Comparing Bayesian Statistics and Frequentist Statistics in Serious Games Research

Comparing Bayesian Statistics and Frequentist Statistics in Serious Games Research

Wim Westera[1]
[1] *Open University of the Netherlands, wim.westera@ou.nl*

Abstract

This article presents three empirical studies on the effectiveness of serious games for learning and motivation, while it compares the results arising from Frequentist (classical) Statistics with those from Bayesian Statistics. For a long time it has been technically impracticable to apply Bayesian Statistics and benefit from its conceptual superiority, but the emergence of automated sampling algorithms and user-friendly tools has radically simplified its usage. The three studies include two within-subjects designs and one between-subjects design. Unpaired t-tests, mixed factorial ANOVAs and multiple linear regression are used for the analyses. Overall, the games are found to have clear positive effects on learning and motivation, be it that the results from Bayesian Statistics are more strict and more informative, and possess several conceptual advantages. Accordingly, the paper calls for more emphasis on Bayesian Statistics in serious games research and beyond, as to reduce the present domination by the Frequentist Paradigm.

Keywords: *Statistics, Methodology, Bayes, Motivation, Learning, Games*

1 Introduction

For assessing the efficacy of games for learning sound empirical studies are indispensable, including rigorous statistical procedures. For a long time, the predominant statistical analysis paradigm in serious games research (and educational research at large) has been Frequentist Statistics, which is the type of statistics covered by most statistical textbooks, statistical software tools and social sciences teaching programmes. However, methodologists have severely disqualified Frequentist Statistics as being a deceptive or even an "embarrassing and mindless ritual" leading to wrong interpretations and false claims [1], for instance unjust claims as simple as "...to accept the null-hypothesis" [2]. In contrast, the alternative option of Bayesian Statistics has a logical foundation that allows for computing and updating probabilities in a straightforward way after obtaining new data; it allows for comparing competing predictions in a superior manner without the risk of deception, misconceptions or false claims. Notwithstanding these claimed advantages of Bayesian Statistics, Frequentist Statistics still is the prevailing paradigm in psychological and educational studies, being presented as the golden standard in lectures, student textbooks and scientific journals. This is not just a matter of established tradition, but also due to the fact that Bayesian Statistics has been less practical to apply, because it requires to obtain its probability distributions in closed form, which is rarely possible. In recent years, however, the introduction of the Markov Chain Monte Carlo method (MCMC) [3], which is a random-walk-based data sampling technique to generate probability distributions, has greatly simplified the practical application of Bayesian Statistics by overcoming the need for closed form analytic solutions. Bayesian procedures using MCMC are now becoming available as standard utilities in various statistical packages. Hence, it is the right moment to explore the opportunities that Bayesian Statistics offers and present practical application cases to the wider research community. The principal research

question under consideration is ”how the outcomes of Bayesian Statistics compare with those obtained from Frequentist Statistics in the serious games' educational research practice”. Although our data originate from serious games, our findings may be of relevance for the wider field of educational media research, since the investigated statistical procedures remain unaware of the nature of the instructional tools used.

To exemplify and substantiate the Bayesian claims, this article presents three original serious game studies on learning outcomes and motivations, while it compares and discusses the outcomes resulting from both statistical methodologies. The focus on serious games is motivated by the fact that 1) games are advanced, rich and dynamic learning environments that can cover a wide variety of learning scenarios and thus effectively represent the wider field of learning and teaching approaches, 2) like any emerging instructional tools games require sound effect studies as to provide empirical evidence about their effectiveness.

So far, the ever-growing popularity of games for leisure and entertainment has positively influenced the use of games for learning and training purposes. Games have successfully been applied in schools to promote media literacy [4,5], to reduce student dropout [6,7], and to enhance motivation [8,9], to accommodate flipped classrooms [10], and many other things. A large body of evidence has become available corroborating the effectiveness of games as instructional tools: many game studies have reported positive effects on learning outcomes and motivation. This scientific evidence is crucial to overcome existing barriers for adoption, as many teachers have their reservations about replacing their traditional materials and exercises with supposed game-based frivolities. Teachers' skepticism is readily fueled by the host of over-enthusiastic game proponents and believers [11], who unceasingly present games as the panacea for solving all contemporary problems in schools. Therefore, unbiased and scientifically grounded validation studies of games are a persistent requirement: the scientific method is the best if not the only form of rational inquiry, providing the best possible, objective and unbiased answers to the questions posed. But even well-established scientific methodologies, such as Frequentist Statistics, have their flaws and should be critically evaluated or eventually be replaced with more reliable alternatives.

This article proposes Bayesian Statistics as favourable alternative to the common Classical or Frequentist Statistics, when assessing the effectiveness of serious games, or even any instructional approach. It introduces and presents Bayesian Statistics and contrasts this with the Frequentist approach. Three original game studies are subjected to both Frequentist data analysis and Bayesian data analysis, and the outcomes are compared and discussed. The first study investigates the learning outcomes resulting from the “SKILLS” game, which is a board game for the training of basic military skills. The second study focuses on the motivational effects of a gamified digital workbook that offers spelling exercises to schoolchildren. The third study uses a statistics game for psychology students to investigate to what extent player personality and in-game player logs can predict the player's learning outcomes. These studies are exemplars of a between-subjects design, a pre-test/post-test within-subjects design and a regression predictor model, respectively. The statistical calculations in these studies are carried out with JASP (https://jasp-stats.org/), which is a free, user-friendly, open source package that accommodates both Frequentist and Bayesian analyses.

Before presenting the three game studies we first introduce and explain the foundations of both statistical paradigms.

2 Background

2.1 Frequentist versus Bayesian Statistics

In many scientific studies statistical methods are being used to process data obtained from observations and measurements and derive real-world insights from these. A principal

controversy in statistical sciences is the one between Classical Statistics (or Frequentist Statistics) and Bayesian Statistics. Although both paradigms make inferences from data, their approach is more or less opposite. In Classical Statistics truth is considered a fixed concept, expressed as fixed models or hypotheses (H_0, H_1, etcetera), while observations are considered random, in fact, conceived as a sample out of many similar samples that could have been drawn from the same population distribution. In Bayesian Statistics, it is the other way round: the data are fixed (facts as observed), but the models or hypotheses are random (e.g. parameterised) [12]. Although both approaches aim to make inferences from data in empirical research and seek evidence to support or reject proposed hypotheses, classical Statistics is about establishing truth and untruth, while Bayesian statistical inference is about belief revision, that is, adjusting one's initial belief about the world to the evidence provided by the data by making probability statements about possible states of the truth. In Table 1 both principles are expressed in a Bayesian way as conditional probabilities [13].

Table 1. *Principles of statistical paradigms.*

Paradigm	Outcome	Explanation
Frequentist	$p(D\|H)$	Probability of the data (D), when the (null-)hypothesis (H) holds
Bayesian	$p(H\|D)$	Probability of the hypothesis (H), given the available data (D)

These principles are further explained in the next sections.

