Mal for sensorveiledning

Emnekode	Psy3121
Emnenavn	Forskning og prosjektutvikling
Emneansvarlig/oppgavegiver	Robert Biegler
Kvalitetssikret av	
Semester, år	Vår 2024
Vurderingsform, lengde	Oppgave
Tillatte hjelpemidler	Alle hjelpemidler tillatt. Including the document on criteria for judging papers included here: **Criteria for judging papers, ordered by what you need to know before you can make sense of the next bit. (Ordering by importance would give you a different ranking.)** **Robert Biegler, NTNU** I assembled this checklist of things to look for when reading scientific literature from various sources over the years, then found something pretty similar in Carey et al (2020) and in Schiermeyer (2016). These criteria for evaluating papers are also useful when writing. Your readers will apply these criteria. Does your writing satisfy them? I then added more fine-grained distinctions, most from Meltzoff & Cooper (2018; point 2 is a quote from that book). If you are reading or writing a review, then the points relevant to methods, data analysis, and results do not apply. 1) How interesting is the research question and why? Why should you care? Why read this and not something else? 2) What is the type of research question? (You don't need to report this in a review, but correctly categorising the research question helps you judge whether the analysis is appropriate, and whether the conclusions deal with the original research questions.) a) Existence: does *x* exist? b) Description and classification: i) What are the characteristics of *x*? ii) To what extent does *x* exist? iii) Does *x* take on multiple values or is it invariant?

	iv) What are the limits of *x*? v) Is *x* unique or does it belong to a known class (taxonomy) of things? c) Descriptive-comparative: Is group *x* different from group *y*? i) Are men more aggressive than women? ii) Are younger people more liberal than older people? iii) Are wealthy people happier than poor people? d) Statistical relationship: is there a relationship between *x* and *y*? i) Is happiness related to income? ii) Is there an association between time spent studying in college and college grades? e) Causal relationship: Does *x* produce, lead to, or prevent changes in *y*? i) Does smoking marijuana reduce epileptic seizures? ii) Does playing violent video games make children more aggressive? iii) Does psychotherapy change behavior (an exploratory question because the behaviors are not specified)? f) Causal-comparative: does *x* cause more change in *y* than does *z*? i) Is studying alone more effective for humanities classes but studying in groups more effective for science classes? ii) Is a certain medication more effective than psychotherapy in treating depression among men than women? iii) Are face-to-face job interviews better than online interviews for making successful hiring decisions depending on the type of position being filled? 3) Is the argument, the introduction's explanation of what is to be researched, logical? 4) How well does the argument fit empirical data that you know? Are relevant data being ignored? 5) Has the research question been translated into testable hypotheses? If yes, do these hypotheses follow from the research question(s), or is there some logical disconnect? Ideally, you get hypotheses by using theory to make predictions. Even a single theory may make multiple detailed predictions, and there may be

	competing theories whose predictions are to be compared. Therefore a single research question can give rise to multiple hypotheses. 6) Are the methods described clearly enough and in enough detail that you can make an informed decision whether the methods are suitable to answer the research question? a) What kind of study is this? i) Observational or experimental? ii) Prospective or retrospective? iii) Cross-sectional or longitudinal? iv) Simulation or real life? v) Laboratory or field? b) Who are the participants? i) How are participants selected? ii) Is the sample representative of the target population? iii) How are participants assigned to groups or conditions? iv) Is there attrition? Is it selective? c) The hypotheses should predict, from identifiable predictor variables, the values or trends (increasing or decreasing, accelerating or decelerating) of identifiable outcome variables. What are these variables? How are they measured? d) Scientific hypotheses are usually concerned with conceptual or latent variables. How do these relate to the actual data? Are the data and whatever is calculated from them valid measures of the conceptual variables? 7) Are the methods suited to answer the research question? Have the authors conducted all appropriate controls? a) Are the observed data reliable and valid measures of conceptual variables that are relevant to the hypotheses? i) If the observed outcomes are scores of a test, has that test been validated previously? If yes, for what population? ii) Is there any report on the reliability of whatever measurement instrument is being used?

iii) If raters are involved, how many? What is their training? How much do they agree? How are disagreements resolved?

b) Have other variables that could affect the observed data been counterbalanced, or at least measured so that the statistical analysis can estimate their influence? For example, if there are two different tests, has their order been counterbalanced to account for order effects such as fatigue or becoming more comfortable with the testing situation? If the same test is applied before and after a treatment, is there also a control group without treatment that can provide a baseline for maturation, spontaneous recovery, or carry-over effects?

8) Is the statistical analysis appropriate to the experimental design?

a) Is the analysis appropriate to the type of measurement (nominal, ordinal, difference, ratio)?

b) Do the data match the criteria needed for this statistical analysis? For example, some analyses need (approximately) normal distributions, and possibly equal variances.

c) Is the analysis appropriate to the predicted relationship between predictor variable and outcome variable? For example, if there are several nominal predictors and one continuous outcome (and normal distributions and equal variance), ANOVA is a suitable analysis. Change the predictors to continuous (with normally distributed residuals), and some form of regression is suitable.

9) Are the findings adequately described and discussed?

a) Is there an analysis of data relevant to every prediction in the hypotheses?

b) Are the data coherent? Do percentages add up? (Note that if, for example, 33.4%, 33.4%, and 33.2% are rounded to integers, then 33%, 33% and 33% add up only to 99%, so you should worry about deviations that can't be explained by rounding.) Are means and standard deviations mutually consistent? Are the degrees of freedom correct?

c) Are measures of dispersion (standard deviations, standard errors, variances, ranges, quartiles) being reported where appropriate, including in figures?

d) Are effect sizes being reported? If yes, is the effect size measure appropriate to the analysis? For example, is it

	important to know relative risk, absolute risk difference, or both? e) Do figures show the important results without misleading you? If there is a ratio measurement on the vertical axis, is zero included? Are there error bars, or box plots or some other visual display of the distribution of data where you need it? Do visual symbols give a misleading impression of outcomes? Is contrast being used to make some results stand out while hiding others? Are important results simply missing from the figures? 10) Is there any sign of p-hacking? a) Were hypotheses chosen after the analysis? One sign of that is more positive results than you would expect from the effect size and sample size in studies with multiple experiments. Another is analyses showing up in the results that are not mentioned in the methods. These may be either entirely new analyses, or pairwise comparisons that are neither justified by significant interactions nor by hypotheses that call for these comparisons. Also, if analyses that are mentioned in the methods, or that are needed to test stated hypotheses, just quietly disappear, the reason might be that the results were inconvenient. b) Is there any reason to suspect unreported degrees of freedom in the analysis? c) Were exclusion criteria defined before data analysis? Are participants excluded without clear reason? d) If the study was preregistered, does the analysis deviate from the published analysis plan? 11) Are the claims and conclusions justified by the data? a) For each conclusion, is there either a statistical analysis to back it up, or a reference to other research? b) Do the results actually support the conclusions? Is a "non-significant trend" being treated as support? Is the absence of a significant effect being treated as evidence that there is no effect, even though the analysis is neither Bayesian nor equivalence testing? 12) Are the conclusions an answer to the original research question? (It is good science to conduct exploratory analyses and report any interesting result, but it should be made clear that this generates a new hypothesis that must be tested with new data. Passing off

