Hannah Gartz, Darren Skidmore,
Drew Wilson, Lindsay McNary

Effects of Scratch Paper and Background Noise on Figure Weights Test Performance

Introduction

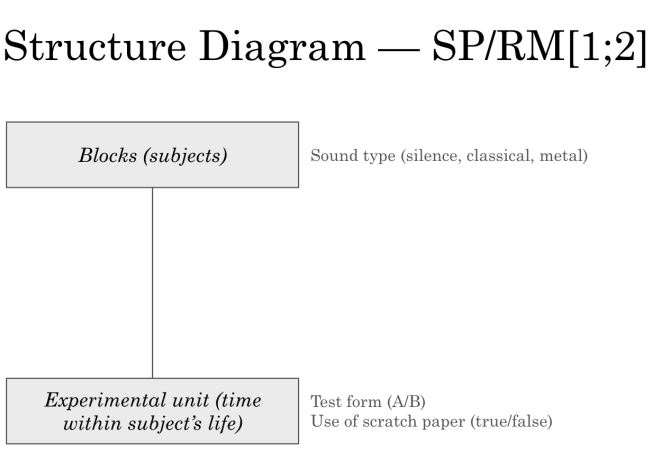
Figure Weights is a puzzle of intellectual rigor initially introduced by Pearson in the *Wechsler Adult Intelligence Scale: Fourth Edition* (WAIS-IV) in 2008. In a Figure Weights puzzle, the examinee is presented with an algebraic system of equations visually represented with at least two balanced weights containing any number and type of two-dimensional polygons. One of the sides in a single balance will be missing (denoted with a question mark) and the examinee must provide one valid answer of exactly five options to satisfy the system of equations.

The Figure Weights task is purported by the test publisher to be an abstract deductive reasoning task which has high correlations to mathematical prowess, fluid reasoning, working memory, and global intelligence. The "official" Figure Weights task on both the WAIS-IV and WISC-V can discriminate between individuals at the 0.1^{st} and 99.9^{th} percentile (σ= -3 to +3). For this experiment, we utilized an "unofficial" version of the Figure Weights task found online which claims to have a ceiling as high as four standard deviations above the mean (roughly one in 10,000 rarity).

We considered several factors which may influence one's performance on Figure Weights, including various types of background noise and if the examinee could use a writing utensil to show their work and reduce working memory demands. We also split the set of twenty-six problems into two groups (A, B) of thirteen problems. We expected that the test score would increase if work was allowed, and we also expected that the test score would decrease as the intensity of noise increased.

Design and Data Collection

We initially used an SP/RM[1;2] design for our data collection. Each subject was the block, and the sound type factor (with levels of listening to 1. classical [Canon in D Major], 2. silence, or 3. Metal [Metallica]) was the between-blocks factor. The experimental unit was a time within each subject's life, and the within-blocks factors were the use of scratch paper factor (with levels of 1. used scratch paper and 2. did not use scratch paper) and the test form factor (with levels of 1. A and 2. B). The hypotheses we were interested in testing included the effect of sound type on test score, the effect of use of scratch paper on test score, and the interaction between sound type and use of scratch paper.



Model: $y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \eta_\ell + (\alpha\gamma)_{ik} + (\alpha\eta)_{i\ell} + (\gamma\eta)_{k\ell} + (\alpha\gamma\eta)_{ik\ell} + \epsilon_{ijk\ell}$

Hypotheses:

(a) $H_0 : \alpha_1 = \alpha_2 = \alpha_3, H_a :$ at least one α_i differs from another
(b) $H_0 : \gamma_1 = \gamma_2 = 0, H_a :$ at least one γ_i differs from another
(c) $H_0 : (\alpha\gamma)_{11} = (\alpha\gamma)_{12} = (\alpha\gamma)_{21} = (\alpha\gamma)_{22} = (\alpha\gamma)_{31} = (\alpha\gamma)_{32} = 0, H_a :$ at least one $(\alpha\gamma)_{ik}$ differs from another

We recruited twenty-four total participants who were all BYU students, within our social and physical proximity, and willing to take the test. The treatments were randomized by typing every combination of the blocks (24 in total) into a Google Sheets spreadsheet. The `RAND()` function was then used to sort each pair. These randomized treatments were then entered into the Google spreadsheet and given to the subjects as they arrived at the testing site. In order to control the environment as much as possible, each subject was tested in a quiet environment—either the examiner's apartment or a private study room in the HBLL.

