Movie Rating Replication

Paola Calle Mario Tan

Introduction

This paper replicates *Strikingly Low Agreement in the Appraisal of Motion Pictures* [WW17] using the Data Analysis Project 1 dataset of 1,097 participants rating 400 films, alongside demographic and viewing-preference variables. A conservative significance threshold of $\alpha = 0.005$ was applied to control false positives.

Exploratory Data Analysis

What Are Our Sample Characteristics? We analyze ratings from **1,097** participants on **400** films. Ratings are on a $[0, 4]$ scale (half-star increments), with missing values where a movie was not seen. The dataset also includes demographics and viewing-preference items (e.g., gender, only-child status, social-watching preference).

Missingness and inclusion. We restrict analyses to valid numeric ratings and, where pairwise statistics are computed, to movie pairs jointly rated by both participants (threshold detailed below). Demographic fields with -1 (no response) are treated as missing.

Rating distribution. Figure 1 visualized the demographics of the respondents and the overall rating behavior. The sample is predominantly male (75%), followed by female (24%), with a small fraction identifying as non-binary ($\approx 1\%$). The average movie rating clusters around 3 stars out of 4, indicating a generally positive evaluation. Most of the participants are not only children ($\approx 81\%$), and a slight majority $\approx 56\%$) report that movies are best enjoyed alone rather than socially. Together, these patterns show a well-diversified sample with moderate variability in viewing preferences and consistent, moderately high average ratings across respondents.

Individual Rating Spearman Correlation. Across 577,691 participant pairs, the average Spearman correlation between rating vectors was $\rho = 0.109$, indicating very weak similarity in movie appraisals Figure 2 and Table 2. On average, the participants shared ratings for 42.8 movies, reflecting a limited overlap in the titles viewed Figure 2 and Table 1. These results replicate the finding of original paper [WW17] that inter-subjective agreement in film evaluation is low, underscoring the individualized nature of movie enjoyment.

Results

Q1: Are movies that are more popular rated higher than movies that are less popular? We assume approximate independence between the movie ratings, given the low average pairwise correlation. A median split on popularity (number of ratings) was used to classify movies into *high-* and *low-popularity* groups because it offers a way to divide the dataset into two equal parts, allowing for comparison of average ratings between more and less popular movies. Let H_0**:** $\mu_{\text{high}} = \mu_{\text{low}}$ and H_1**:** $\mu_{\text{high}} > \mu_{\text{low}}$. Because the group variances differed and the distributions were approximately normal as seen in Figure 3 and Table 3, a Welch's t-test was applied, yielding $df = 378.536$, $t = -17.765$, $p = 0.000$, and $p < 0.005$. We therefore drop H_0 and conclude that more popular movies receive significantly higher average ratings.

Q2: Are movies that are newer rated differently than movies that are older? Because the distributions of average ratings for *new* and *old* movies appear approximately normal and their variances are similar ($var_{new}/var_{old} \approx 1.09$), as shown in Table 4 and Figure 4, a Student's t-test assuming equal variances was applied. Let H_0**:** $\mu_{\text{new}} = \mu_{\text{old}}$ and H_1**:** $\mu_{\text{new}} \neq \mu_{\text{old}}$. The t-test yielded $t = 1.605$, $df = 398$, and $p = 0.1092$. Since $p > 0.005$, we cannot drop H_0, indicating that there is no significant difference in mean ratings; thus, the slight mean increase for new movies (2.66 vs 2.61) is probably due to chance.

Q3: Is enjoyment of *Shrek (2001)* rated differently between male and female viewers? Because the distributions of *Shrek* ratings for females and males are not normal (Figure 5) and their sample sizes differ considerably (Table 5), we apply a Mann–Whitney U test, treating the rating variable as ordinal categorical. Ratings are treated as ordinal rather than continuous because they represent ranked levels of enjoyment–preserving order but not assuming equal intervals between values. We exclude the non-binary group due to its small sample size (Table 5). Let H_0 denote that the distribution of *Shrek* ratings for females and males is the same, and H_1 denote that the distributions differ. The U test yields $U = 96{,}830.5$ with a p-value of 0.0505. Since $p > 0.005$, we cannot drop H_0, indicating that there is no statistically significant difference in the distributions of *Shrek* between female and male respondents.

Q4: What proportion of movies are rated differently by male and female viewers? As in the Q3 analysis, the distributions of movie ratings deviate from normality, and the rating variable is treated as ordinal (see Figure 6). Accordingly, we apply the Mann–Whitney U test to each movie to determine whether the rating distributions differ significantly between male and female viewers. We then calculate the proportion of movies with statistically significant results ($p < 0.005$) to estimate the overall prevalence of gender-based differences in ratings. We found that $\approx 12.5\%$ of movies exhibit significant differences, as shown in Figures 7a and 7b.

Q5: Do people who are only children enjoy *The Lion King* (1994) more than people with siblings? Following the same logic as Q3–Q4, we apply the Mann–Whitney U test since the rating distributions are non-normal and the rating variable is ordinal (see Table 7 and Figure 8). Let H_0 denote that the distributions of *The Lion King* ratings for only children and non-only children are the same, and H_1 denote that only children tend to give higher ratings than non-only children. The test yields $U = 52{,}929.0$ with a p-value of 0.9784. Since $p > 0.005$, we cannot drop H_0, indicating no statistically significant difference in *The Lion King* ratings between only children and those with siblings.

Q6: What proportion of movies show an "only child effect," meaning ratings differ between only children and those with siblings? Following the same procedure as Q4, we apply a two-sided Mann–Whitney U test to each movie to assess whether rating distributions differ across the two groups. Only 1.75% of movies exhibit statistically significant differences, suggesting that the "only child effect" is minimal overall (see Figures 11, 10, and 9, and Table 7).

Q7: Do people who prefer watching movies socially rate *The Wolf of Wall Street* (2013) higher than those who prefer watching alone? Following the same reasoning as in Q3–Q6, and based on the summary statistics in Table 8 and distributions in Figure 12, we apply a one-tailed Mann–Whitney U test. Let H_0 denote that the rating distributions are the same between the two groups, and H_1 denote that the distribution of ratings is higher among those who prefer watching socially. The test yields $U = 56{,}806.5$ with a p-value of 0.0564. Since $p > 0.005$, we cannot drop H_0, indicating no statistically significant difference in ratings between viewers who prefer watching alone and those who prefer watching socially.

