"Uncle Iroh, from Fool to Sage – or Sage All Along?"

Appendix

Eric Schwitzgebel & David Schwitzgebel

Experiment 1

Methods

Participants

All participants were recruited via Prolific (Palan & Schitter, 2018). The title was "Make Judgments about TV Clips (from Avatar: The Last Airbender)." The description follows:

> In this study, you will watch several brief clips from the show Avatar: The Last Airbender, then answer a series of questions about the clips. You are welcome to participate in this study even if you have not seen Avatar: The Last Airbender. You will be compensated regardless.

Several pre-screening filters were applied, limiting eligible participants to those meeting the criteria indicated in **table 1**.

Age	18-25
Nationality	US
Fluent languages	English
Approval rate	98% +
Number of previous submissions	100 +

An initial pilot sample of 20 participants were recruited to ensure the functionality of the experimental design and data collection. These participants were not included in any analyses. Following collection of the initial pilot, 200 participants were gathered in two batches of 100. All participants were compensated $3.33, and participants who were recruited to any of our batches (i.e., pilot, batch 1, batch 2) were automatically excluded from subsequent batches and experiments. All exclusion criteria, planned analyses, and inter-batch procedures are available at the following pre-registration page: https://aspredicted.org/a4kj6.pdf.

Design

The experiment, as it was presented to participants, is available as a *.qsf* (Qualtrics Survey Format) file on Github: https://github.com/dschwitz-PSL/Avatar-Schwitzgebel/tree/Qualtrics. However, note that two substantial changes were made to the survey between our experiments: firstly, in experiment 1, no demographic question was included for age; secondly, in experiment 1, no debriefing question was included for gender. All data were collected internally (i.e., using the Qualtrics data collection features).

Participants were first presented with a simple request for consent. Following their acceptance, participants completed an introductory section in which they self-reported their familiarity with *Avatar: The Last Airbender* on a scale ranging from "not at all familiar" (1) to "very familiar" (7). Next, they completed a six-question multiple choice pretest intended to examine their familiarity with the show. The questions were presented in a random order. All questions had four options, with only one correct option, and the final option always read "I do not know the answer to this question." Incorrect answers (including the final multiple choice option) were coded as 0, and correct answers were coded as 1. Each question concerned a noteworthy detail about the plot, characters, or world of *Avatar: The Last Airbender*. Finally, participants were asked to provide qualitative (i.e., free response) descriptions of Iroh and Katara no longer than 2-3 sentences. However, the questions specified that they could be left blank if the participant was not familiar with the character.

The main section of the experiment consisted of six clips, six semantic differential blocks, two qualitative questions, and two attention checks. Presentation was semi-randomized. Three clips featuring Iroh were presented, and three clips featuring Katara were presented. The

semantic differential questions were presented immediately after the participant completed the corresponding clip, such that one set of questions corresponded to one clip. The semantic differentials included six components, and each component was rated on a seven-point scale (i.e., wise (1), unwise (7)). Each component concerned the perceived wisdom of the character, including (crucially) "wise" and "unwise." Following completion of the semantic differential, the participant was either presented with another clip or asked a qualitative question. The qualitative question asked participants whether the character seemed wise or unwise in the previously-presented scenes, and required a brief (20-80 words) explanation for their answer. The attention checks simply asked participants to provide a one-sentence description of the previous clip.

All three clips and semantic differential questions from a single character (i.e., all clips from Iroh and all clips from Katara) were presented as a sequenced block. In other words, participants were presented with either three clips from Iroh or three clips from Katara, interspersed with three corresponding semantic differential questions. After participants completed all of the clips for a given character, participants were presented with the qualitative question described above (in which they were requested to judge the character's actions from the three prior scenes). The presented order of the three scenes for a given character was randomized. Furthermore, the order of character presentation (i.e., whether the three scenes from Iroh came first, as opposed to Katara) was randomized. Attention checks were randomly presented after completing a clip (before the semantic differential block).

Following completion of the main section, participants completed a brief posttest. This posttest included three optional free-response sections: comments that the participants would like us to consider, technical issues that participants encountered during the study, and guesses or

predictions concerning the purpose of our experiment. After completing this posttest, participants were debriefed and redirected to Prolific.

Results

All data (.csv) and analyses (.R) are available on GitHub: https://github.com/dschwitz-PSL/Avatar-Schwitzgebel/tree/main

Six participants were excluded due to failing the attention checks (i.e., giving no response, an incorrect response, or an irrelevant response). All analyses were performed on the remaining 194 participants. Participants were classified as "naive" or "knowledgeable" based on the criteria described in the preregistration. Of the 194 participants, 122 (62.8866%) were classified as knowledgeable.

Mean participant responses, based on the "wise" semantic differential rating, are illustrated in **table 2.**

Experience (1 = knowledgeable, 0 = naive)	Character	Mean response
0	Iroh	3.08
0	Katara	1.95
1	Iroh	3.02
1	Katara	1.80

Our main analysis was conducted according to the description provided in the preregistration. This analysis failed to find a significant difference between the IMK scores of knowledgeable (M = 1.22, SD = 1.38) and naive (M = 1.13, SD = 1.36) participants ($t(192) = -.46$, $p = .64$).

