Why the Wilhelm Scream, One of the Most Well-Known "Easter Egg" Sound Effects in Video Games, is Used, and How to Test Participants' Extent of Assessment by Using a Bayesian Statistical Approach

Jakub Binter[1,2(✉)], Silvia Boschetti[1], Tomáš Hladký[1], Daniel Říha[1], and Hermann Prossinger[2,3]

[1] Faculty of Humanities, Charles University, Prague, Czech Republic
jakub.binter@fhs.cuni.cz
[2] Faculty of Social and Economic Studies,
University of Jan Evangelista Purkyně, Ústí nad Labem, Czech Republic
[3] Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria

Abstract. The *Wilhelm Scream* is the sound effect that first appeared in 1951 and has been used in an almost uncountable number of movies and video games as an "Easter egg" related to pain, injury, and displeasure. The *Intensity Paradox* theory claims that vocalizations are rated valence-wise with very low accuracy. The learning hypothesis, on the other hand, suggests that repeated exposure allows for overcoming the intensity paradox. To test this, we collected a large sample of two ratings each by 902 raters and developed a novel statistical approach, based on Bayesian statistics, in which we determine the maximum-likelihood probability of Beta distributions of these responses. We asked our participants to rate the Wilhelm Scream mixed in with the other high-intensity vocalizations. The outcome showed an unexpected result—an extremely high precision of rating and consistency in repeated exposure. Furthermore, older men (but not women) rated with lower consistency. This provides, using a natural experiment, novel support for the learning effect of vocalization and the importance of using Easter eggs as a vehicle of familiarity related to gaming attractiveness and audiovisual media creation.

Keywords: Wilhelm Scream · Beta distribution · Vocalization · Game design · Intense affective states

1 Introduction

1.1 A Subsection Sample

The Wilhelm Scream (one of the "Easter eggs" of gaming sound engineers) is the sound effect that was first used in 1951 and it is often included in videogames—among these: *Battlefield 1*, *Call of Duty*, *Modern Warfare 2*, *Fallout 3*, *Grand Theft Auto V*, and *Red*

C. Stephanidis et al. (Eds.): HCII 2023, CCIS 1957, pp. 118–123, 2024.
https://doi.org/10.1007/978-3-031-49212-9_16

Dead Redemption. As the name suggests, it is a loud vocal display from a male actor and is related to a negative emotion of pain or displeasure involving fear.

The term Easter egg was originally used for the hidden function of a program that would be impossible to be called the usual way by the user. Since then, the meaning has shifted and the meaning has become broader.

In order to make audio-visual materials attractive, immersive, and enjoyable, a suite of options needs to be employed. One example of these are so-called Easter eggs, "[the] secret 'goodies'… Used in video games, movies, TV commercials, DVDs, CDs, CD-ROMs and every so often in hardware" (PC Magazine [n.d.]).

Since the early 1980s, digital Easter eggs have become a common phenomenon in the digital world. Today, they are even appearing in the automotive industry. The purpose of these Easter eggs is to enhance the experience by providing some additional excitement and surprise (within the context of familiarity so that the user can relate to the content of the Easter eggs). Academic research about Easter eggs is scarce, even though the concept is intensively used. This paper intends to fill the gap about the lack of academic research into Easter eggs, restricting our study to the Wilhelm Scream; it is one of the first scientific analyses of this legendary sound effect.

Each affective display has two main properties; the *intensity* (low-middle-high), and the *valence* (negative-neutral-positive).

To allow for comparison, we interleaved the Wilhelm Scream with other vocalizations while conducting our research project (Binter et al. 2023) in which we asked participants to rate acoustic signals (short, randomly presented vocal displays) of highly intensive affective states (pain, pleasure and fear), asking them to evaluate their valence. We found that the participants who rated these stimuli as positive, neutral, or negative affective states had extremely low accuracy in their judgment (approximately only 50% correct answers). Through further analysis, we concluded that the participants' ratings were statistically equivalent to being due to chance—the raters were guessing (Boschetti et al. 2022; Binter et al. 2023). In other words, the raters were using a trial-and-error approach unsuccessfully. All ratings were conducted twice to test for *consistency* between the first and second ratings in two randomized trials (Boschetti et al. 2022; Binter et al. 2023). The consistency was extremely low in the case of these high-intensity vocalizations (pain, pleasure, and fear).