Q8: What proportion of movies show an "social-watching" effect, where ratings differ between viewers who enjoy watching alone versus socially? Using the same approach as in Q4 and Q6 and based on the summary statistics and distributions in Table 9 and Figure 13, we apply the Mann–Whitney U test to each movie to assess rating differences by viewing preference. Only

2.5% of movies exhibit statistically significant differences, suggesting that social versus solitary viewing preferences have minimal influence on movie ratings overall (see Figure 14).

Q9: Is the ratings distribution of *Home Alone* (1990) different from that of *Finding Nemo* (2003)? Let H_0 denote that the rating distributions of the two movies are the same, and H_1 denote that they differ. Since the rating variable is ordinal, non-normal, and follow roughly similar shape as seen in Figure 15, we apply a two-sided Mann–Whitney U test to see if there are any distbrutional differenes even if the means are the same (see Table 10). The test yields $U = 510{,}860.0$ with a p-value of 8.81×10^{-12}. As $p < 0.005$, we drop H_0, indicating a statistically significant difference in rating distributions. Viewers tend to assign different ratings to *Finding Nemo* (2003) than to *Home Alone* (1990).

Q10: There are ratings on movies from several franchises in this dataset. How many of these are of inconsistent quality? We test whether movie ratings differ within and across eight major film franchises (*Star Wars*, *Harry Potter*, *The Matrix*, *Indiana Jones*, *Jurassic Park*, *Pirates of the Caribbean*, *Toy Story*, and *Batman*). Because ratings are ordinal and non-normally distributed (see Figure 16 and Table 11), we use the Kruskal–Wallis H test, a nonparametric alternative to one-way ANOVA, to compare rating distributions. Across all franchises, the test yields $H = 965.56$ and $p = 3.32 \times 10^{-204}$, leading us to **drop** H_0 that all share the same distribution and **accept** H_1 that there is a difference in distribution. We then perform separate Kruskal–Wallis tests within each franchise to assess consistency. Seven of the eight franchises show significant within-franchise variation ($p < 0.005$), indicating inconsistent perceived quality (see Table 12). Only the *Harry Potter* series maintains uniform ratings across its films.

QE: Are emotionally immersive viewers more variable in their ratings than less immersive viewers? We computed each viewer's mean emotional engagement score by averaging ten self-report immersion items and measured their rating variability across all rated movies. To examine whether engagement relates to rating consistency, we first tested for a monotonic relationship using a Spearman rank correlation, which was near zero ($\rho = -0.014$, $p = 0.639$), indicating no monotonic relationship. Because both variables exhibited non-normal distributions (Figure 17), we compared high- and low-engagement groups using a Kolmogorov–Smirnov test, which yielded $D = 0.045$, $p = 0.634$. Since $p > 0.005$, we did not drop H_0, suggesting no significant difference in rating variability between high- and low-engagement viewers. These findings suggest that average emotional immersion is not meaningfully associated with the consistency of movie viewers' ratings.

Conclusion

Across all analyses, viewer ratings are largely consistent across groups, with minimal gender, sibling, or social-watching effects. Popular films tend to earn higher ratings, while movie age shows no influence. However, seven of eight major franchises display significant within-series variation, revealing that **inconsistent quality across installments is common**, with *Harry Potter* as the notable exception.

All analysis scripts and code used to generate these results are available online.[1]

[1] See GitHub repository and the preliminary draft on Google Colab.

Supporting Figures

EDA

Figure 1: Counts and distributions for gender, average rating, only-child status, and watching preference per respondent.

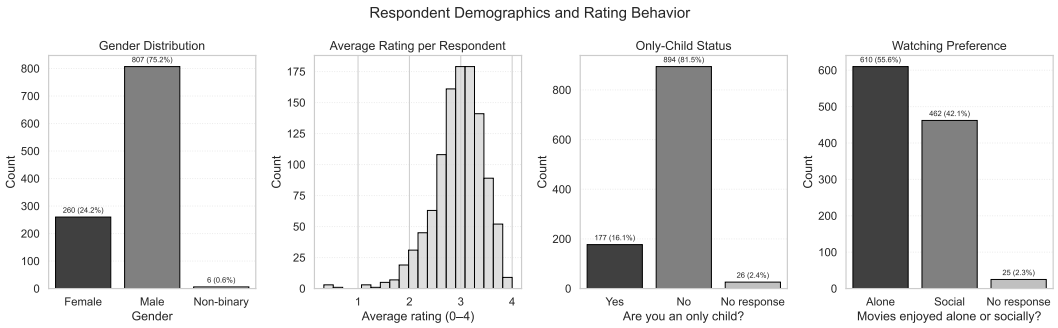
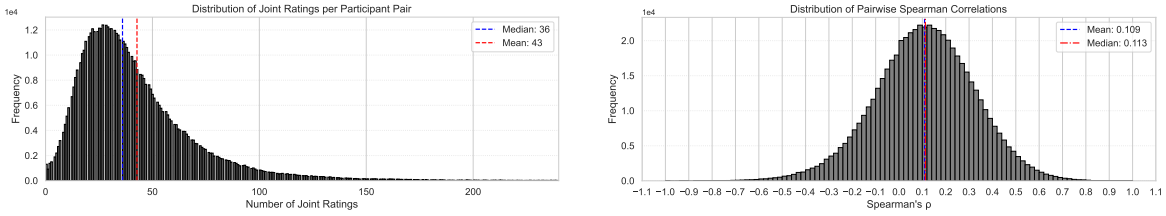


Figure 2: Distributions of overlap (a) and agreement (b) across respondent pairs.