When each subject arrived, they were given instructions regarding proper administration of the Figure Weights task. They were informed about the purpose of the task and the exact treatment they would be subject to. The subject then completed two practice problems with an extremely low difficulty level, allowing them to become acquainted with the basic premise of the task. After that, the examinee was subjected to their first treatment and given six minutes to solve as many as the thirteen problems they wished. Examinees were allowed to move between questions at will and make guesses, but they were prohibited from receiving aid from the examiner.

During the administration of each treatment the examiner queued the song that the subject was assigned to (except for those whose noise level was Silence) and provided the subject over-the-ear headphones. The proctor ensured that each subject knew if they were allowed to use scratch paper or not on their first test (found on the assignment sheet initially given to each subject). They were specifically instructed not to write anything on the sheet of paper besides their answers if they were not allowed to use scratch paper. As the subject pushed play on their song, the proctor started a timer of six minutes for them to complete the thirteen questions on the form of the test. When the timer finished, the subject stopped the song and the test. They were then instructed to switch to the other form of the test and instructed on whether or not they could use scratch paper on the second part of the test (if they used scratch paper on the first part they would not use it on the second part and vice versa). The subjects restarted the song, and the proctor started another timer of six minutes to complete the second portion of the exam. After each subject was tested, their answers were entered into the spreadsheet used for data collection and the raw scores were automatically calculated.

Data Analysis

We used analysis of variance (ANOVA) to determine if effects between factors were significant in this study. In order for ANOVA to be performed, the six Fisher assumptions (known by its acronym CAZSIN) must be met. CAZSIN constraints were checked using the linear model (`lm`) structure in R.

After the conclusion of the test administration, we noticed some issues in performing ANOVA. The primary complication was that the data did not actually reflect a true SP/RM[1;2] because each block only covered two out of four experimental units. In tandem with Dr. Christensen's feedback, we agreed to drop one factor which we were not experimenting upon (the form) as we assumed that the differences between the two forms were negligible. Our model was now an SP/RM[1;1] in the form:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ijk}$$

with the hypotheses unchanged. The structure diagram only changed with the absence of the test form.

- **C:** Unknown true values are constant. This is true.
- **A**: The components go together to make y_{ijk} by adding them. This is true; a factor diagram can be generated with each factor and observation error adding to produce a single observation.
- **Z**: The errors have a mean of zero. This is true.
- **S:** The errors come from the same distribution, with common standard deviation. This is true.

- **I**: The errors are independent. This is true.
- **N**: The distribution of the errors is normal. This is true.

Because CAZSIN requirements were satisfied, we proceeded to perform ANOVA on the data using the analysis-of-variance (`aov`) structure in R.

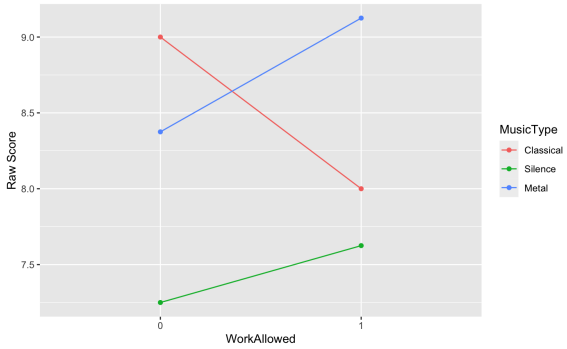
Varying levels of music type did not affect subject performance, nor did varying levels of work allowed or the interaction between music type or work allowed have a significant effect on subject performance. Every p-value was quite high, being greater than or equal to 0.179, indicating that the raw score would remain unchanged in lieu of variations in the level of work allowed or music type.

We also ran a Tukey analysis on the music type to determine if there was a difference between individual music types. After changing α to 0.05 ÷ 3 = 0.0167 through a Tukey correction, we found that there was no significant difference between any of the music types.