These results are an example of the so-called *Intensity Paradox* phenomenon (CIT), which claims that the more intensive the display, the more difficult it is to assign a valid valence to it. We, therefore, investigated whether the *Intensity Paradox* also applies to the Wilhelm Scream, since it, too, is a high-intensity emotional display.

We emphasize that the Wilhelm Scream is solely employed to accompany death, injury, or some other negative experience of the portraying character on the screen. We did find a small number of media where the effect accompanied a pleasurable experience; in those cases, however, the effect was used sarcastically in order to enhance the humorous undertone.

2 Methods

2.1 The Data Set

We asked the 902 participants (526 females aged 18–50 years and 326 males aged 18–50 years) to also (in addition to the affective states mentioned above) rate the Wilhelm Scream—twice. Because, during the same session, the participants were also rating other acoustic signals, we were able to interleave the Wilhelm Scream with these others randomly.

2.2 Statistical Methods

The ratings are categorical variables, which may not be converted to computable numbers (Blalock, 1979). Instead, we use a Bayesian approach (Gelman et al., 2014, Kruschke, 2014; Lambert, 2018). In the Bayesian approach, the probability s is a random variable ($0 \leq s \leq 1$), with a distribution called the likelihood function $☹(s)$ (Bishop, 2006).

In the case of two categorical variables (correct versus incorrect, say), in which there are m correct responses and n incorrect responses, then the likelihood function for the probability s of being correct is (Bishop, 2006).

$$☹(s) = const \cdot s^m \cdot (1-s)^n.$$

This likelihood function is the *pdf* (probability density function) of the Beta distribution. The constant *const* ensures that $\int_0^1 ☹(s)\, ds = 1$, so $const = \left(\int_0^1 s^m (1-s)^n ds\right)^{-1}$. We note that $m + n = N$, , which is the total sample size (the total number of registered ratings). The most likely probability is $\frac{m}{m+n}$ (which, for large sample sizes, approaches the Laplace limit; i.e. the traditional, elementary way of defining probability as a ratio.). In the Bayesian approach, the likelihood function includes the inference of the uncertainty (the confidence interval; see below), which the Laplace limit does not (despite the questionable approaches involving the error of the mean; for details, see Lambert, 2018).

Furthermore, the Bayesian approach allows for the determination of whether an observed outcome is due to the raters guessing. One way of quantifying this determination is to calculate depending on which side of $\frac{1}{2}$ the most likely probability s_{ML} is. If this integral is above 5% (the conventional significance level, and the one we have chosen for this publication—any other choice can be chosen, however), then the raters are guessing at the chosen significance level.

$$\int_0^{\frac{1}{2}} 😕(s)\, ds \text{ or } \int_{\frac{1}{2}}^1 😕(s)\, ds.$$

Another method of determining the significance level of a result not being due to guessing is to find the boundaries s_1 and s_2, such that.

$$\int_{s_1}^{s_2} \mathcal{L}(s)\, ds = 0.95$$

for a significance level of 5%. Finding the boundaries s_1 and s_2 involves solving for $\mathcal{L}(s_1) = \mathcal{L}(s_2)$with the above condition. The interval $\{s_1, s_2\}$ is called $\mathrm{HDI}_{95\%}$ (highest density interval at 95% confidence; Kruschke, 2014).

We use these definitions to determine how reliably raters rated the screams—in other words, whether they were guessing.

Because the sample sizes were so large, we use Wilks' Λ (Wilks, 1938) to determine whether the two distributions (likelihood functions $\mathcal{L}(s_A)$ and $\mathcal{L}(s_B)$) for two populations A and B are significantly different.