	exploratory analyses as confirmatory is either bad science or an outright lie.) 13) Is the answer supposed to be an inference, or an explanation? (An inference is a new expectation, prediction or belief, based on both prior knowledge and the new information obtained through the experiment. "I expect this rock will fall if you let go of it" is an inference. An explanation adds to the inference a set of causal relationships that can generate the observed results. A theory of gravity would explain why the rock falls if you let go of it. It is an inference if you order the items in this list as they are because that closely matches the order in scientific papers, and when lots of smart people agree on something, they probably have a reason. It is an explanation if you give, for each item in turn, a reason why you must have *those* pieces of information before you can understand *this* item, and therefore information must come in this order.) 14) Is the interpretation of the data justified in light of available theory? (Important: this does *not* mean you should only accept interpretations that fit available theory. If the authors soundly argue that data contradict available theory, that is *appropriate in the light of available theory*. It would *not* be appropriate to present contradictory data as if they supported available theory, or to quietly ignore contradictions as if they weren't there.) How are data that contradict theory being treated in the discussion? See where they fit in Chinn & Brewer's (1998) scheme and decide whether you accept that as sound reasoning. a) Are inconvenient data just being ignored, as if they were not valid, or not relevant to the theory, but without explanation of *why* the data are supposed to be invalid or irrelevant? b) If there is an explanation for why data are rejected as invalid, do you agree with the reason for rejection? c) If data are being excluded from consideration because it is uncertain whether they are valid or relevant, do you agree that the uncertainty is justified? d) If data accepted as valid are, without explanation, excluded as irrelevant, is that acceptable? e) If authors say a theory can stand because an explanation for how the data don't actually contradict theory will eventually be found, do they have a reason for that expectation, or only pious hope?

f) If the data are reinterpreted as actually supporting theory (Meltzoff & Cooper's practice papers gave several examples of Freudian reinterpretations), do you agree? If yes, was the initial interpretation of the theory wrong, or is the possibility of reinterpretation a sign that the original theory is vague, or that it was being wrongly applied to a domain for which it is not valid?

g) If the authors argue that the contradictory data can be accommodated with only peripheral theory change, do you agree? An example is astronomers not abandoning Newtonian mechanics when they noticed unexplained perturbations in the orbit of Uranus. They assumed there was another planet, calculated how large it had to be and where, and found Neptune. But perturbations in the orbit of Mercury were not explained by another planet Vulcan, but by relativity.

h) If the authors argue the theory must be abandoned or its core assumptions revised, do you agree? (The authors may just not like the theory and set out to prove it wrong.)

				Current theory altered?		
				No	Partly	Completely
Data accepted as valid?	Yes	Data explained?	Yes	*Reinterpretation*	*Peripheral theory change (Neptune)*	*Complete theory change (relativity, not Vulcan)*
			Not yet, but later	*Abeyance*		
			No	*Exclusion (data irrelevant)*		
	Uncertain	Data explained?	Yes			
			No	*Uncertainty re. Validity*		
	No	Data explained?	Yes	*Rejection*		
			No	*Ignoring Exclusion (data irrelevant)*		

15) Is there adequate replication? (Is the effect size so enormous that a single experiment convinces you? Do you know how robust the reported result is against changes to the method, or the analysis?)

16) What research should be done next?

17) Is the research ethically justifiable?

	a) Were participants being deceived or coerced? If they were being deceived, why, and was the cost to the participants minor enough that deception can be justified by the benefit for research? b) Was participants' mental or physical health at risk? If the participants are not human, should you expect them to be capable of feeling distress? How much might they suffer? How does the experiment balance that against gain in knowledge? c) Did participants consent? How informed was their consent? If the experiment can't be done with prior consent (the experimenter wants to know how people behave when they don't know they are being studied because that knowledge would change how they act), can negative effects be expected to be so minor that experimenters may ask afterwards? As a rule of thumb, if a reasonable person would think the experimenter is an arsehole for doing this, then an ethics committee would have a problem with it, and so should you. An action can be unethical without being illegal, so remember that scientific ethics apply stricter standards than the law. d) Do the authors have conflicts of interest? If yes, were they declared, including to participants? 18) Are all the important claims in the paper supported by either references, or, in discussion and conclusions, by the paper's own data? Do all the references show up in the reference list? Does the reference list contain sources that are not cited in the paper? 19) Now that you know the paper, return to the abstract. Does it fairly describe the content of the paper? Does the paper deliver all the abstract promises? If you review for a scientific journal, you usually have four options for your final recommendation to the editor: 1) Accept as is, without changes. That is very rare. 2) Publish if either changes are made as recommended or the authors give good reason why they won't change, and leave it up to the editor to decide to what extent the authors have met either condition.

	3) The authors should make changes and submit their work again for a new review (you do this if you are not sure whether the recommended changes will make the paper good enough, or if you ask for additional data to be collected). 4) Reject. As a courtesy to both editor and authors, it should be very clear what changes you recommend, and they should be easy to find. You could put each recommendation into a separate paragraph, after you explain the problem you identified and why you think the recommended change is a solution to that problem. Or you could repeat the recommended changes in a list at the end. These options also apply to giving feedback to a colleague, though "reject" becomes "it's better for your reputation if you don't submit that". References Carey, M. A., Steiner, K. L., and Petri, W. A. (2020). Ten simple rules for reading a scientific paper. PLOS Computational Biology, 16(7): e1008032. https://doi.org/10.1371/journal.pcbi.1008032 J Chinn, C. A. and Brewer, W. F. (1998). An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science. *Journal of Research in Science Teaching*, **35(6)**, 623 – 654. Meltzoff, J. and Cooper, H. (2018). *Critical Thinking About Research*. American Psychological Association, Washington, DC. Schiermeier, Q. (2016). Close inspection. *Nature* **533**: 279-281.

