"Measuring inter-subjective agreement on units and attributions in comics with annotation experiments": Annotator competence and reliability

1 Introduction

This document is supplementary material for the article "Measuring inter-subjective agreement on units and attributions in comics with annotation experiments", and describes agreement between each annotator pair, per condition, in more detail than what appears in the main article. The purpose of the study described in the main article is to assess which scale - binary, ordinal or continuous - works best for annotating semantic information within individual panels in comics stories. The overall experiment tests inter-annotator agreement using a between-subjects approach where a unique set of 10 participants annotate a full comic story using either a binary, ordinal or continuous scale. Each scale is tested on two comics stories, for a total of 6 studies.

This document describes qualitative assessments about the extent to which participants tend to agree or disagree with one another, and whether there are annotators that are unreliable - that is, whether they did the annotations earnestly and with an plausible level of competence. To assess competence, a heatmap of Pearson's r scores per annotator pair is shown per study. While other agreement scores, namely Spearman's rank and Krippendorff's alpha, were calculated and provided in the main article, the measures produced similar results so providing a heatmap of one measure is sufficient to assess agreement. In addition, each annotator filled out the Visual Language Fluency Index (VLFI) Cohn [2014] questionnaire to help determine annotator competence. The VLFI measures visual language fluency, and a higher score corresponds to greater fluency.

Participants were recruited using the online crowd-sourcing platform Prolific. Annotators were screened to only include UK or US nationals, fluent English speakers, people who have attained at least a Bachelor's degree, and people between the ages of 21-70.

2 Binary scale annotation task

2.1 Story 1 agreement

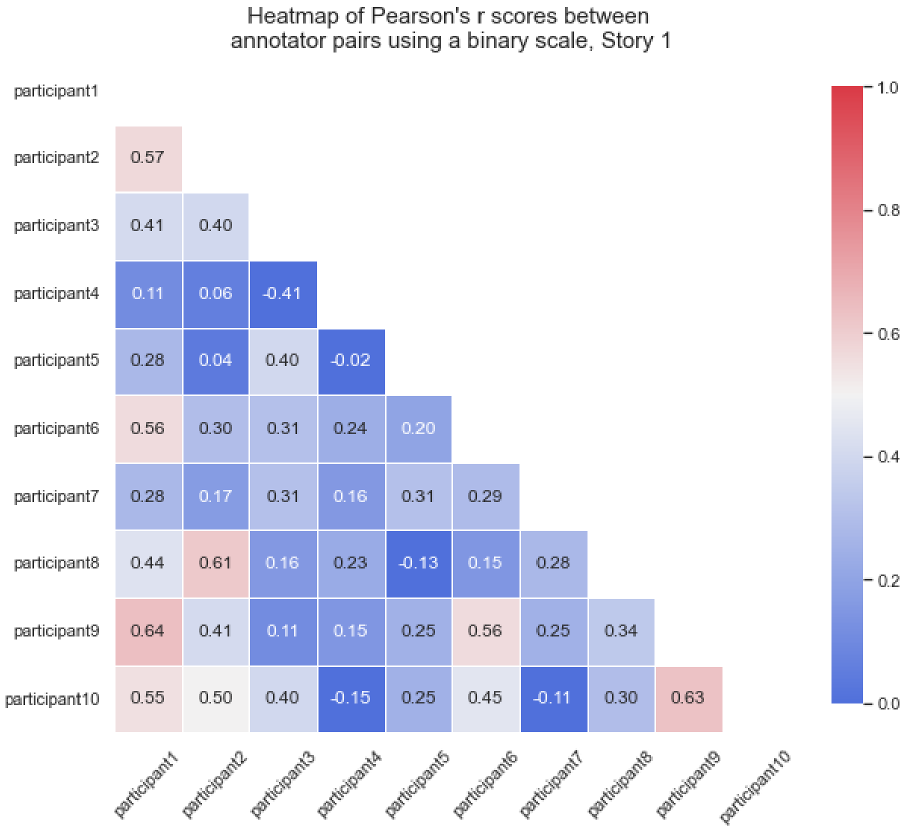
As shown in the heatmap in Figure 1, the agreement scores are generally low between all annotator pairs and rarely reach a Pearson's r score of 0.5.

Table 1 provides an overview of the demographics and VLFI scores. The overall mean VLFI score of 5.94 indicates that this group of annotators is not fluent in reading visual language. All participants have either low or very low VLFI scores, with the exception of participants 3 and 8 which have average fluency scores. Despite their higher scores, participants 3 and 8 have generally low agreement with other annotators and with each other. Participants that consistently display high disagreement (scores below 0.3) with all others are participants 4, 5 and 7, and all have low or very low VLFI scores. In addition, participant 1 is the most agreeable with other annotators, but also has a very low VLFI score.