Specifically, we calculate the log-likelihoods: $LL_A = \ln\big(\mathcal{L}(s_A)\big), LL_B = \ln\big(\mathcal{L}(s_B)\big)$, and $LL_{AB} = \ln\big(\mathcal{L}(s_{AB})\big)$, where AB is the union of both populations. For large sample sizes

$$\Lambda = -2(LL_{AB} - (LL_A + LL_B))$$

is (therefore asymptotically) χ^2-distributed with $df = \big(df_A + df_B\big) - df_{AB}$ degrees of freedom (Wilks, 1938). We note that we can thereby calculate the significance level (as we do here) without specifying it to be 5%

Because the number of males in our sample ($N_{♂} = 376$) is much smaller than the number of females ($N_{♀} = 526$) and the likelihood function $\mathcal{L}(s)$ is sensitive to $m + n$, we use a bootstrap method when comparing male populations with female populations. We randomly choose 376 out of the 526 females and compare their maximum likelihood with the maximum likelihood of the 376 males. We repeat this random selection one thousand times and find the computed comparison distributions.

We used ML methods to find the ML age distributions of participants in a population. We first scaled all ages (by dividing by the maximum age in the data sets—50 years) and then calculated the log-likelihood of the beta distribution, the normal distribution, the log-normal distribution, the Gamma distribution, and the Weibull distribution. (We note that many univariate continuous parametric distributions are subsets of the Gamma distribution with specific values of its parameters.) Again, we used Wilks Λ; to compute significances of comparisons.

3 Results

In contrast to other situations of highly intense affective states (Prossinger et al. 2021; Boschetti et al. 2022; Binter et al. 2021, 2023), we found some unexpected results.

For the 1st rating, 94.7% of the female raters and 91.8% of the male raters correctly rated the Wilhelm Scream. For the 2nd rating, 94.1% of the female raters and 90.2% of the male raters correctly rated the Wilhelm Scream.

In both cases we found, using the described bootstrap method, that the rating difference was not significant at $siglevel = 5\%$. . When analyzing whether the ratings were consistent (i.e. when the ratings were 'doubly correct'), we find that 91.4% of the female raters were 'doubly correct' whereas 'only' 86.2% of the male raters were. The differences between 'doubly correct' and 'not doubly correct' were not significant (females: $siglevel < 4 \times 10^{-4}$; males: $siglevel < 7 \times 10^{-3}$; Beta distribution test).

On the other hand, the distributions of ages showed a remarkable result: the ages of the 'doubly incorrect' females (*mean* = 31.7 years; *stdev* = 8.8 years) were not significantly different from the ages of the 'doubly correct' females (Wilks' Λ; test: *siglevel* = 5%), , while for the males ('doubly incorrect': *mean* = 30.9 years; *stdev* = 8.6 years) there is a highly significant difference (Wilks' Λ; test: *siglevel* $< 4 \times 10^{-6}$).

We observe several outcomes: (1) Neither do all the males, nor all the females, reliably and correctly rate the Wilhelm Scream. Yet the likelihood of correct identification is much higher than expected. All differences are, however, not significant. (2) When asked to repeat the rating, the probability of a 'doubly correct' rating decreases for both the males and the females. (3) The ages of the males rating the Wilhelm Scream incorrectly twice are significantly older (by 2.9 years on average) than those males that rate 'doubly correct'.

4 Discussion and Conclusion

For categorical variables, we could not use point estimators of the most probable q in a binomial test. Nor could we estimate the expectation (via the arithmetic mean) of a distribution. While it is commonplace to use the standard error of the mean to estimate a confidence interval (which is actually illogical, as these are point estimates), it would be useless here: only if we had in excess of 100 samples would we be able to infer that 95% of the point estimators would lie within the intervals bracketing the true mean (the expectation value). Bayesian statistics, which treats all estimators as likelihood functions, allows for a natural estimation of both the confidence interval and the ML true mean—and also the mode (Lambert, 2018).