2.2 Frequentist Statistics

In Frequentist Statistics, also known as Null-Hypothesis Significance Testing (NHST), the fit between the data and the hypothesis is checked, by calculating the probability of the data, given the hypothesis is true: $p(D|H)$. In the original Fisherian approach to NHST only a single hypothesis is considered, the null-hypothesis (H_0). In the Neyman-Pearsson tradition one would compare two pre-fixed hypotheses (the null hypothesis H_0 and the alternative hypothesis H_1), while assuming that the null-hypothesis is true. The calculated p-value is the main criterion for drawing conclusions. Today, a hybrid of both NHST approaches is commonly used [14].

For decades, NHST has been the dominant statistical paradigm in social sciences. This dominance is preserved and continually re-established by the fact that most statistical textbooks fully focus on the classical approach, thus acquainting masses of students and thereby future researchers and lecturers in this tradition [13,14]. The books sometimes briefly mention the basic Bayesian formula of conditional probability, but the resulting Bayesian Statistics approach and its consequences are rarely explained. Also, the fundamental issues of NHST are often neglected. NHST displays three fundamental problems:

- No hypothesis testing
  As can be seen from Table 1, NHST assumes that the null-hypothesis is true. Inferences can be made about the probability of the data only. In principle, NHST cannot tell us anything about the probability of the hypothesis, as the hypothesis is treated as a given fact. If the Null-hypothesis is not true, no inference can be made. In practice, however, obtaining a small p-value (e.g. $p<0.05$) is used for a *reductio ad absurdum* to conclude and accept the alternative hypothesis as being true, while larger p-values ($p>0.05$) are supposed to confirm the null-hypothesis. This is valid only in case H_0 and H_1 are fully complementary: if H_0 does not hold, H_1 must be valid. Even so, the evidence against the null hypothesis is often misinterpreted as evidence in favour of an alternative hypothesis rather than other options such as measurement error or selection bias [15].

- Structural, persistent misconceptions
  Treating *p*-values as evidence for hypotheses being true is one of many wide-spread misconceptions in NHST. Over the years, various studies have revealed shocking results about the persistent lack of statistical knowledge and understanding of psychology students, lecturers and professors, even of faculty teaching statistics [1,13,14,16]. It is often unjustly assumed that the 95% confidence interval of a parameter has 95% chance to contain the parameter [13]. Many sustain the illusion of certainty (believing that statistical significance proves that an effect exists), mistake *p*-values for error states, or suppose that the *p*-value reflects the probability of a successful replication (i.e., 1–*p*) [14,15]. The latter misconception makes replication studies appear to be superfluous. Various authors [21] denounce the so-called Null-ritual, which is the mindless application of NHST, being used as a conflicting mix of the Fisher approach and the Neyman-Pearson approach, presented in textbooks as an objective method magically making statements about truth governed by the sacred number (the *p*-value). Their criticism also concerns the elimination of the researcher's judgment precisely at points where statistical theories demand it, for instance about the grain size of the analysis, the grouping or exclusion of data and the tuning of coefficients, which all may bias toward finding positive results (negative results are not likely to be accepted for publication) or at least producing Type I and Type II statistical errors.
- The significance level debate
  Likewise, the central role of the *p*-value as the criterion for significance (usually $p<0.05$) has been widely criticised. Not just because the meaning and role of the *p*-value are often misunderstood, but also because the threshold value of 0.05 is an arbitrary convention, not in agreement with the required reliability of research [15,17-19]. A comparison of NHST with powerful Bayesian hypothesis testing shows that for a valid interpretation of data the levels of significance should be lowered down to the 0.001 level, or even below [18]. It would explain why so many social sciences studies cannot be reproduced appropriately: statistical significance seems to be obtained just too easily under the current NHST *p*-value regime. Acceptance of this huge correction of the significance threshold, however, would disqualify the majority of social science research as to produce nothing but noise and thus would – worst case - effect a total breakdown of the domain.

2.3 Bayesian Statistics

Most of the drawbacks of NHST are absent in Bayesian Statistics. It considers the data as fixed facts, which is what they are. It uses these data to update prior beliefs about the world, in fact, to infer statements about the probability of hypotheses, which is essentially what most research studies are about. The focus on probabilities of hypotheses implies that the object of analysis is stochastic. This allows to make probability statements on truth, cursed with uncertainty, and thereby it stays away from what is called the "false idol of objectivity" associated with the absolute conclusions and absolute truth in NHST [20]. In contrast to Frequentist 95% confidence intervals, Bayesian 95% confidence intervals for a parameter are what they are supposed to be: the chance that the parameter lies in the interval is 95%. To avoid any confusion with the Frequentist confidence intervals, the Bayesian intervals are called 95% credible intervals. No disputed p-values are needed to draw conclusions, since Bayesian decision criteria are straightforward and more robust than those of NHST. Finally, sample size need not be pre-defined to draw valid conclusions, that is, the "Stopping Rule Principle", which is required in NHST (and prone to violation), is irrelevant in Bayesian Statistics: data sampling may be stopped any time, without affecting the validity of analysis. A comprehensive overview can be found in [22].

In Bayesian Statistics our beliefs about the world are updated by the evidence from collected data. The approach starts off with a set of candidate hypotheses H_i about the world. Beforehand, we may or may not have some beliefs about which hypotheses are most

plausible. This belief system, based on prior knowledge and facts available, is then updated by the data collected. If the data are consistent with a hypothesis, our belief in that hypothesis is strengthened; if the data are inconsistent with the hypothesis, our belief in that hypothesis is weakened.

The process of Bayesian hypothesis testing is entirely based on Bayes' rule:

$$p(H|D) = \frac{p(D|H).p(H)}{p(D)} \quad (1)$$

Here,
$p(D)$ is the probability of data D
$p(H)$ is the prior probability of hypothesis H, not yet taking into account the data D
$p(D|H)$ is the likelihood of the data D, given hypothesis H
$p(H|D)$ is the posterior probability of hypothesis H, taking into account the data D

Bayes' rule, being grounded in formal probability theory, is undisputed and can be easily derived from $p(A,B)=p(A|B).P(A)=p(B,A)=p(B|A).p(B)$.

Bayes' rule can be understood as follows: the prior probability $p(H)$, which reflects our initial belief that hypothesis H is true, is updated by multiplication with $p(D|H)/p(D)$ to obtain the posterior probability $p(H|D)$, which is our revised belief state based on both our initial belief and the evidence from the data D. Comparison of two hypotheses H_1 and H_2 can be done by using equation(1) to calculate the posterior probabilities for each hypothesis separately. The ratio of posterior probabilities can then be written as:

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \cdot \frac{p(H_1)}{p(H_2)} \quad (2)$$

This equation shows that the change from prior odds $p(H_1)/p(H_2)$ to posterior odds $p(H_1|D)/p(H_2|D)$ is given by the ratio of the likelihoods $p(D|H_1)/p(D/H_2)$. The latter ratio is called the Bayes-factor: it quantifies the evidence that comes from the data in favour of H_1 against H_2. A high Bayes factor indicates how much more strongly the data support hypothesis H_1 over hypothesis H_2. The following classification scheme explains the interpretation of the Bayes Factor (Table 2).