Emnets læringsutbyttebeskivelser angitt I kunnskaper, ferdigheter og generell kompetanse. (Henvisning med lenke til emnesiden på NTNUs nettsider er tilstrekkelig)	https://www.ntnu.no/studier/emner/PSY3121#tab=omEmnet
Pensum	Goodman, S. and Greenland, S. (2007). Assessing the unreliability of the medical literature: A response to "Why most published research findings are false". Johns Hopkins

NTNU

Institutt for psykologi

	University, Department of Biostatistics. Available: http://www.bepress.com/jhubiostat/paper135. Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8), 696-701. Lenharo, M. 2023 Consciousness: The future of an embattled field. Meltzoff, J, and Cooper, H. (2018). Critical Thinking about Research. American Psychological Association, Washington DC. Simmons, J. P., Nelson L. D., and Simonsohn, U. (2011). False-Positive Psychology: Flexibility in Data Collection and Analysis Allows Presenting anything as Significant. *Psychological Science*, 2011 (Nov), Vol 22(11), 1359-1366
Eventuelle formelle krav til besvarelsen	Ingen
Hvordan de ulike oppgavene i eksamenssettet er vektlagt	See below

Oppgavetekst:

Dear Dr InsertNameHere

The Journal of Experimental and Inadvisably Applied Theology requests your expert judgement on the manuscript appended below, which has been submitted to the journal. Please indicate your overall recommendation (publish as submitted, publish with minor revisions, publish with major revisions, reject) as well as the specific strengths and weaknesses of the submitted work. For each specific issue you address, state the line number, or range of line numbers, of the text you comment on, or the figure. Quote the relevant text if necessary. Then summarise your most important concerns.

If the submitted work fails to adequately address the research questions, please recommend methods for a better experiment, focusing on flaws that invalidate the research.

Yours sincerely

Tryphona Seemerrie Brain (editor)

NTNU

Institutt for psykologi

An Empirical Inquiry into Abominations Unto Nuggan

Zebulon Q. Wrigglesworth, Marvellous Scanfield, Igor, and Loveday Snashfold

Borogravian Institute for the Study of Heresy

Abstract

Abominations unto Nuggan are manifold, foremost among them chocolate. The Faithful are only allowed to possess chocolate with intent to sacrifice it on holy days. However, the manner of sacrifice itself may give rise to Abomination. Here, we examine how the manner and order of sacrificing different flavours of chocolate may please Nuggan, and whether that depends on species (human, dwarf or troll). Implications for wedding feasts will be discussed.

Introduction

Nuggan is a jealous and most exacting god. The Faithful may find joy in their worship of Nuggan, and must avoid distraction by secular pleasures (Buttery & Quitquit, Century of the Fruitbat 13). The highest of these pleasures is the consumption of chocolate in any of its forms (Genussmensch & Snodgrass, Century of the Fruitbat 17), with the exception of those brands from the rebellious colonies that contain less than 12.5% coacoa and more than 40% chalk or powdered straw, which may instead be consumed as penance for heinous sins (Trubelmacher et al., Century of the Insignificant Mollusc 98). All other chocolate products displease Nuggan and must be sacrificed. The question is what manner of sacrifice pleases Nuggan, whether different flavours of chocolate must be sacrificed in a specific order, and whether that depends on the species to which the Faithful belong.

Methods

Participants

The Most Holy Exquisition apprehended 537 individuals at the Bad Bath Hogswatch Market, 299 humans (128 male, 169 female, 2 other), 174 dwarves (all self-declared male, and even the Most Holy Exquisition knows better than to enquire further[1]), and 56 trolls.

Procedure

All participants were kept for three days in individual cells under the Temple, supplied with only water for the first two days. On the second day, three blocks of chocolate were supplied (white chocolate, milk chocolate, dark chocolate), and participants were instructed to sacrifice the chocolate by throwing it, one block at a time, in the order of their choice, onto either burning coals in a brazier, or to a pig. Either brazier or pig was brought before each cell for that purpose. Participants were detained for a further day, and the degree of Nuggan's

[1] It is reliably believed that dwarves come in enough varieties to make more dwarves, but dwarves tend to respond to queries about their sex with armed displays of masculinity that thoroughly discourage further questions. The only known self-confessed female dwarf serves in the Ankh-Morpork City Watch, and no members of that organisation were apprehended for this experiment.

displeasure was quantified by the number and size of boils observed on the participant's body. At this time, the geological composition of the trolls was also recorded.

The number and diameter of boils was recorded. Dwarves are both modest and heavily armed, so we recorded the number and diameters of boils on their hands and what parts of their faces were not covered by either beard or helmet, estimated body surface area from height and girth, and extrapolated the total number of boils from that.

Results

16 trolls walked through their cell doors and left the Temple during their first night of detention. Of the remaining trolls, 21 consist of sedimentary rock (6 sandstone, 5 chalk, 4 mudstone, 4 siltstone, 2 claystone), 14 of igneous rock (7 granite, 4 gabbro, 3 basalt), and 6 metamorphic rock (2 gneiss, 2 slate, 1 marble, 1 quartzite). None of the igneous or metamorphic trolls showed any sign of boils. Sedimentary trolls showed an average of 14.6 ± 4.9 ammonites, the troll equivalent of boils, covering 0.02 ± 0.0105% of surface area. We found a main effect of the mode of sacrifice (chocolate destroyed by fire: 12.9 ± 5.2 ammonites, chocolate thrown to a pig: 14.5 ± 5.7 ammonites, $t(40) = 2.01$, $p < 0.05$). Further, the number of ammonites decreased as the grain size of the rock increased (Pearson's $r = -0.498$, $p < 0.05$; see Figure 1).

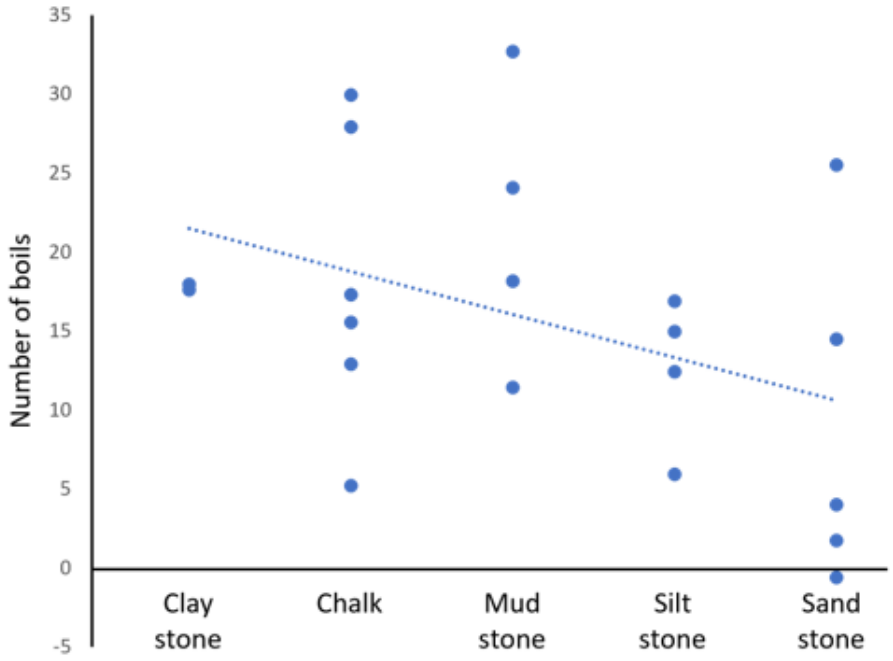

Figure 1. Correlation between the type of rock comprising a sedimentary troll, ordered from smallest to largest grain size, and the number of boils.