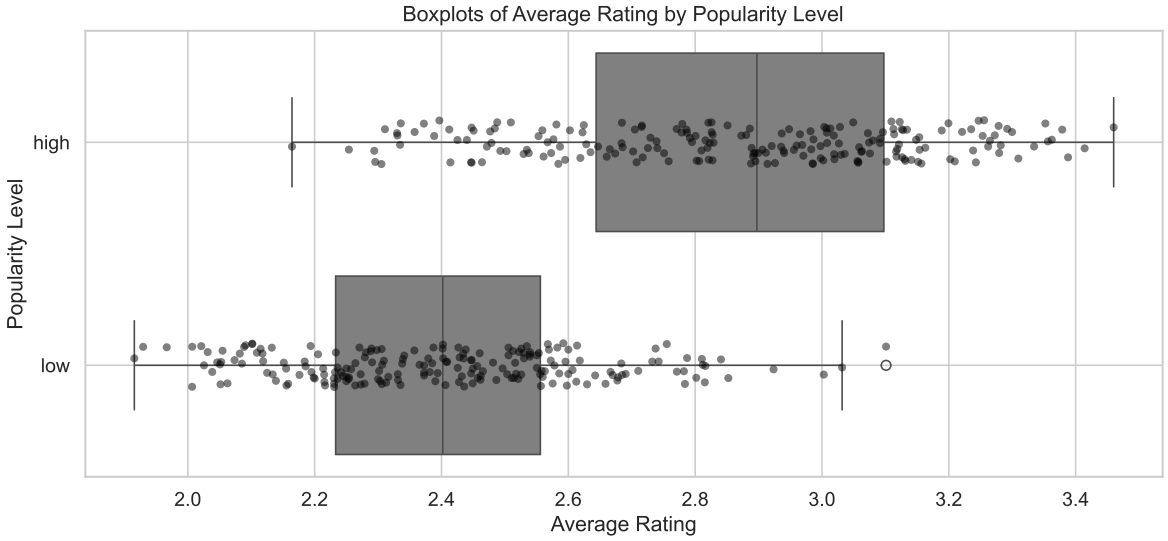


(a) Joint ratings per participant pair

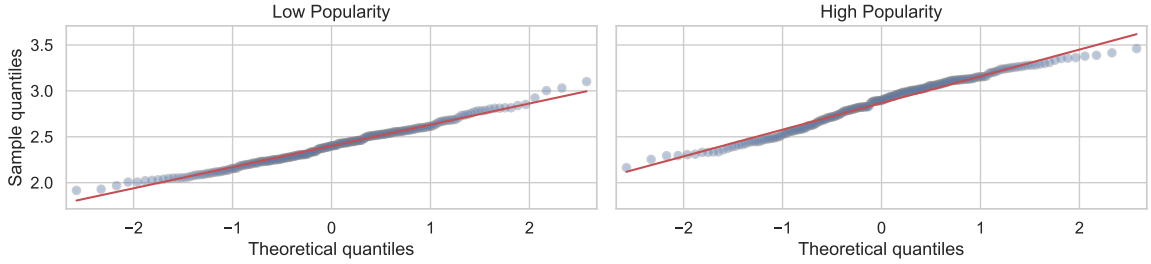
(b) Pairwise Spearman correlations

Results

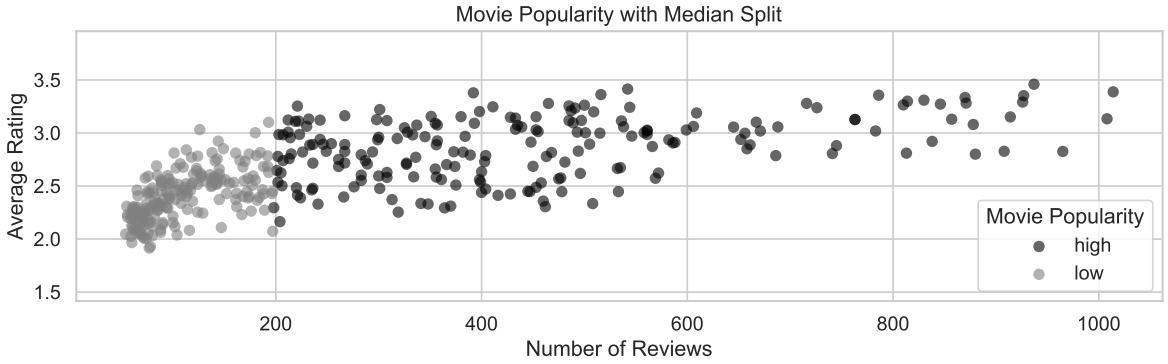
Figure 3: Visual analyses of movie popularity and rating behavior: (a) distribution of average ratings by popularity level, (b) normality assessment for low- and high-popularity groups, and (c) relationship between average rating and review count. Overall, high-popularity movies tend to receive slightly higher ratings. Both distributions are approximately normal, though their variances are heterogeneous.



(a) Boxplots of average movie ratings by popularity level (high vs. low).

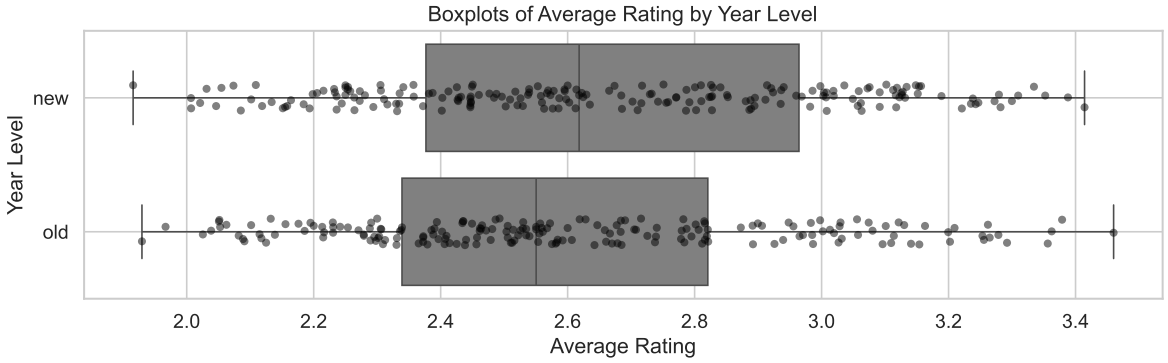


(b) Q–Q plots comparing rating distributions for low- and high-popularity movies.

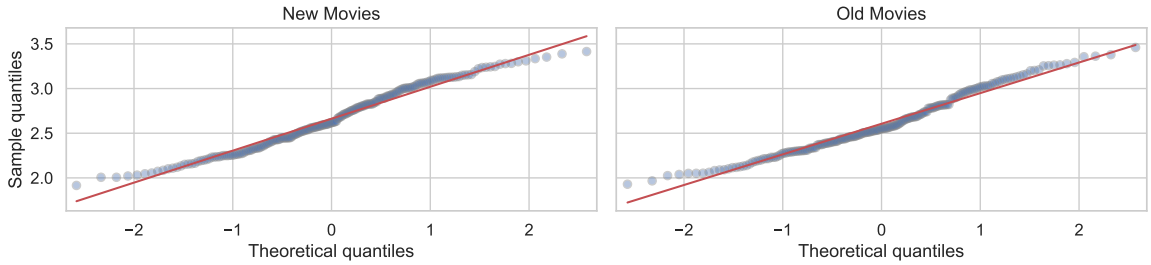


(c) Scatterplot of average rating versus number of reviews, with median-split grouping.