Table 2. *Classification scheme for the interpretation of Bayes factors BF_{12} [23], adjusted from [12].*

Bayes factor BF_{12} for H_1 over H_2	**Evidence category**
> 100	Extreme evidence for H_1 over H_2
30 - 100	Very strong evidence for H_1 over H_2
10 - 30	Strong evidence for H_1 over H_2
3 - 10	Moderate evidence for H_1 over H_2
1 - 3	Anecdotal evidence for H_1 over H_2
1	No evidence over H_2

In practice, deriving the Bayes factor was often a difficult task, because posterior distributions can seldom be obtained in closed form. Only few combinations of prior distribution and likelihood function yield closed form posteriors, which directly allow to derive the Bayes factor. But as hypotheses in Bayesian inference are mostly treated stochastically, including a parameterised hypothesis model, one readily ends up with complex integrals in equation(1), which cannot be solved analytically. As a result, Bayesian inference in research remained the exception. Over the last few decades, however, computational sampling algorithms have successfully been put in place to remove these barriers and enable researchers to fully enjoy the benefits of Bayesian inference.

2.4 Sampling software for Bayesian Inference

The practical application of Bayesian Statistics has been greatly simplified after the introduction of the Markov Chain Monte Carlo method (MCMC) [3]). MCMC allows to generate posterior distributions via a step-wise sampling procedure applied to the right hand side of equation(1), thus removing the need for closed form analytic solutions. In 1989, BUGS (Bayesian inference Using Gibbs Sampling) was the first software programme offering MCMC for Bayesian analysis, later on followed by a free version (WinBUGS) and quite recently an open source version (OpenBUGS). JAGS (Just Another Gibbs Sampler) is an open source MCMC tool that runs natively on Windows, Mac, Linux and several Unix versions. In recent years, mainstream commercial statistics programmes (e.g. SPSS, Stata) have also included MCMC modules to support Bayesian Inference, while benefitting from their well-established user-interface styles. Accessibility to MCMC has been further extended by the recently launched JASP software, which is a free, open source tool that uses the R-package for Bayesian Adaptive Sampling and MCMC (https://jasp-stats.org/). JASP offers both Frequentist Statistics and Bayesian Statistics procedures in a highly user-friendly environment. In the following we have used JASP to perform both Frequentist and Bayesian analyses and compare the two.

2.5 Barriers to applying Bayesian Statistics

These recent developments have certainly contributed to an increased interest of methodologists and researchers in Bayesian approaches as an attractive replacement of Frequentist Statistics. But the perspective on Bayesian Statistics presented above may be easily perceived as too optimistic. Severe doubts have been raised about the subjective elements assumed in Bayesian Statistics, not just because of the concept of subjective belief states, but largely because of the need to specify a prior distribution, which is readily regarded a matter of personal taste [24, 25]. Also, the sampling procedures in MCMC can become computationally inefficient, the convergence of which is hard to control and to understand, in particular when datasets are complex. Default prior distributions are often used to simplify the procedure, but defaults may not quite adequately take into account available knowledge. Simonsohn [26] explains that default priors are prejudiced against small effects, suggesting misleading Bayesian results in particular under the following combination of factors: a small sample size, a small effect size and a prior distribution assuming a large effect size. Due to the ease of default distributions, users may readily regard the Bayesian Sampler as a magical machine that can be thoughtlessly operated to produce outputs from any inputs. In the case of large sample sizes Bayesian hypothesis testing suffers from the fact that even small and practically meaningless effects will be deemed “strongly supported by the data” [22], but this likewise holds for the observation of “significance” in Frequentist Statistics. It is important to note that the Bayes factor reflects a relative measure of performance rather than an absolute measure: this means that overwhelming support in favour of H_1 over H_0 only indicates that the predictive performance of H_1 is superior to that of H_0, even when the absolute performance of H_1 may be insignificant. Both Bayesian Statistics and Frequentist Statistics equally rely on the data

available and they are equally sensitive to biasing effects of selective reporting, ad-hoc use of transformations and outlier removal, which lead to incorrect conclusions.
Apart from these technical issues, the main barrier to a wide uptake of Bayesian Statistics in serious games research (and beyond) is the strong and well-established position of Frequentist Statistics as the prevailing paradigm in today's textbooks and lectures and accordingly in the minds of scholars and students. It is a well-recognised phenomenon that the social dynamics of scientific communities display an inherent resistance against paradigm shifts [27]. Nevertheless, in various domains (e.g. data science, machine learning, natural language processing) Bayesian approaches are gaining momentum and new libraries and tools for Bayesian processing are becoming available.

3 Experiments

3.1 Study 1: Basic military SKILLS game (Independent t-test)

This study uses data collected from an experiment with the "SKILLS" game, which is a game used for the mandatory yearly update of basic military knowledge and skills of military personnel [28]. Topics include basic search, survival, explosive devices, and ammunition awareness. The game was developed to replace classroom lectures. A quasi-experimental between-subjects experiment was setup with 102 participants who were randomly distributed over an experimental group playing the game (46 subjects) and a control group attending the lectures (56 subjects). Group composition did not differ with respect to gender, age and years of service. The sessions took typically 2-4 hours. After the sessions a 20 minutes post-test (10 multiple choice questions and 15 open questions) was administered to assess the learning outcomes. Total performance was expressed as an aggregate metric in the [0,1] interval. For analysing the anonymised data we specify two complementary hypotheses:

- H_0 the null hypothesis: post-test performances do not differ between the two conditions.
- H_1 the alternative hypothesis: post-test performances are different for the two conditions.

The data successfully passed the Shapiro-Wilk normality test (experimental group $W(46)=0.963$, $p=0.154$; control group $W(56)=0.968$, $p=0.141$), and Levene's test for homogeneity of variances ($F(1,100)=3.277$, $p=0.073$). The average post-test score of the group playing the game was 0.644 (standard deviation $SD=0.163$); for the control group attending the lectures the average post-test score was 0.580 ($SD=0.132$). Now, the question to be answered is what can be concluded about the hypotheses?

3.1.1 Frequentist analysis

The Frequentist approach starts with the assumption that the null-hypothesis H_0 is true. A two-sided independent t-test was used to analyse the differences between the two groups. The t-statistic was found to be $t(100)=2.186$, $p=0.031$, which indicates that the difference between groups is significant (because $p<0.05$). The effect size given by Cohen's d is 0.435, which means a small to medium size effect. The mean difference between groups ($\delta=0.064$) is linked to a 95% confidence interval ranging from 0.006 to 0.122.