NTNU

Institutt for psykologi

71 dwarves dug tunnels out of their cells during the first night of confinement. Of these tunnels, 5 led into the cells of trolls who prematurely dropped out of the experiment. There were signs of struggle (pickaxe marks, impacts of hammers, dwarf bread, and rock fists on the walls, on what remains of the cell doors, and on the stairs and corridors leading out of the dungeon), but no remains, body parts, or bodily fluids were found. The remaining 105 dwarves installed new and improved locks in their cell doors as well as in the dungeon's front door, and were only willing to emerge after the Temple signed contracts to pay for both the locks, and the keys to those locks.

48 dwarves sacrificed chocolate to fire, 49 to pigs, 8 were excluded for not following instructions. We found a significant effect of both mode of sacrifice (mean ± standard deviation for sacrifice by fire: 26.7 ± 14.5 boils; sacrifice by pig: 23.2 ± 13.9 boils; $F(1, 93) = 6.77$, $p < 0.01$) and of the order of sacrifice (mean ± standard deviation for lightest to darkest chocolate: 30.9 ± 17.6 boils; darkest to lightest chocolate: 19.5 ± 6.6 boils; $F(1, 93) = 17.83$, $p < 0.001$). Further, we found there was no difference between sacrifice by fire and sacrifice by pig if the order of sacrifice progressed from lightest to darkest chocolate ($t(44) = 1.15$, $p > 0.25$), but there was a difference when the order of sacrifice progressed from darkest to lightest chocolate ($t(49) = 3.75$, $p < 0.001$). See Figure 2.

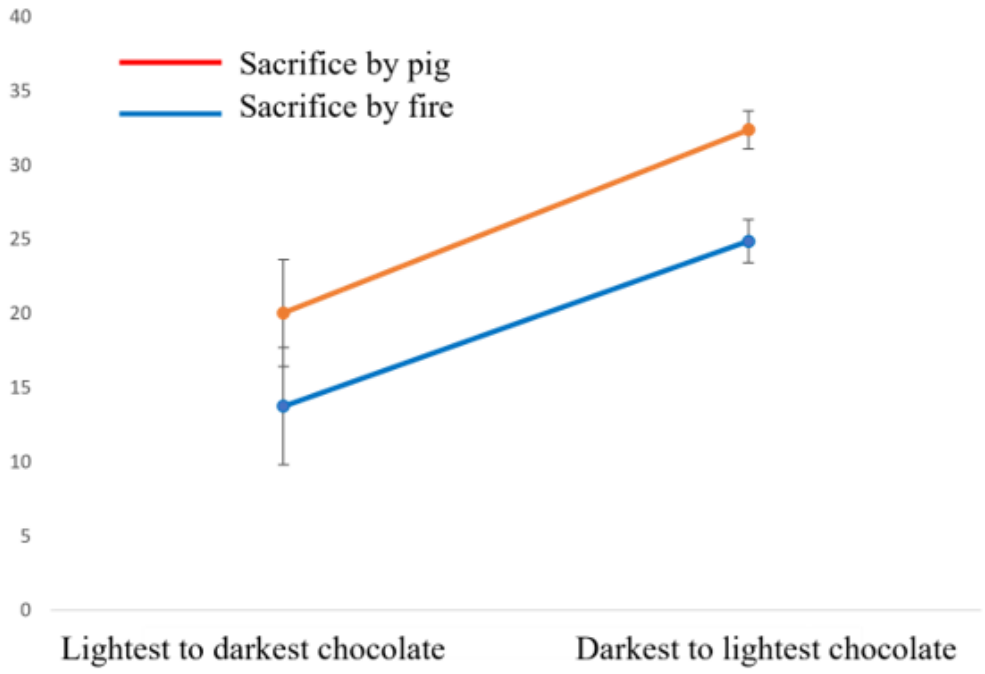

Figure 2. Dwarf data

134 humans sacrificed chocolate to fire, 132 to pigs, 35 were excluded for not following instructions. We found a significant effect of both mode of sacrifice (mean ± standard deviation for sacrifice by fire: 26.7 ± 14.5 boils; sacrifice by pig: 23.1 ± 14.0 boils; $F(1, 258) = 6.77$, $p < 0.01$) and of the order of sacrifice (mean ± standard deviation for lightest to darkest chocolate: 13.8 ± 17.7 boils; darkest to lightest chocolate: 19.5 ± 6.6 boils; $F(1, 258) = 54.90$, $p < 0.001$, $\eta^2 = 0.168$). Further, we found there was no difference between sacrifice

by fire and sacrifice by pig if the order of sacrifice progressed from lightest to darkest chocolate ($t(128) = 1.89$, $p > 0.05$), but there was a difference when the order of sacrifice progressed from darkest to lightest chocolate ($t(134) = 2.23$, $p < 0.05$). Finally, men had more boils than women ($F(1, 258) = 4.00$, $p < 0.05$). See Figure 3.