Nevertheless, the best explanation for high disagreement between participants is that the binary scale is not amenable to the annotation task of deciding whether a panel shows background information or not. Consistent disagreement is a strong indicator that the annotation task itself was not well-formed, as varying levels of agreement introduces the possibility of unreliable annotators or an annotation task that has a wider range of interpretation. While the VLFI scores are low, other experiments described in the next sections also have low VLFI scores and achieve much higher annotator agreement. Finally, a binary scale was used for the same annotation task in a previous experiment on the same story, and produced similar low agreement results.

No.	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	27	M	2	2.5	2.75	Very low
2	25	U	1.75	2	2.375	Very low
3	52	M	8	18.75	14.625	Average
4	23	F	3	6.75	6.375	Low
5	51	F	2	3.75	3.875	Very low
6	33	F	1.5	1	4.375	Low
7	30	F	1	1	2.75	Very low
8	25	M	6	17	12	Average
9	26	F	7	9	8.75	Low
10	46	F	1	1	1.5	Very low
	Mean age:	33.8		VLFI mean:	5.9375	Low
				VLFI std:	4.17	

Table 1: Participant demographics and VLFI scores for Story 1, binary scale



	participant1	participant2	participant3	participant4	participant5	participant6	participant7	participant8	participant9	participant10
participant1										
participant2	0.57									
participant3	0.41	0.40								
participant4	0.11	0.06	-0.41							
participant5	0.28	0.04	0.40	-0.02						
participant6	0.56	0.30	0.31	0.24	0.20					
participant7	0.28	0.17	0.31	0.16	0.31	0.29				
participant8	0.44	0.61	0.16	0.23	-0.13	0.15	0.28			
participant9	0.64	0.41	0.11	0.15	0.25	0.56	0.25	0.34		
participant10	0.55	0.50	0.40	-0.15	0.25	0.45	-0.11	0.30	0.63	

Figure 1: Story 1 Binary scale heatmap

2.2 Story 2 agreement

The heatmap in Figure 2 shows mixed agreement between annotators, with most Pearson's r scores ranging from 0.3-0.7. Participants 5 and 7 appear to consistently disagree with other annotators. According to Table 2, participant 5 has the lowest VLFI score at 1.5. She selected the lowest score of 1 for all questions on the VLFI, which indicates unreliability as it is unlikely that someone would have no interaction with comics, books or film throughout their young or adult life. Participant 7 also has a very low VLFI score at 1.75, however unlike participant 5, participant 7 gave a range answers across the questions on the VLFI which does not suggest unreliability.

Interestingly, participants 4 and 7 completely disagree, with a Pearson's r score of 0.0. Participant 4 otherwise has a range of 0.3-0.7 Pearson's r scores, but also has a VLFI score in the lower range at 5.75.

Participants 6 and 8 have substantially higher VLFI scores of 17.13 and 16, respectively. Both participants have relatively high Pearson's r scores except for those with participants 5 and 7, which supports the latter's lack of reliability.

Given the above, the middling agreement for this annotation task is best explained through the binary scale being only weakly appropriate for the annotation task of choosing whether a panel shows any background information for this story. While there is clear unreliability in the case of participant 5, the middling Pearson's r scores remain when ignoring the lower scores of participants 5 and 7. Furthermore, as with story 1 above, the binary scale task in a previous experiment produced similar agreement on the same story, and other experiments have an overall low VLFI score and achieve higher agreement.

No.	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	28	F	1	2	2.75	Very low
2	24	M	2.5	5.25	4.375	Low
3	22	M	3	7.5	7.25	Low
4	46	M	6	4.5	5.75	Low
5	30	F	1	1	1.5	Low
6	29	F	9	16.25	17.125	High average
7	36	F	1	1.5	1.75	Very low
8	32	M	6	19	16	High average
9	32	F	2	1.5	3	Low
10	43	M	6.75	9.75	8.75	Average
	Mean age:	32.2		VLFI mean:	6.83	Low
				VLFI std:	5.35	