The significant t-test technically means, that if H_0 is true (viz. no differences between groups), the probability of obtaining a result as extreme as the data observed is only 3.1%. Although this is generally accepted as a positive outcome in favour H_1, the t-test does not quantify the evidence in favour of each hypothesis, which hampers a better substantiated conclusion about the two hypotheses. The Frequentist 95% confidence interval is often mistaken for the chance that the value of the mean difference has a 95% chance to be in that interval. However, this is a misconception, because long-run probabilities cannot be

assigned to individual events without specifying what the long run practically entails (the reference class problem, see e.g. [29,30]). Instead, the confidence intervals represent the following: repeatedly drawing samples from the probability distribution (which is the premise of Frequentist Statistics) would each time produce a different confidence interval, while only 95% of those intervals would contain the population mean [31].

3.1.2 Bayesian analysis

Next, a Bayesian independent *t*-test was applied to study the difference between groups. The approach entails the assessment of the effect size δ, which is the standardised mean difference between the post-test scores of the two groups. A principal difference with the Frequentist approach is that Bayesian Statistics does not assume that H_0 is correct. Instead, it involves the comparison of two competing hypotheses for effect size δ, where H_0 in this case happens to refer to zero effect size (δ=0) and H_1 refers to non-zero effect size ($\delta \neq 0$). The first thing to do is to decide upon the prior distribution of δ, that is, the probability distribution of the effect size before taking into account the data of the experiment. Here, we may not have a clear clue whether δ would be positive or negative, so the prior distribution should be symmetrical. Also, it is fair to say that very high positive or negative effect sizes (e.g. $|\delta|>2$) are not quite likely. Hence, the prior distribution would somehow be bell-shaped and centred around δ =0. To this end, JASP proposes as a default a Cauchy distribution (also known as Lorentz distribution), which is very similar to the normal distribution, but has thicker tails. A default value of the Cauchy width parameter *r* of $\frac{1}{2}\sqrt{2}$ (0.707) reflects the reasonable implication that δ is between -2 and 2 (probability of 78%). The choice of the prior may seem a bit arbitrary, but the influence of the prior on the Bayes factor can be tested afterwards with a separate robustness check.

Running the Bayesian independent *t*-test in JASP shows that the Bayes factor is 1.714 in favour of H_1 over H_0. This qualifies as anecdotal evidence of H_1 over H_0, which means "weak" (cf. Table 2), "inconclusive" [23,30] or even "worth no more than a bare mention" [12]. Figure 1 shows how the posterior distribution of δ differs from the prior.

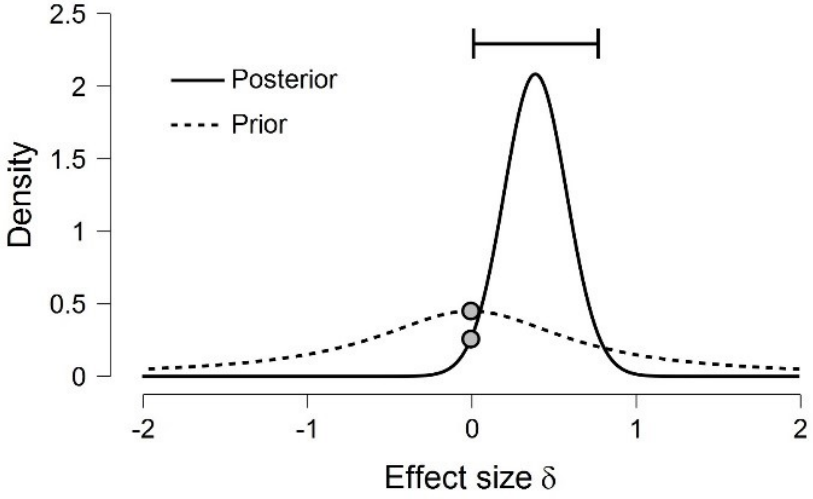

Figure 1. *Prior and posterior distribution of standardised effect size δ.*

The dashed curve is the prior distribution as specified by the Cauchy distribution. The solid curve is the calculated posterior distribution. Most of the posterior mass is on the positive side. The posterior peaks at δ=.064, which is the mean difference between groups. The Bayesian 95% credible interval, indicated as the solid interval above the posterior curve, ranges from 0.002 to 0.078: in the Bayesian framework this truly means that the mean effect size δ has a 95% chance to be in that interval. The two grey dots in figure 1 mark the prior and posterior densities at δ=0, that is, 0.5 and 0.3 respectively. It shows that the data have decreased the support for δ=0 (which is H_0) with a factor of about 0.6 (viz.

0.3/0.5). This means that the support for H_1 over H_0 is 1/0.6, which, obviously, corresponds with the calculated Bayes factor of 1.714.

3.1.3 Comparison of statistical methodologies

The clearly significant result from the Frequentist analysis (p=0.031) cannot be fully confirmed by the Bayesian analysis ("worth no more than a bare mention"). Put differently, regarding ratio of the Bayesian probabilities of H_0 and H_1 (the Bayes factor) the positive Frequentist result confirmed by the low p-value turns out to be less flourishing, if not deceptive. The analysis confirms that Bayesian criteria are more strict than p-values to reduce Type 1 errors (falsely rejecting the null-hypothesis), which may be the case here.

3.2 Study 2: KPITO game-based spelling workbook (mixed factorial ANOVA)

This case uses data collected in an experimental study on motivational effects of "KPITO", which is a game-based, digital workbook targeting spelling skills for school children [32]. The objective of KPITO is to make repetitive spelling exercises more attractive in order to raise the children's motivations. The content and structure of KPITO are fully based on a well-established original paper-based workbook, which is not gamified. Game elements in the digital version include challenges, direct feedback, adaptivity, animated objects, retrials, rewards (coins, permissions) and storytelling. A quasi-experimental pre-test/post-test study was arranged with 94 schoolchildren to compare children's motivation in the KPITO condition and the paper-based workbook condition. The experimental group (KPITO) was populated with 46 children; the control group (paper-based) with 48 children. Both groups performed the spelling exercises 1 hour per week, during a full period of 6 weeks.

Pre-test and post-test questionnaires for motivation used 20 items of the Intrinsic Motivation Inventory [33], representing the interest/enjoyment scale, the perceived competence scale, and perceived choice scale. The ordinal data from the 5-point Likert scales were combined into an overall motivation score. The study uses two independent variables: the within-subjects factor *Time* and the between subjects factor *Group*, respectively; the dependent variable is *Motivation*. The statistical analysis requires a mixed factorial ANOVA. Three null hypotheses need to be specified:

- $H_{0,A}$: there are no within-subjects differences between pre-test and post-test scores (ignoring *Group*).
- $H_{0,B}$: there are no between-subject differences between test scores of the two groups (ignoring *Time*).
- $H_{0,C}$: there are no interaction effects between *Time* and *Group*.

The associated alternative hypotheses $H_{1,A}$, $H_{1,B}$ and $H_{1,C}$ are the negations of the respective null-hypotheses. Descriptive statistics are in Table 3.