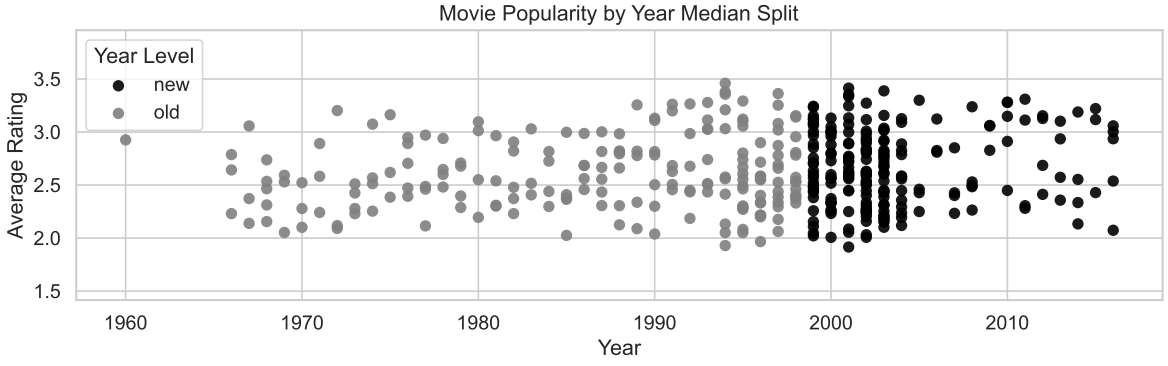
Figure 4: Visual analyses of movie popularity and rating behavior: (a) boxplots of average ratings by year level (new vs. old), (b) Q–Q plots assessing normality for new and old movie groups, and (c) scatterplot of average ratings versus release year with a median-split grouping. Overall, newer movies tend to receive slightly higher ratings, though both distributions appear approximately normal with slightly heterogeneous variances.



(a) Boxplots of average movie ratings by popularity level (high vs. low).

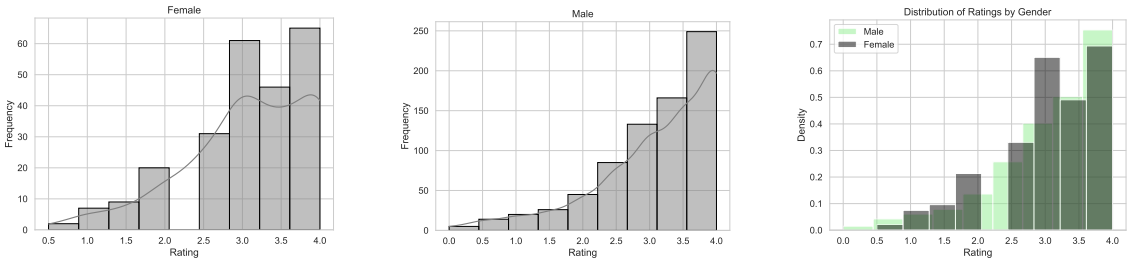


(b) Q–Q plots comparing rating distributions for low- and high-popularity movies.



(c) Scatterplot of average rating versus number of reviews, with median-split grouping.

Figure 5: Visual comparison of *Shrek* rating distributions across gender identities.



(a) Female **(b)** Male **(c)** Avg. rating density

Figure 6: Ratings by gender identity. Although males submitted more ratings overall (left), both genders exhibit comparable relative rating patterns (right), with slightly higher proportions of high ratings (3.5-4.0) in both groups.

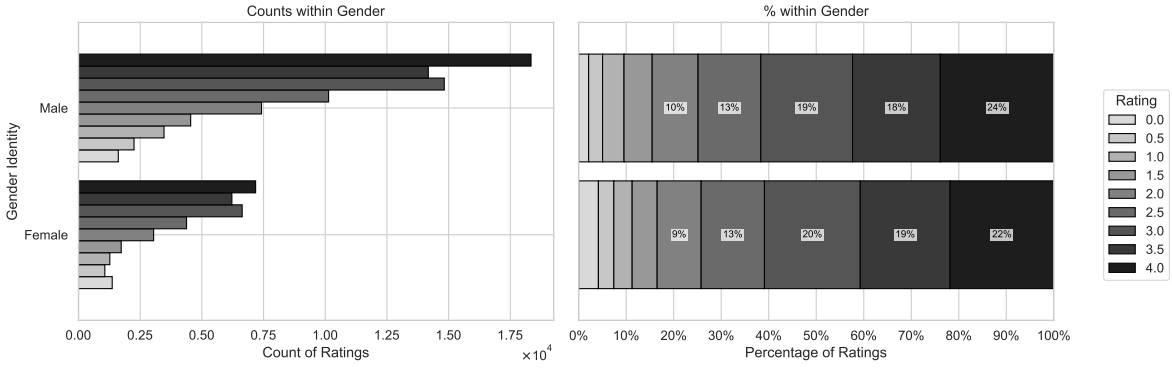
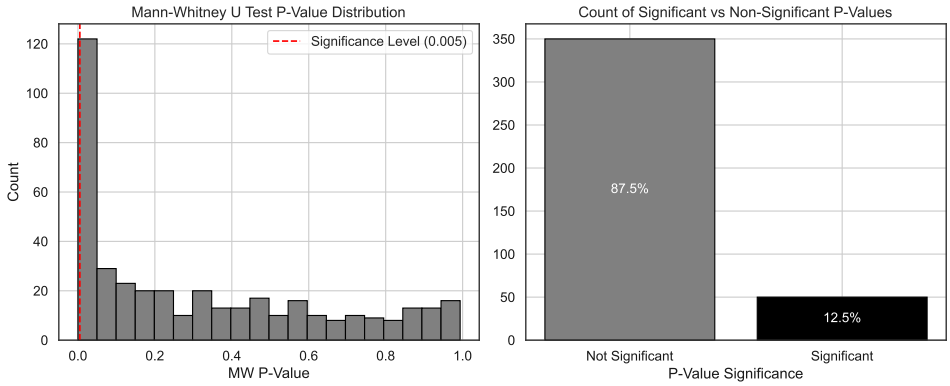
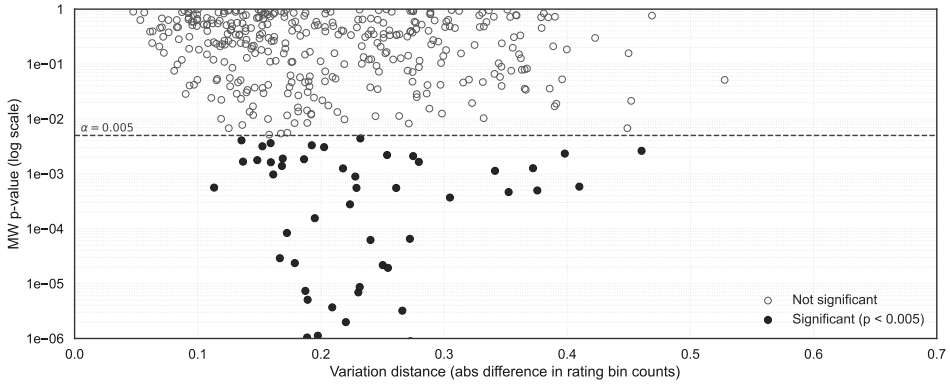


Figure 7: Gender differences in rating distributions. Panel (a) summarizes the distribution and frequency of Mann–Whitney test results across movies; panel (b) relates effect size (variation distance) to statistical significance.