Our secondary analysis was also conducted according to the description provided in the preregistration. The model successfully converged, indicating that no further adjustments were required (Bates et al., 2015). In accordance with the first analysis, this analysis failed to find the crucial interaction between character (Iroh or Katara) and experience (knowledgable or naive) ($\chi 2(1) = .42$, $p = .52$). The full output of the model is provided in **figure 1** (note that "Q3" corresponds to "participant").

```
Linear mixed model fit by REML ['lmerMod']
Formula: response ~ clip_character * experience + (1 | PID) + (1 | clip)
   Data: all_data

REML criterion at convergence: 3925.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1599 -0.6464 -0.0977  0.4906  4.0392

Random effects:
 Groups   Name        Variance Std.Dev.
 PID      (Intercept) 0.3674   0.6062
 clip     (Intercept) 0.2709   0.5205
 Residual             1.4337   1.1974
Number of obs: 1164, groups:  PID, 194; clip, 6

Fixed effects:
                                Estimate Std. Error t value
(Intercept)                      3.08333    0.31943   9.653
clip_characterKatara            -1.12963    0.44030  -2.566
experience                      -0.06421    0.13664  -0.470
clip_characterKatara:experience -0.09441    0.14529  -0.650

Correlation of Fixed Effects:
            (Intr) clp_cK exprnc
clp_chrctrK -0.689
experience  -0.269  0.110
clp_chrctK:  0.143 -0.208 -0.532


Analysis of Deviance Table (Type III Wald chisquare tests)

Response: response
                          Chisq Df Pr(>Chisq)
(Intercept)             93.1715  1     <2e-16 ***
clip_character           6.5821  1     0.0103 *
experience               0.2208  1     0.6384
clip_character:experience  0.4223  1     0.5158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both of our planned analyses (one t-test and one generalized linear model) failed to identify the crucial effect. Therefore, we cannot conclude that there is any statistically significant interaction between clip character and reported wisdom. Further interpretations of the results, including discussion of the qualitative data, are available in the main report.

Experiment 2

Methods

Participants

All participants were recruited via Prolific (Palan & Schitter, 2018). The title was "Make Judgments about TV Clips." The description follows:

> In this study, you will watch several brief clips from a cartoon, then answer a series of questions about the clips. We will also ask you some questions to assess your familiarity with this cartoon, but you will be compensated regardless of your familiarity.

Several pre-screening filters were applied, limiting eligible participants to those meeting the criteria indicated in **table 3**.

Age	40-100[1]
Nationality	US
Fluent languages	English
Approval rate	95% +
Number of previous submissions	100 +

[1] Due to a data collection error, the ages of six participants were erroneously recorded as 25 years. Although we were unable to identify the precise ages of these six participants, we have verified that they were within range of the prescreened ages.

An initial sample of 20 participants were recruited to ensure the functionality of the experimental design and data collection. These participants were combined with the main sample for subsequent analyses. Following collection of the initial pilot, an additional 60 participants were gathered in one batch. Other than the indicated changes, we retained all procedures indicated in the pre-registration for experiment 1.

Design

The experimental design is largely identical to the design of experiment 1. However, the two features not present in the prior experiment—the demographic questions for gender and age—are implemented in this version of the experiment.

Results

This experiment was not pre-registered, and it was performed post-hoc given the results of the prior experiment. Therefore, the results are purely exploratory and should be considered with caution. All data (.csv) and analyses (.R) are available on GitHub: https://github.com/dschwitz-PSL/Avatar-Schwitzgebel/tree/data

Six participants were excluded due to failing the attention checks (i.e., giving no response, an incorrect response, or an irrelevant response). All analyses were performed on the remaining 74 participants. Participants were classified as "naive" or "knowledgeable" based on the criteria described in the preregistration. Of the 74 participants, 5 (6.76%) were classified as knowledgeable.

Mean participant responses, based on the "wise" semantic differential rating, are illustrated in **table 4.**

Experience (1 = knowledgeable, 0 = naive)	Character	Mean response
0	Iroh	3.00

0	Katara	1.90
1	Iroh	2.40
1	Katara	2.13

Due to the small sample of knowledgeable participants, the mean response cannot be used for within-experiment analyses. However, an independent two-sample t-test did not identify any significant difference in Iroh wisdom ratings between naive participants in experiment 2 ($M = 3.00$, $SD = 1.06$) and naive participants in experiment 1 ($M = 3.08$, $SD = 1.24$), $t(139) = .43$, $p = .67$. Note that there is no significant difference between the mean response for naive participants from this sample (experiment 2) and from the previous participants (experiment 1). Further interpretations of the results, including discussion of the qualitative data, are available in the main report.

References

Bates, Douglas & Kliegl, Reinhold & Vasishth, Shravan & Baayen, Harald. (2015). Parsimonious Mixed Models. *arXiv.* 1506

Palan, Stefan & Schitter, Christian (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*(C), pages 22-27