Table 3. *Mean motivation scores from the KPITO study*

Time	***Group***	***Motivation***	***SD***	***N***
Pre-test	Experiment	3.4	0.8	46
	Control	3.5	0.8	48
	Overall	3.4	0.8	94
Post-test	Experiment	3.8	0.6	46
	Control	3.4	1.0	48

	Overall	3.6	0.8	94

Face-value, it seems that motivation goes up in the experimental group, whereas it remains at a low level in the control group. The question to be answered is what can be concluded from the data about the hypotheses.

3.2.1 Frequentist analysis

For testing the respective hypotheses, the Frequentist mixed factorial ANOVA produces the following results. With respect to $H_{0,A}$, which assumes that there are no within-subjects differences between pre-test and post-test scores (ignoring *Group*), we find $F(1, 92)=5.190$, $p=0.025$. This suggests a significant effect in favour of the alternative hypothesis $H_{1,A}$: participants show increased motivation after the lessons. The overall effect size is given by partial eta squared $\eta_p^2=0.053$, which means it is a medium size effect. As can be seen from Table 3 above, this increase can be fully attributed to the experimental condition.

With respect to $H_{0,B}$, which assumes that there are no motivation differences between the two groups (ignoring *Time*), we find $F(1,92)=0.952$, $p=0.332$. This means corroboration for this null hypothesis: overall motivations in the groups do not differ significantly.

For $H_{0,C}$, which assumes that there is no interaction effect between the factors *Time* and *Group*, we find $F(1,92)=19.691$, $p<0.001$. This is strong evidence in favour of the KPITO case: after the experiment, children in the game condition display significantly higher motivations, both in absolute terms and in terms of motivation growth. The effect size metric $\eta_p^2=0.176$, which indicates a large effect.

3.2.2 Bayesian analysis

In all comparative studies, which aim at discovering differences of an observed variable between groups, prior distributions will inevitably be bell shaped (for instance, the Cauchy distribution). For the Bayesian Repeated measures ANOVA (allowing for mixed factors) we preserve the default Cauchy priors in JASP, which use a width factor of 0.5 for fixed effects, 1.0 for random effects and ¼√2 (0.354) for interaction effects. The output generated by JASP is summarised in Table 4.

Table 4. *Model comparison for the Bayesian mixed factors ANOVA applied to the KPITO case.*

Models (*M*)	***P(M)***	***P(M|data)***	***BF $_M$***	***BF $_{10}$***	**error %**
Null model	0.200	0.002	0.012	1.000	0
Time + Group + Time ✻ Group	0.200	0.991	450	321	1.6
Time	0.200	0.003	0.012	0.9	1.9
Group	0.200	0.001	0.006	0.4	0.5
Time + Group	0.200	0.001	0.006	0.4	2.8

The output in Table 4 shows a comparison between 5 models (*M*). *P(M)* denotes the prior model probability for each of the five candidate models, each equally loaded with the same probability. The *P(M|D)* column shows the posterior model probabilities (cf. the left-

hand side of Equation(1)). The BF_M column displays the change from prior model odds to posterior model odds. The Bayes factor BF_{10} indicates how much more strongly the data support each model over the null model (viz. the null hypothesis). A large Bayes factor is found only for the model that includes both two main factors (*Time* and *Group*) and the interaction term (*Time*Group*). The Bayes factor of 321 provides extreme evidence in favour of this model (cf. qualifications in Table 2).

3.2.3 Comparison of statistical methodologies

Both methodologies reveal strong evidence in favour of the model that includes the interaction between both factors *Time* and *Group*. The Frequentist model produces a highly significant result ($p<0.001$). The Bayesian model output is more informative showing a convincing Bayes factor of 321 and a small error (1.6%). It unambiguously specifies how strong the evidence is in favour of the proposed model as compared to the null hypothesis as well as compared to the other hypotheses.

3.3 Study 3: Playground game (multiple regression)

This case uses data collected in a study of the Playground game, which – coincidently - deals with misconceptions in statistics, targeting psychology students [34]. In the game the player investigates potentially unjust statistical claims from "alleged experts" in a practical case about determining the best location for laying out a playground. The game is composed of a set of challenges, each requiring well-considered decision taking. Correctness of decisions is expressed in a performance rate. The number of participants was 112. A pre-questionnaire was used to test general prior knowledge in statistics. It included 15 self-assessment items on statistics and a set of five test questions, which were combined into an aggregated metric on a 1-10 scale. Also, the pre-questionnaire used 5x3 items from the Big Five Inventory [35] to assess the player's personality traits. A post-questionnaire included a post-test very similar to the prior knowledge test to assess the participants' post-game knowledge. Full sessions of the experiment took typically 1.5 hours. The average time for completing the Playground game was 65 minutes with a large spread (*SD*=54 minutes). Anonymised log files of students were available for extracting some key indicators, such as success rate and time spent.

The main question in this study is to what extent player data can provide a predictor model for post-test performance. Candidate model variables are the student's *Success rate*, *Playing time* and *Pre-test score*. Also, the *Conscientiousness score* obtained from the personality test is expected to be a relevant factor, as it readily relates to dedication, precision and the aim to avoid mistakes. Descriptive statistics of these variables are presented in Table 5.

Table 5. *Descriptive statistics of the Playground experiment.*

Variable	***N***	***Mean***	***SD***	***SE***
Post-test score	112	6.16	1.0	0.10
Playing time (seconds)	112	3897	3252	307
Success rate	112	0.64	0.05	0.005
Conscientiousness score	112	6.99	1.38	0.13
Pre-test score	112	5.53	0.93	0.09

3.3.1 Frequentist analysis

First, a paired *t*-test of pre-test and post-test scores revealed that the observed differences (cf. Table 5) are significant ($t(111)=7.479$, $p<0.001$). Next, a stepwise multiple regression analysis was carried out using *Playing time, Success rate, Pre-test score* and *Conscientiousness score* as predictors for the *Post-test score*. In the stepwise approach the predictor showing the highest correlation with the outcome variable is entered first into the

model. Additional predictors are then added one by one based on their correlations, while at every step any redundant predictors are traced and removed. The basic assumptions underlying linear regression have been successfully confirmed. These include homogeneity of variance (by visual inspection of a residuals-versus-predicted plot), normality and linearity (by visual inspection of the Q-Q residuals plot), and minimal multi-collinearity (checked with the *Variance Inflation Factor VIF*<10 and *Tolerance*>0.1). The JASP regression output presents 3 proposed models (Table 6). Model 1 only uses the *Pre-test score* as predictor, model also uses *Playing time*, while model 3 uses both *Pre-test score, Playing time* and *Conscientiousness score*.