(a) Mann-Whitney p-value diagnostics. Left: histogram of p-values for male vs. female comparisons across movies with the dashed line at $\alpha = 0.005$. Right: counts of significant vs. non-significant tests.



(b) Effect size vs. significance. Each point is a movie; the x-axis is the variation distance between gender-specific rating distributions, and the y-axis is the Mann–Whitney p-value (log scale). Filled markers indicate $p < 0.005$; the dashed line marks the threshold.

Figure 8: Show that both only-child and non-only-child viewers rate The Lion King highly, though non-only children exhibit a slightly greater concentration near the top rating, as seen in the density comparison.

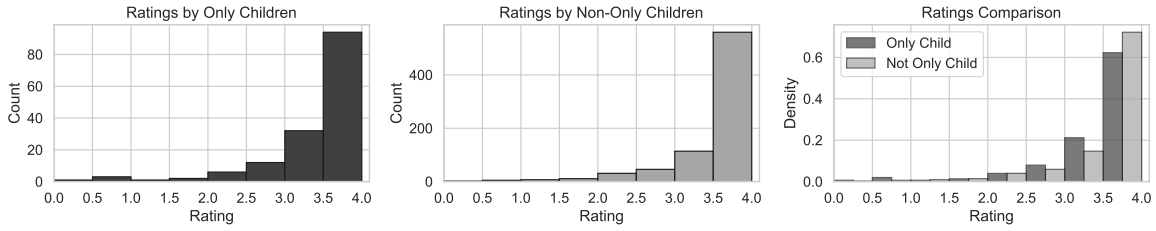


Figure 9: Ratings by only-child status and distributional differences. The top panel shows the count of movie ratings given by only children and those with siblings. Ratings of 3–4 are the most common for both groups, though non-only children submit far more ratings overall. The middle panel plots the cumulative distribution functions of the difference in rating proportions for each score, comparing only-child versus non-only-child viewers. Most CDFs cluster tightly around zero, indicating that rating distributions across groups are largely similar. The bottom panel displays the overall probability mass function of these differences. It is centered near 0 and approximately symmetric, confirming that no systematic only-child effect is evident in the aggregated data.

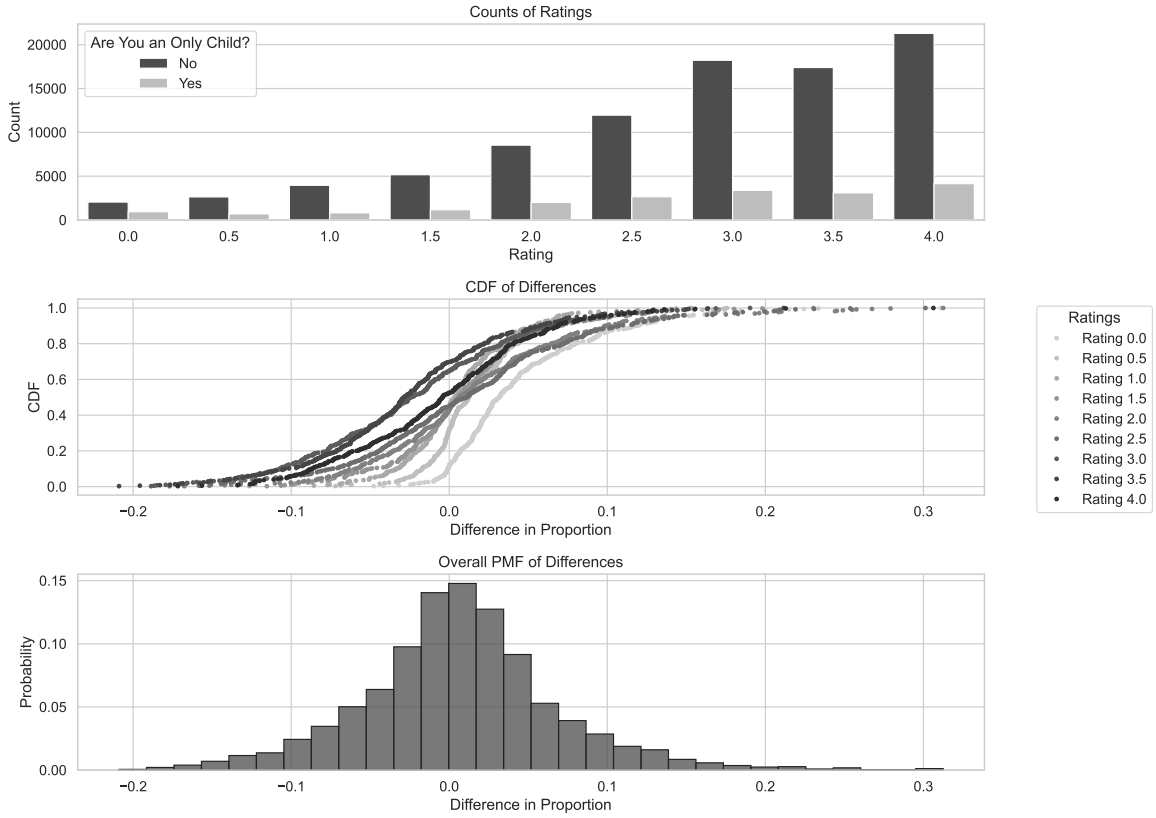


Figure 10: Distributions of movie ratings by only-child status. While rating frequencies differ between groups, the relative percentages across rating levels remain largely similar.