Table 2: Participant demographics and VLFI scores for Story 2, binary scale

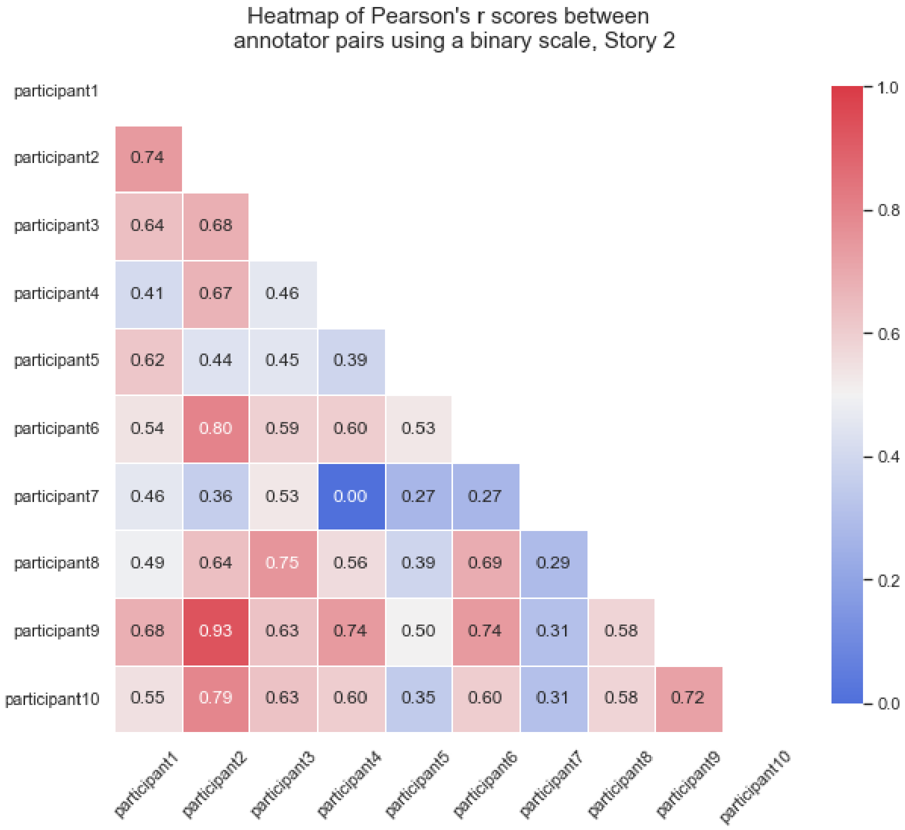


Figure 2: Story 2 Binary scale heatmap

2.3 Binary scale suitability

The inadequacy of using a binary scale for the location information amount annotation task is the best explanation as the cause of low agreement scores. Although there are several evident unreliable annotators - especially participant 5 in Story 2, who gave the same response to all questions on the VLFI questionnaire - the Pearson's r scores indicate middling agreement at best even when taking unreliable annotators into account. Story 1 in particular exhibits consistent low agreement, which is strong evidence against using a binary scale. While Story 2 exhibits general higher agreement, the scores are weaker when compared to the agreement scores in the ordinal and continuous scales.

Finally, the VLFI scores are low for both stories, which may contribute to a general trend towards disagreement as this may contribute to a wide range of semantic information interpretation in comics. Participants 6 and 8 in story 2 were the only annotators to have a high VLFI score, and both achieved higher levels of agreement across the board except when paired with consistently low agreement annotators. While a previous annotation experiment using the binary scale produced similar results for both stories, the role of the overall low visual language fluency of the annotators cannot be ruled out as a cause of disagreement. The purpose of testing this annotation scheme is to determine how to capture an average persons semantic judgments, it may nevertheless be a pitfall to use readers who haven't been screened for visual language fluency or comics reading experience.

No..	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	40	F	4.5	3	6.25	Low
2	40	F	3.75	3.75	7.75	Low
3	26	F	1	9.75	5.875	Low
4	27	M	10	4	8.25	Average
5	23	F	2.5	10	8.125	Low
6	35	M	17.25	10	20.625	High
7	22	M	26.25	12	21.125	High
8	28	F	1.25	1.5	1.875	Very low
9	29	F	3	6	11.25	Average
10	61	M	20	16	18.5	High average
	Mean age:	33.1		VLFI mean:	10.9	Average
				VLFI std:	6.44	

Table 3: Participant demographics and VLFI scores for Story 1, ordinal scale

3 Ordinal scale annotation task

3.1 Story 1 agreement

The heatmap in Figure 3 shows mid to high Pearson's r scores between annotators, except for participants 9 and 10 who exhibit more disagreement between most other annotators. Interestingly, participant 10 has the highest VLFI score of 18.5, and participant 9 has an average score of 11.25, as shown in Table 3. It could be the case that these participants interpret background detail information amount in a way that lends to more conservative responses, and refrain from assigning very high or very low numbers to panels. Still, the range of lower scores exhibited by these annotators is between 0.25-0.5, which does not indicate as much disagreement as the binary scale experiments, especially the first story.