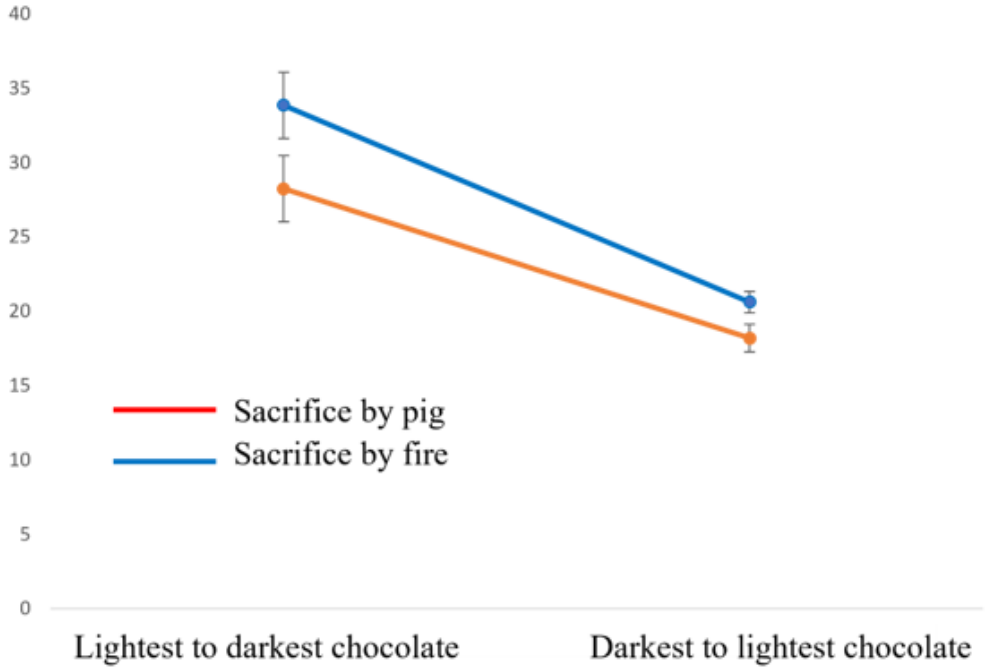

Figure 3. Human data

Discussion

A sacrifice is a sacrifice in name only if the sinner does not suffer for it. That explains the absence of boils on the igneous and metamorphic trolls, who fail to distinguish between different types organic matter, or oograh, and consume it only occasionally for the mineral content of whatever bones may be present. Chocolate is unlikely to be a temptation, and foregoing chocolate is no sacrifice. Future research should examine whether fireworks flavoured with different salts may serve as the trollish equivalent of chocolate (Kabloom & Alvarez, Century of the Fruitbat 04).

Sedimentary trolls, however, are different, and provided meaningful results. Organic inclusions in sedimentary rock make organic matter nutritious. The resulting craving for chocolate makes the sacrifice relevant.

In both dwarves and humans, it makes no difference whether chocolate is sacrificed to fire or a pig, provided that the order of sacrifice proceeds from lightest to darkest. In contrast, when the order of sacrifice proceeds from darkest to lightest, Nuggan is less displeased when dwarves sacrifice by fire, and humans sacrifice by pig. The species also differ in that only humans show a sex difference.

NTNU

Institutt for psykologi

101 **Conclusions**

102 Nuggan is not particularly pleased with anyone. Abominations unto Him are many.
103 Nevertheless, methods of sacrifice have measurable influence on the degree of mercy that
104 Nuggan may bestow.

105 Finally, the Temple of Nuggan the Most Holy and Pernickety requests generous donations for
106 the purpose of rebuilding the temple edifice, which fell into a hole that mysteriously appeared
107 underneath the building soon after the experiment ended.

108 **References**

109 Buttery, Y. & Quitquit, X. (Century of the Fruitbat 13). Reinterpretation of the Revelations of
110 the Prophet Nigel in the Light of the Edict of Lorenzo the Magnificent. *Journal of Expedient*
111 *Theology, 27(2)*, 254 – 267.

112 Genussmensch, O. and Snodgrass, P. (Century of the Fruitbat 17). Absolute Hedonic Values
113 of 113 Experiences: the Dark Chocolate Scale. *Annual Review of Pleasure, 71(9)*, 17 – 38.

114 Snashfold, L. (Century of the Fruitbat 16). *The Hammer of the Heretics: On the Importance*
115 *of Rooting Out Deviations from the Pure and True Faith.* Temple of Nuggan Press,
116 Borogravia. ISBN-13: 666-3141592654

117 Trubelmacher, A. G., D'Belfrey, B., Nasruddin, W. F. and Dobbs. E. (Century of the
118 Insignificant Mollusc 98). A Meta-analysis of the Relationship Between Penance and
119 Redemption. *Theological Review, 1125(1), 998 – 1014.*

Sensurveiledning:

Instructions for examiners: below, and also in the accompanying Excel file, there is a list of line numbers and figures, and the flaws I inserted there. The Excel file contains an importance weighting: 1 for minor points, 2 for something moderately important, 4 for a fatal flaw that invalidates the whole study. Students receive points based on importance and difficulty, where difficulty is measured by the proportion of students who spot each of the documented flaws according to the following formula:

Points for a true positive = importance * (1/(proportion of students who spotted this flaw)$^{0.3}$

If a student reports a flaw that I missed, I treat that as discovered by a proportion of 0.01, no matter how many students report this, meaning 3.98 points * importance. I will have to judge importance when I see what a student has spotted.

A false positive costs 2 points, else the incentive is to treat everything as a flaw, without discrimination.

I will then use the worst of the exams that I still consider a C as my reference to set the boundary between C and D, setting the boundary 1 point below whatever that student

gets. D – F then will each have an equal range of points down to 0, and likewise A – C up to the maximum. If, collectively, all flaws are reported, the upper boundary of an A is the best point score. If, across all students, not all flaws are discovered, I will add to the point range the points awarded for one more flaw, discovered by one student, weighted by importance.

Note that it is impossible to know the range of points in advance because the more that is being missed, the greater the point reward for those who do report the more difficult to notice flaws. Also, ***if a student appeals, consistent grading is only possible by consulting the spreadsheet to find the proportion of students who noticed each flaw and the corresponding calculation of points awarded. Given that this information is not available before the exam, it will be necessary to ask for this information after initial grading.***

This part of the exam counts for two thirds of the final grade.

This file also contains my version of overall comments and recommendations for a revised study. The purpose of asking the students to write this and focus on the most important issues is to find out what they find most important. They should deal with all the fatal flaws, and perhaps the more important of those that are not quite fatal. Note that one recommendation may deal with several related flaws that are separate items in the list and spreadsheet. The number of recommendations does not have to match or exceed the number of fatal flaws that are listed separately.

This part of the exam counts for one third of the final grade.

Lines 1 – 3, importance = 0 because this is an administrative matter, not scientific: Who is the corresponding author?

Lines 3 and 99, importance = 4: The authors fail to explain whether there are conflicts of interest. Is the Borogrovian Institute *for the Study of Heresy* a neutral research institute, or is part of its remit the suppression of heresy? The title of Snashfold's book in the references suggests the latter. If the Institute is tasked with guarding the purity of the Faith, then there is a conflict of interest, and one of the consequences of that conflict will be strong pressure on participants to avoid anything that could give the appearance of heresy. That last issue os what makes this so important.

Lines 9 – 10, importance = 1: "Implications for wedding feasts will be discussed." There is no such discussion.

Lines 20 – 22: "The question is what manner of sacrifice pleases Nuggan, whether different flavours of chocolate must be sacrificed in a specific order, and whether that depends on the species to which the Faithful belong."

Importance = 2. How did that become the research question? Is it empirically motivated, either by previous data or by anecdote. What is the connection between Nuggan insisting on chocolate being sacrificed and the manner of sacrifice? If the research question is theoretically motivated, then there should be an explanation of what mediates between manner of sacrifice and Nuggan's displeasure. See line 85.

Importance = 2. The research question should have been translated into specific hypotheses. There are none.