Table 6. *Comparison of candidate regression models from the Frequentist analysis.*

Model	Factors	Sum of Squares	*df*	*F*	*p*
1	*Pre-test score*	39.021	1	59.252	< .001
	Residual	72.443	110		
	Total	111.464	111		
2	*Pre-test score* *Playing time (s)*	42.747	2	33.903	< .001
	Residual	68.717	109		
	Total	111.464	111		
3	*Pre-test score* *Playing time (s)* *Conscientiousness score*	45.232	3	24.585	< .001
	Residual	66.233	108		
	Total	111.464	111		

This table shows that all three models are highly significant (p<0.001), that is, the models are a better predictor than the null-models, which use the mean values of the predictors. Note that *Success rate* although declared as an input is not preserved in the model, due to an apparent lack of predictive power. The highest Sum of Squares is in model 3: this model accounts for 45/111*100%= 40.6% of the variance in post-test scores. Table 7 lists the (unstandardized) regression coefficients (intercept and slopes) for the 3 models.

Table 7. *Regression coefficients of three models (M) obtained from the Frequentist analysis. Also, the Standard Errors (SE), t statistics, p-values and the bounds of the 95% confidence interval (CI) are given.*

						95% Confidence Interval	
M	**Factors**	**Coefficient**	**SE**	***t***	***p***	**Lower**	**Upper**
1	*Intercept*	2.633	0.464	5.674	< .001	1.713	3.552
	Pre-test score	0.636	0.083	7.698	< .001	0.472	0.800
2	*Intercept*	2.913	0.468	6.219	< .001	1.985	3.841
	Pre-test score	0.625	0.081	7.720	< .001	0.465	0.786
	Playing time (s)	$-5.642*10^{-5}$	$2.321*10^{-5}$	-2.431	0.017	$-1.024*10^{-4}$	$-1.042*10^{-5}$
3	*Intercept*	2.203	0.581	3.790	< .001	1.051	3.355
	Pre-test score	0.616	0.080	7.691	< .001	0.457	0.774
	Playing time (s)	$-5.555*10^{-5}$	$2.289*10^{-5}$	-2.427	0.017	$-1.009*10^{-4}$	$-1.017*10^{-5}$
	Conscientiousness score	0.109	0.054	2.013	0.047	0.002	0.216

In sum, Model 3 produces a highly significant result ($F(1,101)= 24.585$, $p<0.001$) and a Frequentist regression equation given by

$$Post\text{-}test\ score = 2.203 + 0.616 \cdot Pre\text{-}test\ score - \\ +5.555 \cdot exp^{-5} \cdot Playing\ time + \\ +0.109 \cdot Conscientiousness\ score \quad (3)$$

3.3.2 Bayesian analysis

A Bayesian paired *t*-test produces a large Bayes factor of $BF_{10}=4.5*10^{8}$, which indicates extreme evidence for post-test scores being larger than pre-test scores. The Bayesian regression analysis does not require a specific predictor entry mechanism, as it simply allows for a comparison between multiple models. In contrast with Frequentist Statistics, the model parameters, that is, the intercept and coefficients of the regression equation, are not estimated by a single value but are drawn from probability distributions and optimised step by step. In JASP, we used a multivariate Cauchy distribution and default prior scale (width ¼ V2) [36]. Table 8 shows the most relevant models.

Table 8. *Bayesian model comparison of the Playground case.*

Model	**Factors**	***P(M)***	***P(M|data)***	***BF $_M$***	***BF $_{10}$***	***R²***
1	*Null model*	0.200	$3.624*10^{-10}$	$1.449*10^{-9}$	1.000	0.000

2	*Pre-test score* *Playing time (s)* *Conscientiousness score* *Success rate*	0.200	0.317	1.859	$8.756*10^{+8}$	0.411
3	*Pre-test score* *Playing time (s)* *Conscientiousness score*	0.050	0.257	6.583	$2.840*10^{-9}$	0.406
4	*Pre-test score* *Playing time (s)*	0.033	0.145	4.925	$2.404*10^{-9}$	0.384
5	*Pre-test score*	0.050	0.093	1.953	$1.029*10^{-9}$	0.350
6	*Pre-test score* *Conscientiousness score*	0.033	0.062	1.921	$1.028*10^{-9}$	0.373

From the Bayes factors (BF_{10}) in Table 8 it can be concluded that there is extreme evidence against the null model for models that include the *Pre-test score*. Models that do not include the *Pre-test score* are excluded from the Table, as they turned out to display low Bayes factors indicating anecdotal evidence. Model 3, which is based on *Pre-test score, Playing time* and *Conscientiousness score*, has the largest Bayes factor ($2.840*10^{9}$). Although it has a slightly lower explanatory power than model 2 (R^2=.406 versus R^2=.411), model 3 is three times more likely than model 2. Apparently, the inclusion of the *Success rate* as a factor slightly inflates the model. Therefore, and for reasons of parsimony (Ockham's razor), we select Model 3 as the best option. The decision is supported by the Marginal Inclusion Probability of the *Access rate* variable being lower than 50% (not displayed in Table 8; in fact it is 44%). Table 9 shows the coefficients of the favoured Bayesian regression model as well as their credible intervals.

Table 9. *Bayesian regression coefficients of Model 3, including the standard errors (SE) and the bounds of the Credible Intervals.*

			95% Credible Interval	
Coefficient	**Mean**	**SE**	**Lower**	**Upper**
Intercept	2.384	0.074	2.237	2.531
Pre-test score	0.587	0.078	0.432	0.742
Playing time (s)	$-5.295*10^{-5}$	$2.235*10^{-5}$	$-9.724*10^{-5}$	$-8.661*10^{-6}$
Conscientiousness score	0.104	0.053	$-8.751*10^{-4}$	0.208

The regression equation of Model 3 can thus be written as:

$$Post\text{-}test\ score = 2.384 + 0.587 \cdot Pre\text{-}test\ score - \\ +5.295 \cdot exp^{-5} \cdot Playing\ time + \\ +0.104 \cdot Conscientiousness\ score \quad (4)$$

The result turns out to be insensitive to the prior's width: changing from narrow prior (width ¼V2) to a wide prior (width 1.0), produces similar results for all models, be it that the narrow prior provides larger Bayes Factors. As a side remark, it is noted that JASP uses the Bayesian Adaptive Sampling (BAS) Package from R, which uses centralised scales for all predictors in order to make sure that the resulting Bayes factors are location-scale invariant, and to disconnect interaction effects from main effects. As a consequence, a reversed transformation is needed to obtain the original intercept scale.