Factor	df	SS	MS	F	P
MusicType	2	15.54	7.771	1.866	0.179
Residuals (Subject)	21	87.44	4.164		

Factor	df	SS	MS	F	P
WorkAllowed	1	0.02	0.021	0.011	0.916
MusicType:WorkAllowed	2	6.79	3.396	1.843	0.183
Residuals	21	38.69	1.842		

Comparison	diff	P	Sig?
Silence:Classical	-1.063	0.092	X
Metal:Classical	0.250	0.862	X
Metal:Silence	1.313	0.032	X



Conclusion

Our study examined the effects of scratch paper use and background noise on performance in the Figure Weights task. Contrary to our expectations, neither the type of background noise nor the allowance of scratch paper significantly influenced test scores. There was also no discernible effect observed regarding the interaction between those two factors.

The lack of significant findings suggests that scratch paper use and background noise do not affect performance on deductive reasoning tasks. While this study does not unveil a cause-and-effect relationship between these factors, their interactions, and test performance, it would be reasonable to conclude that such a causal relationship may exist had the experiment's factors been significant. We can assume, however, that these results may generalize to the population of BYU students taking the Figure Weights test under comparable testing conditions.

There are a number of ways to refine the experiment performed in this study in future iterations. We only assigned each subject two of the four possible treatment combinations to make our study more convenient for our volunteer subjects. With more time and ability to incentivize subjects to participate, we could have both recruited a larger pool of subjects, increasing the validity of our results, and assigned each subject all four treatments under a true SP/RM[1;2] design. More subjects would reduce the effect of chance error in our data and allow us to more confidently draw a meaningful conclusion. Assessing a more representative sample of BYU's student body would also provide better generalizability than our study, which was largely composed of students that were convenient to test. Further, while we attempted to standardize environmental factors such as administration procedures, more can be done to control for other potentially confounding variables, such as varying levels of cognitive fatigue throughout the day.

In summary, while this study provides insight into the effects of background noise and scratch paper usage, more thorough testing should be performed before the observed results are confirmed.

Appendix

R Code for Analysis

```
# ANOVA
data = read_csv("https://darrenskidmore.com/stat230/fw-data.csv", show_col_types=FALSE)
data$EID ← as_factor(data$EID)
data$MusicType ← as_factor(data$MusicType)
data$WorkAllowed ← as_factor(data$WorkAllowed)
data$Form ← NULL

anova_model_a1 = aov(RawScore ~ MusicType + Error(EID) + WorkAllowed + MusicType:WorkAllowed,
data=data)

summary(anova_model_a1)

# Residuals
anova_model_rs1 = aov(RawScore ~ MusicType + EID + WorkAllowed + MusicType:WorkAllowed,
data=data)

residuals = anova_model_rs1$residuals

hist(residuals)
sum(residuals)

# Tukey Pairwise Comparison
TukeyHSD(anova_model_rs1, conf.level=0.95)

# Interaction Plot
data_rs1 %>% group_by(WorkAllowed, MusicType) %>%
  summarize(mygroups = mean(RawScore)) → ip_tmp
ip_tmp %>%
  ggplot() +
  aes(x = WorkAllowed, y = mygroups, color = MusicType) +
  geom_line(aes(group = MusicType)) +
  geom_point() +
  ylab("Raw Score")
```