Line 25: “The Most Holy Exquisition apprehended 537 individuals”.

Importance = 4. Does not meet ethical standards. People who were apprehended are unlikely to have given free and informed consent.

Importance = 1. Also, haphazard sampling gives low confidence that the sample was representative of any population but those who were apprehended.

Lines 25 – 27, importance = 2: The text states that 537 individuals were apprehended, 299 humans, 174 dwarves, and 56 trolls, adding up to only 529. 8 are not accounted for.

Lines 25 – 27, importance = 2: The research question is specifically concerned with the actions of “the Faithful”, so where is the assessment of the faith of participants? That should be included with the demographic information.

Line 26: “174 dwarves (all self-declared male”.

Importance = 1. There had better not be an analysis of sex differences in dwarfs, both because the reported sex is all male and because the footnote makes clear that this report is unreliable. However, the discussion (lines 97 – 98) mentions that only humans show a sex difference, so an unjustified claim is being made. [That unjustified claim is scored separately later.]

Line 27, importance = 2: Trolls’ sex is not mentioned.

Line 29, importane = 4: Ethics. What are the health implications of fasting for that long? Were participants screened for conditions that could exacerbate the effects of fasting this long? Were medical personnel available to deal with any complications resulting from this fasting?

Lines 29 – 34, importance = 2: Inconsistent timeline. Three days detention, supplied with only water for the first two, yet chocolate is supplied on the second day, one day after being apprehended, and people were only kept for one day beyond that, for a total of two days.

Line 29: “..supplied with only water for the first two days.”

Importance = 4. Ethics. It is unlikely that people agreed to this fasting. There is also no mention of food being supplied between sacrifice and release.

Importance = 4. Why were people supplied with only water? Is it to increase the craving for any food, and so increase the sacrifice made by giving up the chocolate? If so, that would be motivated by the mediating variable hinted at in line 85, and should have been made explicit.

Importance = 2. Do species or individuals differ in how quickly the feeling of hunger grows as they are deprived of food? If the mediating variable is subjective suffering, as suggested on line 85, that matters, and must be recorded and analysed.

Lines 30 – 31: “On the second day, three blocks of chocolate were supplied (white chocolate, milk chocolate, dark chocolate)”.

Importance = 2. How were these handed over? One at a time, or in a stack?

Importance = 2. In either case, what was the order? People might have thrown chocolate into either the fire or to the pig in the order in which they had the blocks of chocolate in their hand, without paying attention.

Importance = 2. Was the chocolate packaged in any way, in which case, did participants even know there were different kinds of chocolate, or was the chocolate handed over without packaging?

Importance = 1. What was the weight of each bar of chocolate?

Lines 31 – 32: “participants were instructed to sacrifice the chocolate by throwing it, one block at a time, in the order of their choice”.

Importance = 4. The assignment to treatment is not random. It will be impossible to distinguish an effect of the order of sacrifice from an effect of a pre-existing group static variable that biases people towards some order.

Lines 31 – 32: “participants were instructed to sacrifice the chocolate by throwing it, one block at a time, in the order of their choice”

Importance = 4. The assignment to treatment is not random. It will be impossible to distinguish an effect of the order of sacrifice from an effect of a pre-existing group static variable that biases people towards some order. This invalidates the whole experiment.

Lines 31 – 33: “participants were instructed to sacrifice the chocolate by throwing it … onto either burning coals in a brazier, or to a pig.”

Importance = 2. Did participants have a choice, both brazier and pig being brought to each cell, or were they assigned the brazier or the pig? If they were assigned a mode of sacrifice, was that assignment random? The assignment process, if any, is not described.

Importance = 2. Did participants have a choice, both brazier and pig being brought to each cell, or were they assigned the brazier or the pig? If they were assigned a mode of sacrifice, was that assignment random? The assignment process, if any, is not described.

Importance = 2. If there was only one pig, and hundreds of participants, all tested on the same day, the pig probably stopped being interested in chocolate at some point. The typical bar of chocolate weighs 100 g, and each participant is instructed to sacrifice three blocks. If they did (49 dwarves, 134 humans and an unknown number of trolls sacrificed to pigs), then at least 54.9 kg of chocolate were thrown to pigs. That is a lot of chocolate, even for the largest pig. So were there multiple pigs? If so, that should have been reported, and the identity of the pigs should be included in the analysis. Whether there was one pig or several, how much chocolate a pig had already eaten should also have been included as a predictor in the analysis.

Lines 34 – 35: “the degree of Nuggan’s displeasure was quantified by the number and size of boils observed on the participant’s body.”

Importance = 4. Is that a valid measurement of Nuggan's displeasure? No prior results are referenced, so if this is the first time this indicator of Nuggan's displeasure is being used, it needs to be validated. There should be control groups (plural because the experimental groups are being divided by species) that are not exposed to temptation and are not being made to sacrifice, and for both control and experimental groups there should be measurements of boils before and after the experimental groups' sacrifice manipulation. Assignment to control or experimental groups must be random.

Line 36: “At this time, the geological composition of the trolls was also recorded.“

Importance = 2. Are there associated ethical concerns? Is geological composition a sensitive issue for trolls? How was the geological composition determined? By asking the trolls, having a geologist look at them, taking a sample?

Line 37: What counts as a boil?

Importance = 2. How was the diameter measured? How reproducible is that measurement?

Importance = 2. Are those who measure and count the boils blind to the treatment (sacrifice by fire versus sacrifice by pig)? They are certainly not blind to the species of the participant.

Lines 37 – 40: How was the area of face and hands estimated?

Importance = 2. Does the extrapolation assume uniform distribution of boils, and if yes, is there any reason to believe that assumption is justified? What is the uncertainty associated with extrapolating from a small sample of boils?

Importance = 2. Did humans and trolls strip to allow boils being counted anywhere on the body, or was an analogous extrapolation done for surface covered by clothes?

Importance = 4. Where is the analysis plan?

Line 43: “16 trolls walked through their cell doors and left the Temple during their first night of detention.“

Importance = 2. Not enough to address concerns over involuntary confinement. Some trolls could just walk away, but did the trolls who stayed behind do so because they chose to stay, or because they could not punch through the doors?

Lines 44 – 46: “Of the remaining trolls, 21 consist of sedimentary rock (6 sandstone, 5 chalk , 4 mudstone, 4 siltstone, 2 claystone), 14 of igneous rock (7 granite, 4 gabbro, 3 basalt), and 6 metamorphic rock (2 gneiss, 2 slate, 1 marble, 1 quartzite).”

Importance = 2. Geological composition is not mentioned in the introduction, and no hypothesis has been proposed to which it is relevant. Looks like harking. Further, geological composition was not determined until after some trolls had left, so there may have been differential attrition.