3.3.3 Comparison of statistical methodologies

Both statistical methodologies propose a regression model with *Pre-test score, Playing time* and *Conscientiousness score* as predictors. However, the intercepts and coefficients of the two models are slightly different. To decide which is the best model, is not a straightforward task. One may naively want to compare the mean squared errors of the predictions. The mean squared error of the Frequentist model is 0.66 (mean error 0.83). It is smaller than the mean squared error of the Bayesian model, which is 1.19 (mean error 1.09). Therefore it may be tempting to prefer the Frequentist solution. However, when we would use the mean squared error as the decisive criterion, the Frequentist regression will always win the game, exactly because the Frequentist fitting procedure is based on minimising the mean squared error (ordinary least squares). It would still be a mistake to conclude that given the lower error the Frequentist model provides the best fit to the data, because the mean squared error is no more than an arbitrary utility function or loss function. This arbitrariness was even recognised by Carl Friedrich Gauss, when he proposed the mean squared error as an accuracy metric. Because of the squares the mean squared error has the disadvantage of disproportionately weighting outliers. As can be read from Tables 7 and 9, respectively the standard errors of the predictor coefficients are slightly lower for the Bayesian case, for the intercept even substantially lower. Therefore, propagation of the predictor errors produces systematically lower standard errors in the predicted *Post-test scores*. When, we use the predictors' mean value coordinates provided by Table 5, we find a standard error of 0.83 for the Frequentist regression and a standard error of 0.58 for the Bayesian regression, This means that the Bayesian regression model is more accurate than the Frequentist regression model. The Bayesian credible intervals are very similar to the Frequentist confidence intervals, be it that the Frequentist intercept range is much wider. As explained before, however, the two interval types represent different things and thereby cannot be easily compared.

4 Discussion and conclusion

The comparison of Frequentist Statistics and Bayesian Statistics in three separate studies has revealed some interesting differences. In study 1 (the SKILLS board game) the Frequentist analysis yields a significant result in favour of the game group, showing a significantly higher score on the post-test than the lecture group (p=0.031). The Bayes factor of 1.714, however, indicates that the evidence is "weak" or even "worth no more than a bare mention". This difference between the two approaches confirms conclusions from existing research [18] that the Bayesian criteria are more strict than p-values and reduce Type 1 errors: the Frequentist result may be deceptive. In study 2 (the KPITO gamified work book) both the Frequentist and the Bayesian approach to a mixed factors ANOVA both show strong evidence in favour of the KPITO case: after the experiment, children in

the game condition display significantly higher motivations, both in absolute terms and in terms of motivation growth. In study 3, it was established that playing the Playground game leads to higher test scores. The Frequentist regression coefficients for predicting the post-test scores were found to be different and less accurate than the Bayesian regression model. Although the outcomes of two competing approaches should ideally be projected on a fixed reference standard in order to decide on the "winner", our three studies have demonstrated that Bayesian Statistics is more informative about the hypotheses under study as compared to Frequentist Statistics. In particular, Bayesian Statistics allows to directly compare different models or hypotheses given the data available and attaches relative probabilities to these, rather than enforcing a yes-or-no decision about a single hypothesis being true. From a theoretical perspective it is worthwhile to mention decisions theory's complete class theorem, which states that theoretically the Bayesian procedure performs at least as well as the non-Bayesian procedure in all cases with certainty [37]. Still, in some practical conditions Frequentist Statistics might be preferred, e.g. because of computational efficiency. With the emergence of statistical packages, such as JASP, that provide Bayesian approaches through simple user-interfaces, principal barriers to applying Bayesian Statistics have vanished. This opens up new opportunities to amplify the discourse about the Frequentist dominance and propagate the Bayesian case. It seems that given its conceptual superiority, its straightforward interpretation and its capacity to make probability statements about the truth, the Bayesian paradigm may gradually take the upper hand at the expense of Frequentist paradigm. Nonetheless, this will take quite some time considering current firm and established position of Frequentist Statistics both in textbooks, educational programmes and scientific articles. The importance of user-friendly packages such as JASP in this transition can hardly be underestimated, as they greatly simplify the application of Bayesian Statistics. In JASP as well as other packages, default functions suffice in most cases. Nevertheless, sufficient understanding of the principles and procedures in Bayesian Statistics are a precondition for its appropriate application. For further reading we recommend the excellent introduction to the Bayesian background from Wagenmakers and colleagues [22] and their practical example cases with JASP [38]. Increasingly, textbooks, tutorials and examples on Bayesian analysis are becoming available to be included into educational programmes. Also, a growing community of adopters can be witnessed along with a growing volume of articles exposing the Bayesian paradigm. It has been the aim of current article to contribute to this transition by developing and exposing representative Bayesian application cases and discussing the underlying concepts, mechanisms and arguments. Although confined to the distinct and well-delineated domain of serious games, the cases are relevant for the wider field of learning and teaching, and beyond.

5 References

[1] G. Gigerenzer, "Mindless statistics", *The Journal of Socio-Economics*, Vol. 33, No. 5, pp.587-606, 2004a, doi: 10.1016/j.socec.2004.09.033.

[2] L. Wilkinson, and The Task Force on Statistical Inference, "Statistical methods in psychology journals: Guidelines and explanations", *American Psychologist*, Vol. 54, pp. 594–604, 1999, doi: 10.1037/0003-066X.54.8.594.

[3] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice,* Boca Raton, FL, USA: Chapman & Hall/CRC, 1996, doi: 10.1201/b14835.

[4] D. G. Oblinger, and J. L. Oblinger, Eds., *Educating the net generation*, Washington DC., Educause, 2012. Accessed Nov. 10, 2020. [Online]. Available: http://www.educause.edu/educatingthenetgen/.

[5] M. Prensky, "Digital natives, digital immigrants", *On the Horizon*, Vol. 9, No. 5, pp. 1–6, 2001, doi: 10.1108/10748120110424843.

[6] J. Stewart, L. Bleumers, J. van Looy, I. Mariën, A. All, D. Schurmans, K. Willaert, F. De Grove, A. Jacobs. and G. Misuraca,) *The Potential of Digital Games for Empowerment and Social*

Inclusion of Groups at Risk of Social and Economic Exclusion: Evidence and Opportunity for Policy, C. Centeno, Ed., 2013, Brussels: Joint Research Centre, European Commission.

[7] C. B. Swanson, "U.S. graduation rate continues decline" *Education Week*, June 2, 2010 [online edition]. Accessed Nov. 10, 2020. [Online]. Available: http://www.educause.edu/educatingthenetgen/.

[8] C. N. Quinn, *Engaging learning. Designing E-Learning Simulation Games*, San Francisco, CA, USA: Pfeiffer, John Wiley and Sons Inc., 2005.

[9] S. A. Barab, M. K. Thomas, T. Dodge, B. Carteaux, and H. Tuzun, "Making learning fun: Quest Atlantis, a Game without Guns", *Educational Technology Research and Development*, Vol. 53, No. 10, pp. 86–107, 2005, doi: 10.1007/BF02504859.

[10] H.-T. Hung, "Gamifying the flipped classroom using game-based learning materials", *English Language Teaching Journal*, Vol. 72, No. 3, pp. 296–308, 2018, doi: 10.1093/elt/ccx055.

[11] W. Westera, "Games are motivating, aren´t they? Disputing the arguments for digital game-based learning", *International Journal of Serious Games*, Vol. 2, No. 2, 2015, [online edition], doi: 10.17083/ijsg.v2i2.58. Accessed Nov. 10, 2020. [Online]. Available: http://journal.seriousgamessociety.org/index.php/IJSG/article/view/58.