Overall, the annotators appear reliable and disagreement can be accounted for as differing interpretations of the interval between the numbers on the ordinal scale. This is typically the case with ordinal or Likert-scale based responses. In addition, this experiment has the highest overall VLFI score of 10.9 which indicates average visual language fluency, which may also contribute to higher inter-annotator agreement in general.

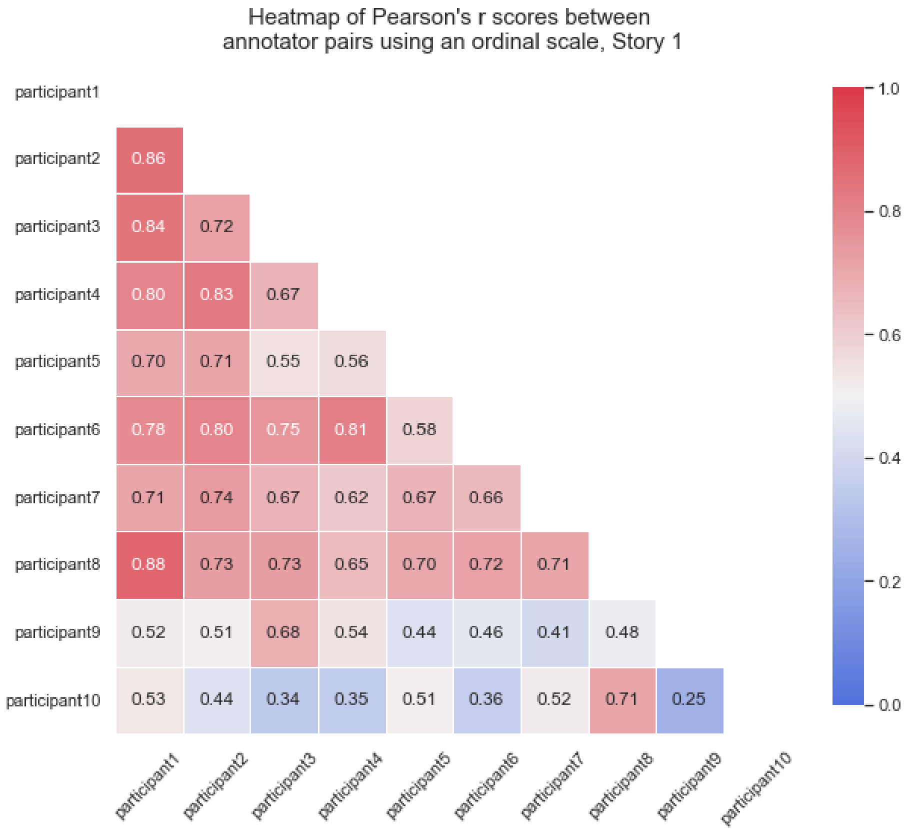


Figure 3: Story 1 Ordinal scale heatmap

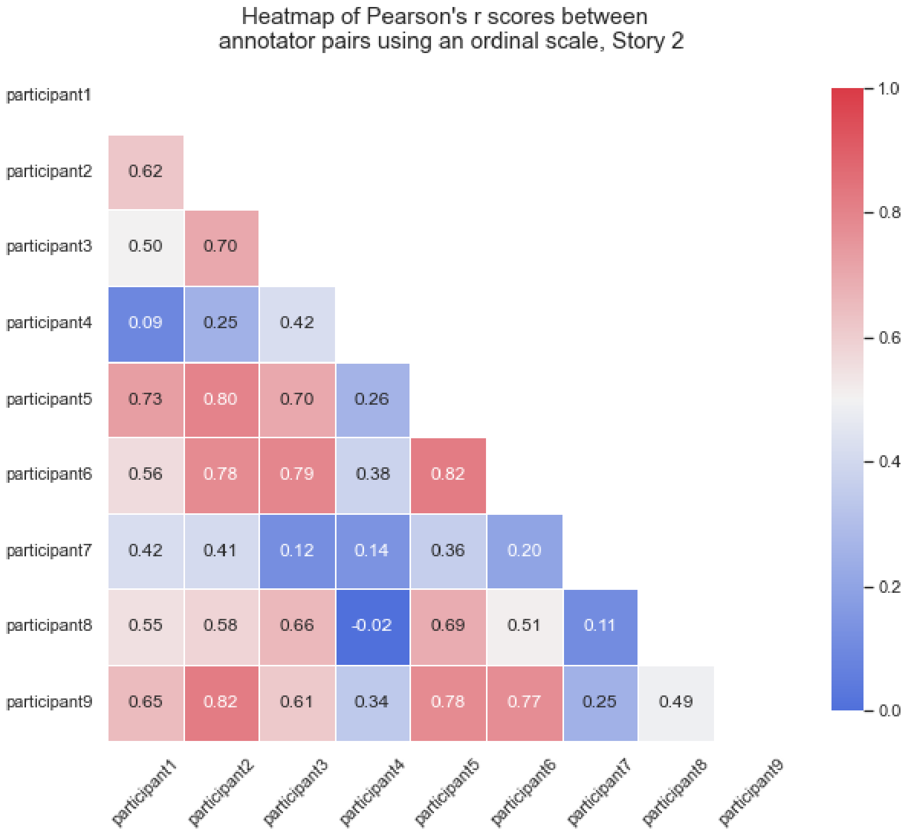
No.	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	28	F	2	5.25	9.25	Average
2	27	F	3.75	22.5	13.875	Average
3	29	F	3.5	4.5	7	Low
4	28	F	4.5	4.5	7.625	Low
5	37	F	3	8	7.25	Low
6	28	F	3.75	3	3.875	Very low
7	25	F	1	2	3.5	Very low
8	47	F	3	9	9.75	Average
9	25	M	8	1	7.5	Low
	Mean age:	30.4		VLFI mean:	7.75	Low
				VLFI std:	3.27	