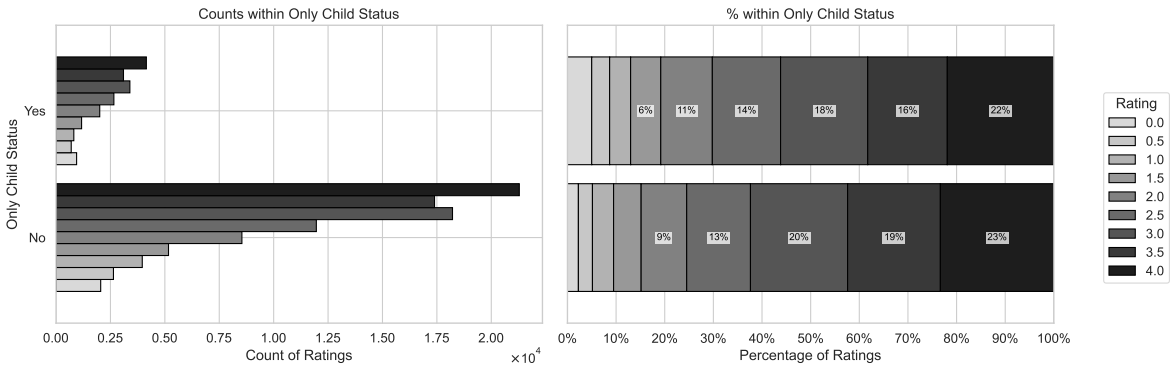


Figure 11: Distribution and significance of Mann–Whitney U test results comparing ratings from only children and non-only children. The left panel shows the distribution of p-values across all movies, with a red dashed line marking the $\alpha = 0.005$ significance threshold. The right panel summarizes the results: only 1.8% of movies show statistically significant differences, indicating minimal evidence of an "only-child effect."

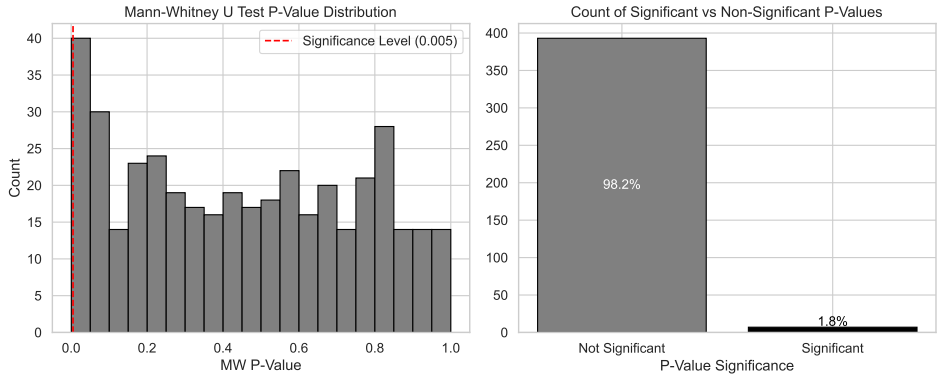


Figure 12: Distributions of *The Wolf of Wall Street* (2013) ratings by viewing preference. Left: viewers who prefer watching alone; middle: viewers who prefer watching socially; right: overlaid densities. Both groups are skewed toward high ratings (3–4) with similar shapes, indicating little evidence of a pronounced "social-watching" effect for this title.

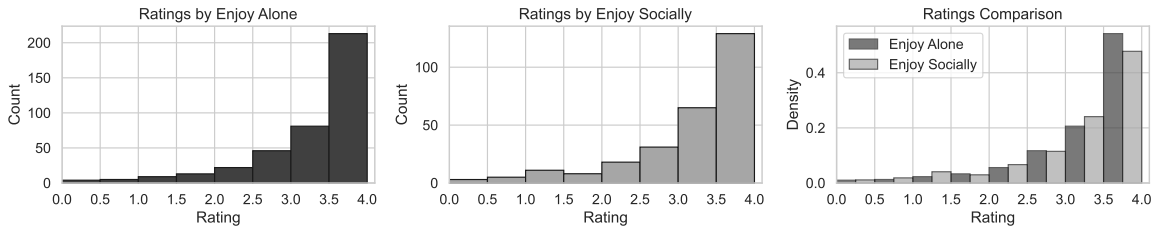


Figure 13: Both groups show similar rating patterns, with most ratings between 3 and 4, indicating comparable enjoyment regardless of watching alone or socially.

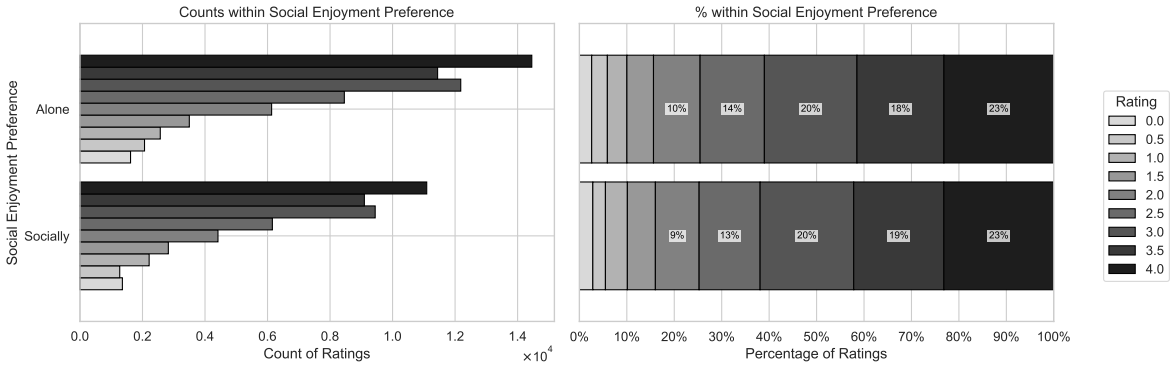


Figure 14: Distribution and significance of Mann–Whitney U test results comparing ratings between viewers who prefer watching movies alone versus socially. The left panel shows the distribution of p-values across all movies, with the red dashed line marking the $\alpha = 0.005$ significance threshold. The right panel summarizes the results: only 2.5% of movies show statistically significant differences, suggesting minimal evidence of a "social-watching" effect.