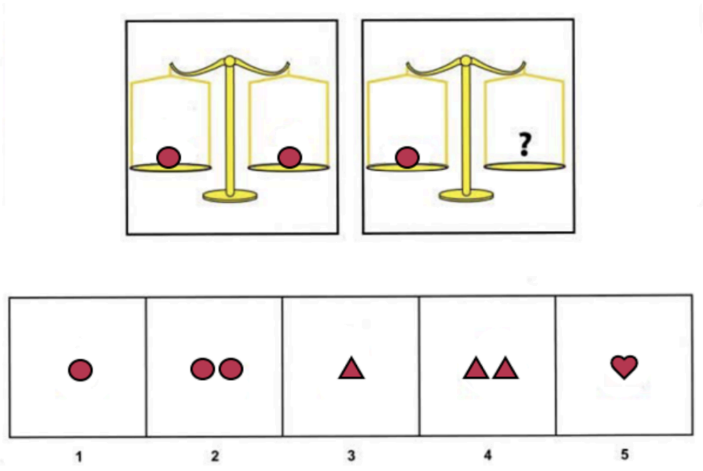
Raw Data

EID	WorkAllowed	MusicType	Form	RawScore
0	1	Classical	A	6
0	0	Classical	B	8
1	1	Classical	B	7
1	0	Classical	A	8
2	1	Silence	A	9
2	0	Silence	B	8
3	0	Metal	A	8
3	1	Metal	B	11
4	1	Metal	B	6
4	0	Metal	A	10
5	0	Metal	B	8
5	1	Metal	A	9
6	0	Classical	B	7
6	1	Classical	A	7
7	0	Metal	B	5
7	1	Metal	A	9
8	0	Classical	A	9
8	1	Classical	B	7
9	0	Silence	B	4
9	1	Silence	A	6
10	1	Metal	A	9
10	0	Metal	B	10
11	1	Silence	A	5
11	0	Silence	B	6
12	0	Classical	A	12
12	1	Classical	B	9
13	1	Classical	B	8
13	0	Classical	A	10
14	0	Metal	A	9
14	1	Metal	B	10
15	0	Silence	A	7
15	1	Silence	B	7
16	0	Silence	B	10

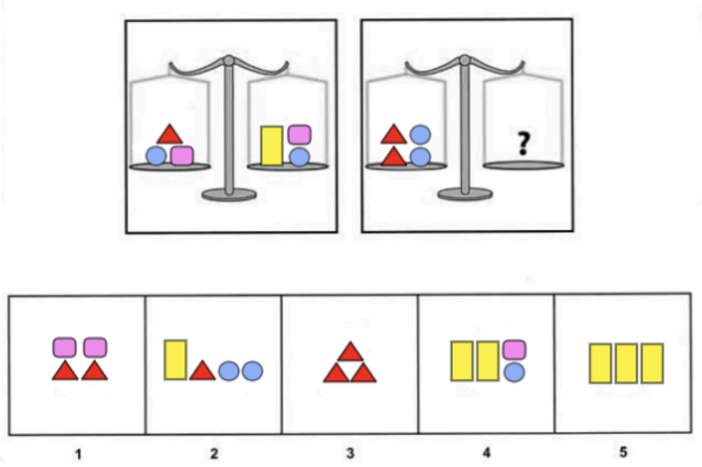
EID	WorkAllowed	MusicType	Form	RawScore
16	1	Silence	A	9
17	0	Silence	A	6
17	1	Silence	B	8
18	1	Silence	B	8
18	0	Silence	A	10
19	1	Metal	A	12
19	0	Metal	B	10
20	1	Classical	A	10
20	0	Classical	B	9
21	1	Silence	A	9
21	0	Silence	B	7
22	1	Metal	B	7
22	0	Metal	A	7
23	0	Classical	B	9
23	1	Classical	A	10