Line 44: “6 sandstone, 5 chalk“

Importance = 2. Doesn’t fit the figure, which has 6 chalk and 5 sandstone.

Line 48: “ammonites, the troll equivalent of boils”

Importance = 4. Are ammonites the troll equivalent of boils? If that claim is not being justified here or in the methods, where such an equivalence should have been mentioned and justified, there should at least be a reference to back up the claim. As it is, it is not at all clear whether this measures what it is supposed to measure.

Line 48: "covering 0.02 ± 0.0105% of surface area"

Importance = 2. The methods mentioned number and *diameter* of boils, not proportion of body surface area covered by boils.

Importance = 2. Calculating that proportion depends on estimating the body surface area, too. How was that done? The information provided in connection with the extrapolation from dwarfs' faces to the rest of their bodies is not enough.

Lines 47 – 50: "14.6 ± 4.9 ammonites, the troll equivalent of boils, covering 0.02 ± 0.0105% of surface area. We found a main effect of the mode of sacrifice (chocolate destroyed by fire: 12.9 ± 5.2 ammonites, chocolate thrown to a pig: 14.5 ± 5.7 ammonites"

Importance = 2. Inconsistent means and standard deviations. If the overall mean is 14.6 ammonites, then the subsets which contribute to the overall mean can't both have means that are smaller than 14.6, but that is what is being reported here: 12.9 and 14.5.

Importance = 1. Also, it is not made clear whether the 4.9 in 14.6 ± 4.9 is a standard deviation or a standard error. By convention, it should be a standard deviation. Then it is also impossible for the standard deviations associated with sacrifice by fire and associated with sacrifice by pig to be larger than the standard deviation for the whole set.

Importance = 2. Further, it is not reported how many sacrifices were by fire and how many by pig. Did some participants not sacrifice chocolate at all, but eat it instead? Perhaps unlikely for trolls, who are clearly not carbon-based life forms and might not derive any nutrition from organic molecules, but that should have been reported.

Importance = 2. Number of boils are reported separately as if some difference had been predicted, but there is no statistical analysis of that difference.

Line 51: "the number of ammonites decreased as the grain size of the rock increased"

Importance = 2. This hypothesis was not mentioned in either the introduction or in the analysis plan. Is this supposed to mean that Nuggan holds trolls made of fine-grained sedimentary rock to a higher standard, or just plain doesn't like them? This doesn't follow from anything that was mentioned. Looks like harking.

Line 51 and Figure 1, importance = 2: Pearson's r is a parametric test, assuming difference or ratio measurements. The figure puts only labels on the x-axis, suggesting that the rocks are ranked by grain size, but that the differences from one type of rock to the next may not be the same. Students should notice that. The appropriate analysis for types of rock ranked by grain size would be a Spearman's rank order correlation.

Figure 1, importance = 2: There is a negative value in the figure. A negative number of boils is impossible. The reviewer will need to see the raw data.

Lines 43 – 52, importance = 2: What happened to the order in which the different flavours of chocolate were sacrificed? Did anyone refuse to sacrifice?

Line 55: "71 dwarves dug tunnels out of their cells during the first night of confinement."

Importance = 1. Between that and the dwarves who changed the locks on their cell doors, it seems all dwarves had a choice whether to stay beyond the first night, but being forced to put in the required work is still not ethical.

Importance = 1. And there is a chance of differential attrition.

Lines 55 – 59: "Of these tunnels, 5 led into the cells of trolls who prematurely dropped out of the experiment. There were signs of struggle…"

Importance = 2. There is an ethical issue, namely whether such conflicts were predictable, whether they should have been prevented (are these fights enjoyed by mutually and freely consenting participants or are they attacks on unwilling victims?), and how they might have been prevented.

Importance = 2. Further, there is the question of what the authors counted as remains. Did they look for gravel or sand? Do trolls have bodily fluids?

Lines 63 – 64: "8 were excluded for not following instructions."

Importance = 4. Which means what? If it means they ate the chocolate, shouldn't that displease Nuggan more than the most inept sacrifice? In that case, this would be an important sacrifice condition. What exactly were the instructions anyway?

Lines 66 – 67, importance = 2: There are six different orders of three blocks of chocolate. What happened to the other four, besides lightest to darkest and darkest to lightest?

Lines 68 – 71: "we found there was no difference between sacrifice by fire and sacrifice by pig if the order of sacrifice progressed from lightest to darkest chocolate ($t(44) = 1.15$, $p > 0.25$), but there was a difference when the order of sacrifice progressed from darkest to lightest chocolate"

Importance = 2. Pairwise comparisons are only justified if either there is an interaction or if there was an a priori hypothesis regarding those pairwise comparisons.

Importance = 2. No interaction is reported here, and the figure makes quite clear that there can't be an interaction. The slopes are much the same, and the absence of a difference between fire and pig in the light to dark order comes from larger standard errors.

Importance = 2. No planned comparison was mentioned.

Importance = 2. The other two pairwise comparisons are not reported, so how was that choice made?

Importance = 2. An absence of a significant effect is not evidence for there being no effect.

Figure 2, importance = 1: The vertical axis has no label.

Importance = 1. The figure legend should contain more information.

Importance = 1. The font has changed compared to Figure 1. A common visual style is encouraged.

Line 73: “134 humans sacrificed chocolate to fire, 132 to pigs, 35 were excluded”

Importance = 2. That now adds up to 301 humans, 2 more than reported in the methods.

Line 78, importance = 2: Inconsistent reporting of statistics. This eta squared is the only effect size being reported, but they should always be reported.

Lines 81 – 82: “men had more boils than women”

Importance = 2. The sex difference is not in the hypotheses or the methods, suggesting harking. It is not shown in the figure, either.

Lines 73 – 82, importance = 2: None of the interactions was mentioned. Does that mean that they are not significant? Not theoretically interesting and never intended to be reported? The introduction is vague on this issue, but ”what manner of sacrifice pleases Nuggan, whether different flavours of chocolate must be sacrificed in a specific order, and whether that depends on the species to which the Faithful belong” does sound like the interactions should be relevant.

Importance = 4. There is no statistical analysis of species differences, even though the research question explicitly mentions them.

Importance = 2. The stated research question is concerned with the actions of the Faithful. There are no data on the faith of those forced into this experiment.

Figure 3, importance = 1: The vertical axis has no label.

Importance = 1. The figure legend should contain more information.

Importance = 1. The font has changed compared to Figure 1. A common visual style is encouraged.

Line 85: “A sacrifice is a sacrifice in name only if the sinner does not suffer for it.”