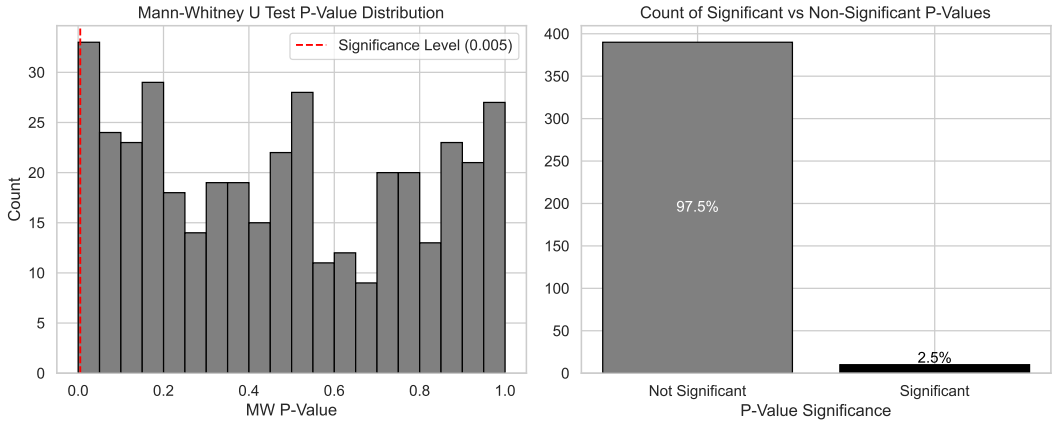
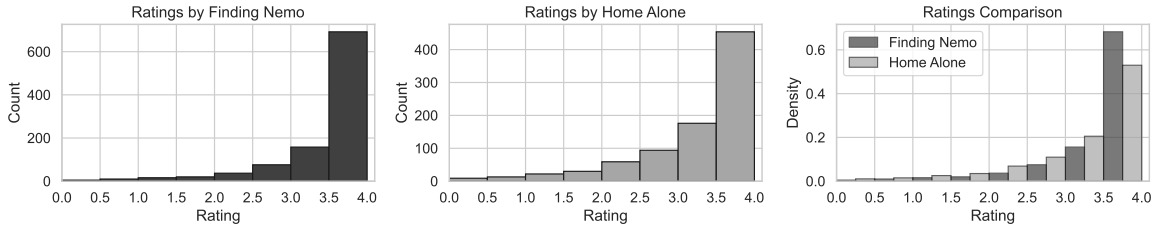


Figure 15: Distributions of ratings for *Finding Nemo* (2003) and *Home Alone* (1990). Both films show right-skewed rating patterns with most viewers assigning ratings of 3.5 or 4.0, though *Finding Nemo* displays slightly higher density at the top end, suggesting marginally greater audience favorability.



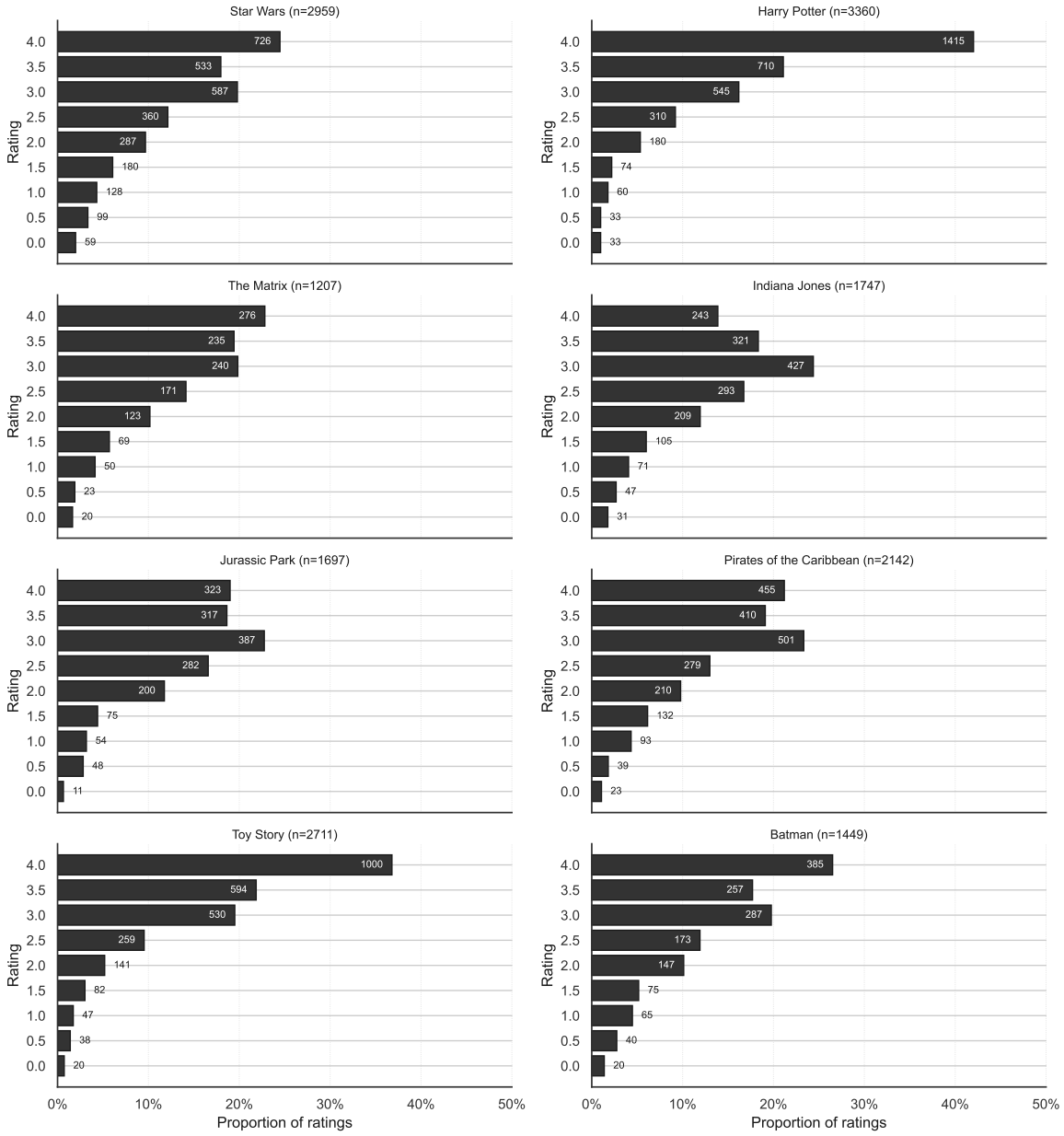


Figure 16: Distributions of viewer ratings across major movie franchises. Each panel shows the proportion of ratings for individual score levels (0–4), with count annotations on each bar. While most franchises exhibit a central tendency around 3.0–4.0, some franchises display greater dispersion.