Problem Examples

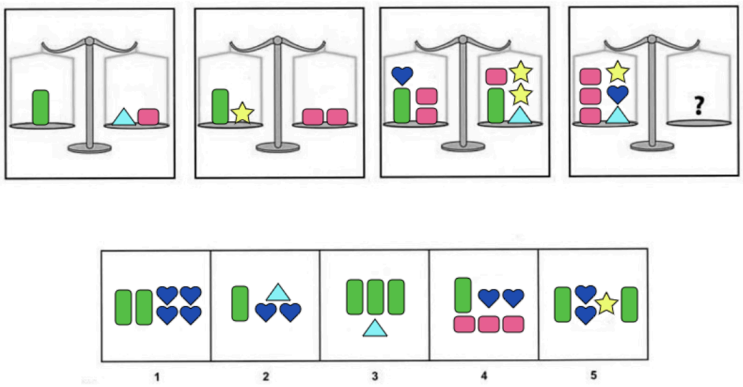
Sample Problem A



Form A Problem 6



Form B Problem 13

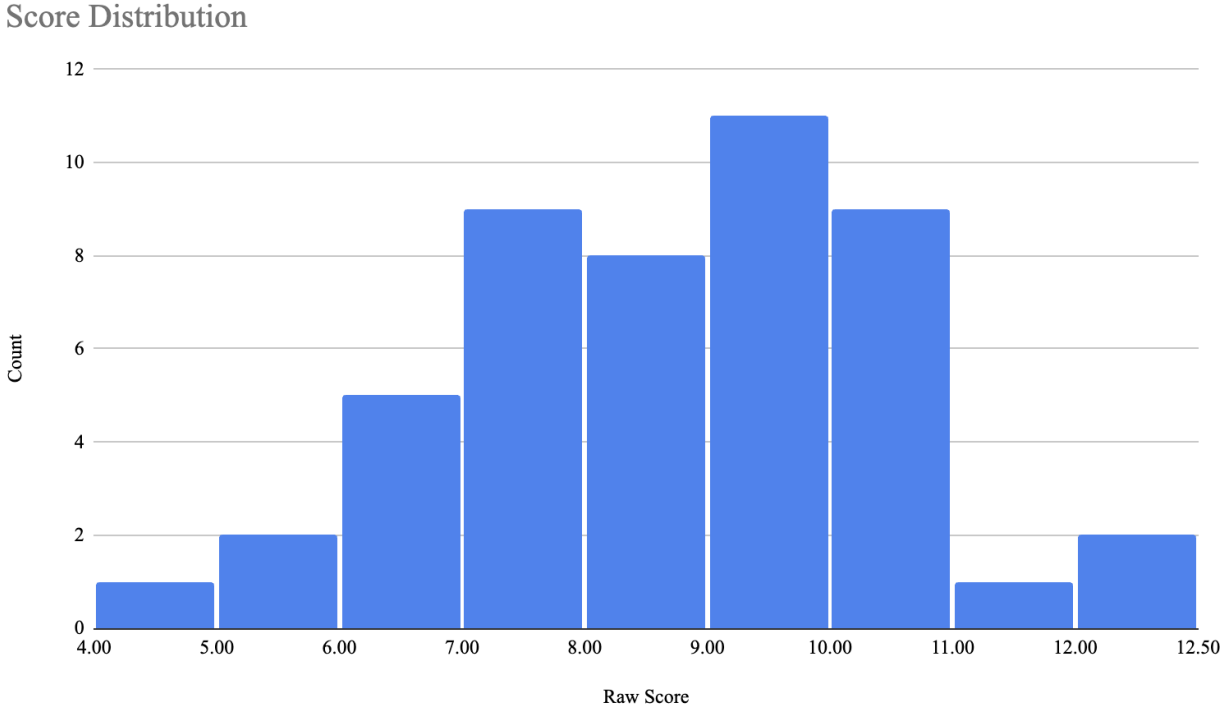


Instructions

- You will be given two or more scales which contain various shapes. One of the weights in the scales will be missing. Your objective is to choose which option balances the scales. This is considered a strong test of mathematical reasoning, having around a 70% correlation with SAT/ACT Math.
- There are thirteen problems in this set, and you will be given a maximum of six minutes.
- The problems will start out very easy and progressively increase in difficulty. Less than 1% of the population can get all the questions correct, so it is not wise to try to finish all the questions with poor accuracy. **Focus on accuracy first, speed second.**
- You are allowed to guess, and incorrect answers will not result in deducted points.

High-Level Analysis

Raw ID	Form ID	Total Correct	% Correct
1	A1	24	100%
2	B1	24	100%
3	A2	24	100%
5	A3	24	100%
8	B4	24	100%
6	B3	23	96%
9	A5	23	96%
13	A7	23	96%
4	B2	21	88%
10	B5	21	88%
14	B7	20	83%
11	A6	19	79%
7	A4	18	75%
12	B6	18	75%
15	A8	18	75%
16	B8	18	75%
17	A9	13	54%
18	B9	8	33%
20	B10	7	29%
19	A10	6	25%
23	A12	6	25%
21	A11	5	21%
22	B11	3	13%
25	A13	3	13%
24	B12	1	4%
26	B13	1	4%


Score Distribution
12
10
8
6
4
2
0
Count
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00
12.00
12.50
Raw Score