Importance = 4. That suggests a mediating variable between sacrifice and Nuggan’s displeasure, namely how much the sinner suffers. Does Nuggan want more or less suffering? Do people differ in whether they regret chocolate thrown to a pig more than chocolate thrown into a fire? In that case, the degree of regret should have been reported. Does Nuggan prefer the Faithful to get over the worst part first, or does He prefer them to build up to their greatest regret? To find out, people’s preferences for different kinds of chocolate should have been reported.

Lines 85 – 86: “A sacrifice is a sacrifice in name only if the sinner does not suffer for it. That explains the absence of boils on the igneous and metamorphic trolls”

Importance = 1. There was no statistical analysis of differences between trolls of different geological composition.

Importance = 2. The conclusion not follow from the data in this study, and no reference is being offered for this assertion.

Lines 90 – 91: “(Kabloom & Alvarez, Century of the Fruitbat 04)”

Importance = 1. Not in the list of references.

Line 92: “Sedimentary trolls, however, are different, and provided meaningful results”

Importance = 2. Given the dubious assumption that ammonites are equivalent to boils, and the absence of any ammonite count and measurement before the experimental manipulation, and that assignment may not have been random, that is debatable.

Lines 92 - 93: “Organic inclusions in sedimentary rock make organic matter nutritious.”

Importance = 2. Assumption not supported by either evidence from this study or by a reference.

Lines 93 – 94: “The resulting craving for chocolate makes the sacrifice relevant.”

Importance = 2. The experiment provides no data supporting the conclusion that sedimentary trolls crave chocolate, or that such a craving makes the sacrifice relevant.

Lines 85 – 94, importance = 2: Where is the discussion on the relationship between ammonites and grain size? It popped up in the results despite the lack of any hypothesis, now there is no discussion.

Lines 95 – 96: “In both dwarves and humans, it makes no difference whether chocolate is sacrificed to fire or a pig, provided that the order of sacrifice proceeds from lightest to darkest”

Importance = 2. Repeats the statistically unjustified claim that the absence of statistical significance is the same as there being no difference, based on pairwise comparisons that are not justified by there being an interaction between sacrifice mode and sacrifice order.

Lines 96 – 99: “In contrast, when the order of sacrifice proceeds from darkest to lightest, Nuggan is less displeased when dwarves sacrifice by fire, and humans sacrifice by pig”

Importance = 2. The analysis of that interaction is missing from the results.

Lines 98 – 99: “The species also differ in that only humans show a sex difference.”

Importance = 2. That interaction of species with sex is not only missing from the results, it would be impossible to perform anyway, because all dwarfs in this study reported as male. There are no data on dwarven sex differences, and therefore it is impossible to tell whether the species differ in their patterns of sex differences.

Importance = 2. If species comparisons are being made, why are the trolls not being included? Why was their sex not recorded?

Lines 43 – 99, importance = 2: The size of boils is a measure reported in the methods, but does not show up in either results or discussion.

Lines 114 – 166, importance = 1: Snashfold’s book is not cited in the text.

Overall comments:

The study is deeply flawed and must be rejected. Listing just the most important points, participants were coerced, the motivation for the research question is not clearly stated, methods and discussion imply a mediating variable that is not mentioned in connection with the research question, there are no specific hypotheses derived from the research question, there is no information on participants’ faith, even though all hypotheses that can be inferred concern the Faithful, assignment to one of the treatment conditions was uncontrolled, rather than random, and assignment to the other treatment conditions is not explained. Measurements were not

adequately described. Analyses that should follow even from vaguely implied hypotheses (the interactions of species, sacrifice mode, and sacrifice order) were not reported. Other analyses that are unrelated to the stated research question show up in the results, suggesting that exploratory analyses are passed off as confirmatory. Data on size of boils, mentioned in the methods, completely vanished from the results. Instead, there was a report on the proportion of skin area, a measure not mentioned in the methods. The numbers of participants are inconsistent. Some of the reported results are impossible. The discussion relies on an interaction that was not reported in the results and claims that dwarves and humans differ in that only humans show sex differences, despite all dwarfs self-reporting as male and therefore not providing data on sex differences. It is not clear whether ammonites are equivalent to boils, and so it is not clear whether any of the troll data are even relevant.

Finally, the authors fail to provide a statement on possible conflicts of interest. Specifically, it is not clear whether there is a conflict of interest arising from people working at an Institute *for the Study of Heresy* possibly being tasked with suppressing heresy, or providing research that is useful for that purpose. The title of a book written by one of the authors (*The Hammer of the Heretics: On the Importance of Rooting Out Deviations from the Pure and True Faith*) suggests that there is such a conflict of interest. Any conflict of interest must be clearly stated both in any research submitted for publication *and* in information supplied to participants to make sure that participants' consent is *informed*.

Due to fatal flaws in design, the data reported in this study are worthless. I present the following recommendations for a revised study of the stated research questions:

First, participation must be voluntary. Any fasting must be voluntary, health implications must be considered, and medical personnel must be on standby. Potential participants must be informed of what can be reasonably expected to matter to their decision whether to participate. That includes conflicts of interest on the part of the researchers.

Second, the motivation for the research question must be made clear. Is it based on anecdotes, previous data, or theory? The discussion mentions that "A sacrifice is a sacrifice in name only if the sinner does not suffer for it", suggesting that suffering is a mediating variable. If so, then it is relevant how much each participant suffers from sacrificing each variety of chocolate to either a pig or to fire, and this must be recorded and analysed.

Third, the research question should be translated into specific hypotheses.

Fourth, the research question and some aspects of the analysis suggest that suffering is assumed to mediate between participants' actions and Nuggan's response. This should be made clear, and if this is a mediating variable, the design must be altered accordingly. There must be appropriate controls of people who are not tempted, people who do not sacrifice. Further, the faithful are mentioned in a manner suggesting that his is a relevant variable. If the research question concerns how Nuggan treats His *followers*, then faith must be measured, and those who have other faiths or none should serve as a comparison, taking into account whatever may be known about whom a god may smite.

Fifth, until it is clear whether the proposed measure of Nuggan's displeasure even applies to trolls, they should be excluded.

Sixth, assignment to treatment conditions must be random.

Seventh, authors must report what it means not to follow instructions. Did participants eat the chocolate? That would seem relevant to the research question and should not just be dismissed without giving further information.

Eighth, there should be an analysis plan.

Ninth, there are several analyses that do not clearly relate to the research question and give the impression of hypothesising after the results are known. Authors should clearly state their hypotheses, and ideally preregister them, along with the analysis plan. They should then clearly distinguish between confirmatory and exploratory data analyses.

Tenth, authors must report data that are complete and coherent, instead of incomplete and logically impossible. I strongly recommend that all raw data and analysis scripts should be made available, to the extent that this is consistent with ethical considerations and data protection.

Karakterskala som er benyttet

Bokstavkarakter: https://innsida.ntnu.no/wiki/-/wiki/Norsk/Karakterskalaen