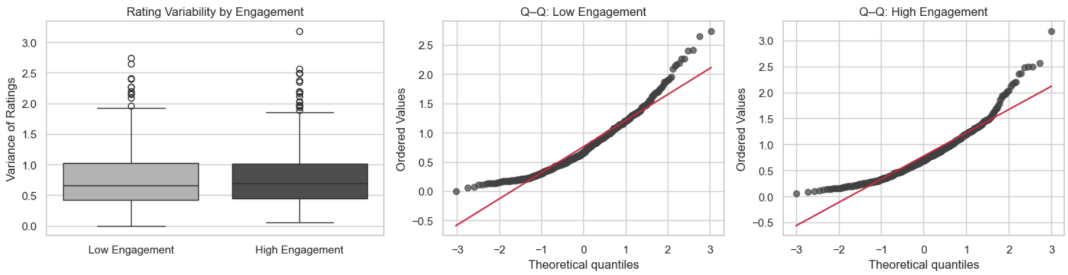


Figure 17: Rating variability by emotional engagement.
The boxplot (left) compares per-viewer rating variance across low and high engagement groups, while the Q–Q plots (right) illustrate non-normal, positively skewed distributions for each group. Both groups show similar distribution shapes, and statistical tests (Spearman $\rho = -0.014$, Mann–Whitney $U = 142{,}685.50$, Kolmogorov-Smirnov $D = 0.045$, all $p > 0.5$) indicate that emotional engagement is not associated with greater variability in movie ratings. Mean rating variance was 0.76 for low-immersion and 0.78 for high-immersion viewers.

Supporting Tables

EDA

Table 1: Summary statistics for the distribution of joint ratings per participant pair.

Statistic	Value	Interpretation
Mean	42.77	Average number of jointly rated movies
Standard deviation	31.65	Substantial variability across pairs
Minimum	0	No shared ratings between some pairs
25^{th} percentile	23	Lower quartile of overlap
Median (50^{th})	36	Typical number of shared ratings
75^{th} percentile	53	Upper quartile of overlap
Maximum	400	Complete overlap in all movies rated

Table 2: Summary statistics of pairwise Spearman correlations between participants.

Statistic	Value	Interpretation
Number of pairs	577,691	Valid participant pairs analyzed
Mean ρ	0.109	Average pairwise correlation (low positive agreement)
Standard deviation of ρ	0.213	Moderate variation across pairs
Minimum ρ	-1.000	Perfect inverse correlation (rare)
25^{th} percentile	-0.027	Lower quartile of correlation values
Median ρ	0.113	Typical pairwise agreement
75^{th} percentile	0.252	Upper quartile of correlation values
Maximum ρ	1.000	Perfect agreement (rare)
Mean p-value	0.389	Average test significance across pairs

Results

Table 3: Summary statistics of average movie ratings by popularity level (median split).

Popularity	Count	Mean	Median	Var	Std
High	200	2.868	2.897	0.085	0.292
Low	200	2.401	2.402	0.054	0.231

Table 4: Summary statistics of average movie ratings by year level (median split).

Year	Count	Mean	Median	Var	Std
New	203	2.662	2.618	0.129	0.358
Old	197	2.606	2.550	0.118	0.344

Table 5: Summary statistics of average movie ratings by gender identity.

Gender Identity	Count	Mean	Median	Var	Std
Female	241	3.083	3.00	0.681	0.825
Male	743	3.155	3.50	0.822	0.907
Non-binary	6	3.250	3.25	0.275	0.524

Table 6: Summary statistics of average movie ratings by only-child status for Lion King.

Only Child?	count	mean	median	var	std
No response	10	3.450	4.0	0.858	0.926
No	776	3.482	4.0	0.516	0.718
Yes	151	3.348	3.5	0.667	0.816

Table 7: Summary statistics of movie ratings by only-child status.

Only Child?	count	mean	median	var	std
No	91,220	2.858	3.0	1.072	1.035
Yes	18,942	2.704	3.0	1.305	1.142

Table 8: Summary statistics of movie ratings by respondents that enjoy *The Wolf of Wall Street (2013)* alone or socially.

Enjoy alone?	count	mean	median	var	std
No Response	4	3.625	3.75	0.230	0.479
No	270	3.033	3.00	0.848	0.921
Yes	393	3.144	3.50	0.757	0.870

Table 9: Summary statistics of movie ratings by respondents that enjoy alone or socially.

Enjoy alone?	count	mean	median	var	std
No Response	1,926	2.670	3.0	1.145	1.070
No	47,874	2.840	3.0	1.116	1.057
Yes	62,414	2.830	3.0	1.112	1.054

Table 10: Summary statistics of ratings for *Home Alone* (1990) vs. *Finding Nemo* (2003).

Movie	count	mean	median	var	std
Finding Nemo (2003)	1,014	3.388	3.5	0.621	0.788
Home Alone (1990)	857	3.130	3.5	0.827	0.909

Table 11: Summary statistics of viewer ratings across major movie franchises.

Franchise	count	mean	median	var	std
Harry Potter	3,360	3.304	3.5	0.729	0.854
Star Wars	2,959	2.856	3.0	1.096	1.047
Toy Story	2,711	3.241	3.5	0.733	0.856
Pirates of the Caribbean	2,142	2.888	3.0	0.901	0.949
Indiana Jones	1,747	2.736	3.0	0.903	0.950
Jurassic Park	1,697	2.863	3.0	0.840	0.916
Batman	1,449	2.915	3.0	1.028	1.014
The Matrix	1,207	2.887	3.0	0.965	0.983

Table 12: Kruskal–Wallis test results on viewer ratings by franchise. Franchises with $p < 0.005$ are labeled *inconsistent*.

Franchise	H	p-val	Films	Ratings	Inconsistent
Star Wars	230.6	8.0×10^{-48}	6	2,959	Yes
Batman	190.5	4.2×10^{-42}	3	1,449	Yes
The Matrix	48.4	3.1×10^{-11}	3	1,207	Yes
Jurassic Park	46.6	7.6×10^{-11}	3	1,697	Yes
Indiana Jones	45.8	6.3×10^{-10}	4	1,747	Yes
Toy Story	24.4	5.1×10^{-6}	3	2,711	Yes
Pirates of the Caribbean	20.6	3.3×10^{-5}	3	2,142	Yes
Harry Potter	3.3	3.4×10^{-1}	4	3,360	No

References

[WW17] Pascal Wallisch and Jake Alden Whritner. Strikingly low agreement in the appraisal of motion pictures. *Projections: The Journal for Movies and Mind*, 11(1):102–120, 2017.