186. <https://doi.org/10.1038/nn.2374>.
- Luksys, G., & Sandi, C. (2011). Neural mechanisms and computations underlying stress effects on learning and memory. *Current Opinion in Neurobiology*. <https://doi.org/10.1016/j.conb.2011.03.003>.
- Luksys, G., Ackermann, S., Coynel, D., Fastenrath, M., Gschwind, L., Heck, A., de Quervain, D. (2014). BAIAP2 is related to emotional modulation of human memory strength. *PLoS One*, 9(1), e83707. <https://doi.org/10.1371/journal.pone.0083707>.
- Luksys, G., Fastenrath, M., Coynel, D., Freytag, V., Gschwind, L., Heck, A., de Quervain, D. J.-F. (2015). Computational dissection of human episodic memory reveals mental process-specific genetic. *Proceedings of the National Academy of Sciences of the United States of America*.
- Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, 4(2), e11. <https://doi.org/10.2196/games.5888>.
- Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. S. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology*, 65(2), 252–267. <https://doi.org/10.1080/17470211003668272>.
- MATLAB Optimization toolbox. (2017). Retrieved from <https://ch.mathworks.com/products/matlab.html>.
- Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*. <https://doi.org/10.1016/j.cobeha.2016.04.003>.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>.
- Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25, 46–59. <https://doi.org/10.1002/hbm.20131>.
- Palau, M., Marron, E. M., Viejo-Sobera, R., & Redolar-Ripoll, D. (2017). Neural basis of video gaming: A systematic review. *Frontiers in Human Neuroscience*, 11, 248. <https://doi.org/10.3389/fnhum.2017.00248>.
- R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>.
- Salmon, J. P., Jones, S. A. H., Wright, C. P., Butler, B. C., Klein, R. M., & Eskes, G. A. (2017). Methods for validating chronometry of computerized tests. *Journal of Clinical and Experimental Neuropsychology*, 39(2), 190–210. <https://doi.org/10.1080/13803395.2016.1215411>.
- Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Neuroscience: Representation of actionspecific reward values in the striatum. *Science*, 310(5752), 1337–1340. <https://doi.org/10.1126/science.1115270>.
- Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S. C., Yamawaki, S., & Doya, K. (2008). Low-Serotonin Levels Increase Delayed Reward Discounting in Humans. *Journal of Neuroscience*, 28(17), 4528–4532. <https://doi.org/10.1523/JNEUROSCI.4982-07.2008>.
- Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, 346(6207), 340–343. <https://doi.org/10.1126/science.1256254>.
- Set, E., Saez, I., Zhu, L., Houser, D. E., Myung, N., Zhong, S., et al. (2014). Dissociable contribution of prefrontal and striatal dopaminergic genes to learning in economic games. *Proceedings of the National Academy of Sciences of the United States of America*, 111(26), 9615–9620. <https://doi.org/10.1073/pnas.1316259111>.
- Smaldino, P. E., & McElreath, R. (2016). The Natural Selection of Bad Science. *Royal Society Open Science*, 3(9). <https://doi.org/10.1098/rsos.160384>.
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15(3), 1–18. <https://doi.org/10.1371/journal.pbio.2000797>.
- Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7, 887. <https://doi.org/10.1038/nn1279>.

- Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2016). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. In Behavioral Economics of Preferences, Choices, and Happiness (Vol. 7, pp. 593–616). [https://doi.org/10.1007/978-4-431-55402-8\\_22](https://doi.org/10.1007/978-4-431-55402-8_22).
- Vinkers, C. H., Tjebk, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, *351*, h6467. <https://doi.org/10.1136/bmj.h6467>.
- Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences*, *109*(5), 1419–1424. <https://doi.org/10.1073/pnas.1116783109>.