[12] H. Jeffreys, *Theory of probability,* 3rd *ed.*, Oxford: Oxford University Press, 1961.

[13] H. Haller, and S. Krauss, "Misinterpretations of Significance: A problem students share with their teachers?", *Methods of Psychological Research Online*, Vol. 7, No. 1, pp. 1–20, 2002. Accessed Nov. 10, 2020. [Online]. Available: https://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

[14] G. Gigerenzer, "Statistical Rituals: The Replication Delusion and How We Got There", *Advances in Methods and Practices in Psychological Science*, Vol. 1, No.2, pp. 198 –218, 2018, doi: 10.1177/2515245918771329.

[15] A. Gelman, "P-values and statistical practice", *Epidemiology*, Vol. 24, No. 1, pp. 69-72, 2013, doi: 10.1097/EDE.0b013e31827886f7.

[16] M. Oakes, *Statistical inference: A commentary for the social and behavioral sciences,* New York NY, USA: Wiley, 1986.

[17] M. J. Lew, "To P or not to P: On the evidential nature of P-values and their place in scientific inference", *arXiv:1311.0081 [stat.ME],* 2013. Accessed Nov. 10, 2020. [Online]. Available: http://arxiv.org/abs/1311.0081.

[18] V. E. Johnson, "Revised standards for statistical evidence", in *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 110, No. 48, pp. 19313–19317, 2013, doi: 10.1073/pnas.1313476110

[19] J. Cohen, "The earth is round (p <.05)", *American Psychologist*, Vol. 49, No. 12, pp. 997-1003, 1994, doi: 10.1037/0003-066X.49.12.997.

[20] E. E. Leamer, *Specification Searches, Ad-hoc Inference with Nonexperimental Data,* New York, NY, USA: Wiley, 1978.

[21] G. Gigerenzer, S. Krauss, and O. Vitouch, "The Null Ritual. What You Always Wanted to Know About Significance Testing but Were Afraid to Ask", in *The Sage handbook of quantitative methodology for the social sciences*, D. Kaplan, Ed. Thousand Oaks, CA, USA: Sage Publications, 2004b, pp. 391–408.

[22] E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q. F. Gronau, M. Šmíra, S. Epskamp, D. Matzke, J. N. Rouder, and R. D. Morey, "Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications", *Psychonomic Bulletin & Review*, Vol. 25, No. 1, pp. 35-57, 2018, doi: 10.3758/s13423-017-1343-3.

[23] M. D. Lee, and E.-J. Wagenmakers, *Bayesian cognitive modeling: A practical course,* Cambridge MA, USA: Cambridge University Press, 2013.

[24] J. M. Bernardo, and A. F. M. Smith, *Bayesian Theory*. Chichester: Wiley, 1994, doi: 10.1002/9780470316870.

[25] A. Gelman, Objections to Bayesian statistics, *Bayesian Analysis*, Vol. 3, No. 3, pp. 445–450, 2008, doi: 10.1214/08-BA318.

[26] U. Simonsohn, "The Default Bayesian Test is Prejudiced Against Small Effects", Datacollada.org, http://datacolada.org/35 (accessed December 26, 2020).

[27] T. Kuhn, *The Structure of Scientific Revolutions (1st ed.)*, Chicago, USA: The University of Chicago Press, 1962.

[28] J. Van Hofwegen-'t Lam, *Serieus Spel bij Defensie: Een Onderzoek naar de Invloed van Competitie op de Motivatie en het Leerresultaat bij het Jaarlijks Oefenprogramma Militaire Basisvaardigheden.* Master's thesis Welten Institute, Heerlen: Open Universiteit, 2018.

Accessed Nov. 10, 2020. [Online]. Available: https://research.ou.nl/en/studentTheses/serieus-spel-bij-defensie-een-onderzoek-naar-de-invloed-van-compe.
[29] H. Reichenbach, *The theory of probability,* Berkely, USA: University of California Press, 1949.
[30] R. D. Morey, R. Hoekstra, J. N. Rouder, and E.-J. Wagenmakers, "Continued misinterpretation of confidence intervals: Response to Miller and Ulrich", *Psychonomic Bulletin and Review*, Vol. 23, No. 1, pp.131-140, 2016, doi: 10.3758/s13423-015-0955-8
[31] O. J. Berger, and R. L. Wolpert, *The likelihood principle,* 2nd *ed*. Hayward, CA, USA: Institute of Mathematical Statistics, 1988.
[32] G. Van Egmond-van Beelen, *Spel en Spelling. Effecten van een Gegamificeerd Digitaal Werkboek op Motivatie en Spellingvaardigheid van Basisschoolleerlingen,* Master's thesis Welten Institute, Heerlen: Open Universiteit, 2018. Accessed Nov. 10, 2020. [Online]. Available: https://research.ou.nl/en/studentTheses/spel-en-spelling.
[33] R. M. Ryan, and E. L. Deci, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being", *American Psychologist*, Vol. 55, No. 1, pp. 68-78, 2000, doi: 10.1037/0003-066X.55.1.68.
[34] W. Westera, A. Slootmaker, and H. Kurvers, "The Playground Game: Inquiry-Based Learning About Research Methods and Statistics", in *Proceedings of the 8th European Conference on Games Based Learning* C. Busch, Ed. Sonning Common, UK: ACPIO, 2014, pp. 620-627.
[35] P. John, and S. Srivastava, "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives", in *Handbook of personality: Theory and research,* Vol. 2, L. A. Pervin and O. P. John, Eds. New York, NY, USA: Guilford Press, pp. 102–138, 1999.
[36] J. N. Rouder, R. D. Morey, P. L. Speckman, and J. M. Province, "Default Bayes factors for ANOVA designs", *Journal of Mathematical Psychology*, Vol. 56, pp. 356-374, 2012, doi: 10.1016/j.jmp.2012.08.001
[37] J. Steinhardt, Beyond Bayesians and Frequentists, Standord University, 2012. Accessed December 26, 2020. [Online]. Available: https://cs.stanford.edu/~jsteinhardt/stats-essay.pdf
[38] E.-J. Wagenmakers, J. Love, M. Marsman, T. Jamil, A. Ly, J. Verhagen, R. Selker, Q. F. Gronau, D. Dropmann, B. Boutin, F. Meerhoff, P. Knight, A. Raj, E.-J. van Kesteren, J. van Doorn, M. Smíra, S. Epskamp, A. Etz, D. Matzke, T. de Jong, D. van den Bergh, A. Sarafoglou, H. Steingroever, K. Derks, J. N. Rouder, and R. D. Morey, Bayesian inference for psychology. Part II: Example applications with JASP, *Psychonomic Bulletin & Review*, Vol. 1, No. 25, pp. 58–76, 2018, doi: 10.3758/s13423-017-1323-7.