Because the raters were very much more successful when rating the Wilhelm Scream than when rating the acoustic signals of highly intensive affective states (Boschetti et al. 2022; Binter et al. 2023), we conclude that the Wilhelm Scream is not a 'victim' of the *Intensity Paradox*. This supports the learning hypothesis of valence evaluation suggested by Corvin et al. (2022) and Boschetti et al. (2022)—both using the Bayesian methodology.

The insignificant differences, both for single ratings and combined ratings, as well as for males and females, indicate a high rating precision by the raters, ascribable to a learning effect acquired *before* participating in our study. Arguably, the 902 participants in our sample have been exposed to the Wilhelm Scream in many films and video games; consequently, an association with the negative experiences (depicted in those) will have unavoidably occurred. If so, we must infer that the deviation from 100% is most likely due to random (yet rare) concentration lapses while the participants were asked to rate not only the Wilhelm Scream but also other highly affective states in the same session.

The significant difference in ages between the 'doubly correct' and 'not doubly correct' males is quite large (2.9 years on average; 9.0%), necessitating a more detailed statistical investigation.

Both the statistical methodologies we used and the results we found have implications for game design, because of the involved evolutionary-developmental mechanisms the prospective gamer (unwittingly) underlies during game development.

The use of known sound effects and situations that have been termed "Easter eggs" can support the orientation in novel environments (such as game environments) and

would then serve as building blocks so as to avoid ambiguity—consequently improving and enhancing the users' experiences.

The results show that four stages of the Easter egg experience were identified: awareness, trigger, delivery, and longevity, which are all important phases the game developer needs to incorporate into his/her design.

Acknowledgements. The study was preregistered on the OSF portal: https://osf.io/bhk6m/. We thank the reviewer for very helpful comments and suggestions.

Funding. This research was funded by the Czech Science Foundation in the project with number GACR 19-12885Y and titled "Behavioral and Psycho-Physiological Response on Ambivalent Visual and Auditory Stimuli Presentation".

This work was supported by the Cooperatio Program, research area ARTS.

Conflicts of Interest. The authors declare no conflict of interest.

References

Bishop, C.M. Pattern Recognition and Machine Learning. Springer Science + Business Media, NY (2006)

Binter. J., Boschetti, S., Hladký, T., Prossinger, H.: Ouch!" or "Aah!: Are Vocalizations of 'Laugh', 'Neutral', 'Fear', 'Pain' or 'Pleasure' Reliably Rated? Under Review (2023)

Binter, J., et al.: Quantifying the rating performance of ambiguous and unambiguous facial expression perceptions under conditions of stress by using wearable sensors. In: International Conference on Human-Computer Interaction, pp. 519–529. Springer, NY (2022)

Blalock, H.M.: Social Statistics, 2nd edn. McGraw-Hill, NY (1979)

Boschetti, S., Prossinger, H., Hladký, T., Machová, K., Binter, J.: "Eye can't see the difference": facial expressions of pain, pleasure, and fear are consistently rated due to chance. Human Ethology **37**, 46–72 (2022)

Corvin, S., Fauchon, C., Peyron, R., Reby, D., Mathevon, N.: Adults learn to identify pain in babies' cries. Curr. Biol. **32**(15), R824–R825 (2022)

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. Bayesian Data Analysis. 3rd edn. CRC Press, Boca Raton

Kruschke, J.K.: Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Academic Press, NY (2014)

Lambert, B.: A Student Guide to Bayesian Statistics. Sage Publications, London, UK (2018)

Prossinger, H., Hladky, T., Binter, J., Boschetti, S., Riha, D.: Visual analysis of emotions using ai image-processing software: possible male/female differences between the emotion pairs "neutral"–"fear" and "pleasure"–"pain". The 14th PErvasive Technologies Related to Assistive Environments Conference, pp. 342–346 (2021). https://doi.org/10.1145/3453892.3461656

Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**(1), 60–62 (1938). https://doi.org/10.1214/aoms/1177732360