Table 4: Participant demographics and VLFI scores for Story 2, ordinal scale

3.2 Story 2 agreement

The Pearson's scores between participants shows cases of both very high and very low agreement, as can be seen in Figure 4. Participants 4 and 7 exhibit consistently low scores between all other annotators, which range from -0.02-0.42. -0.02 indicates an inverse correlation, meaning that when one annotators score increases the other annotators score decreases - this is therefore very strong indication of disagreement! Pearson's r scores between all other annotator pairs range from 0.5-0.82, with many instances of high scores between 0.65-8.0.

The stark split between participants who consistently agree and participants who consistently disagree either suggests that the annotation task was interpreted in two very distinct ways, or that participants 4 and 7 are less reliable than the others. It could be the case that participants 4 and 7 understood the intervals between numbers on the ordinal scale very differently from all other annotators, as with participants 9 and 10 in the previous experiment on story 1. However, unlike the previous experiment, the disagreement here is much more drastic. Participants 4 and 7's low agreement may be attributed to low VLFI scores as shown in Table 4, but this is not a full explanation as other annotators also have low visual fluency. Given that only 2 out of 10 annotators consistently produced disagreement and that there is high overall agreement for the ordinal scale in story 1, participants 4 and 7 performance is plausibly considered to be unreliable or at least incompetent.



	participant1	participant2	participant3	participant4	participant5	participant6	participant7	participant8	participant9
participant1									
participant2	0.62								
participant3	0.50	0.70							
participant4	0.09	0.25	0.42						
participant5	0.73	0.80	0.70	0.26					
participant6	0.56	0.78	0.79	0.38	0.82				
participant7	0.42	0.41	0.12	0.14	0.36	0.20			
participant8	0.55	0.58	0.66	-0.02	0.69	0.51	0.11		
participant9	0.65	0.82	0.61	0.34	0.78	0.77	0.25	0.49	

Figure 4: Story 2 Ordinal scale heatmap

3.3 Ordinal scale suitability

The ordinal scale produces high levels of inter-annotator agreement, which is especially apparent when compared with the binary scale results. The high agreement is attributed to the annotation task being well-formed. Areas of mild disagreement may be explained by different interpretations of the interval between the integers on the given scale, which is typical. There are also two cases of annotators that consistently disagree for story 2, which can be attributed to unreliability, incompetence, or low visual language literacy.

A contributing factor to the high agreement in story 1 is the visual language fluency mean being the highest among all studies. This may increase the amount of agreement compared to the results of other scales. Nevertheless, story 2 was annotated by a group of participants with lower visual language literacy, and high agreement between annotator pairs was achieved except for participants 4 and 7. Still, as suggested in Section 2.3, it might be necessary to screen for the amount of visual language fluency when selecting annotators in the future.

No.	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	38	F	6.75	7	11.375	Average
2	26	F	1	1	3	Very low
3	22	F	1	1	1.5	Very low
4	24	M	1	1	4	Low
5	24	M	3	3	4.5	Low
6	27	F	6.75	11	13.25	Average
7	29	M	1	4	3	Very low
8	26	F	2.75	2	3.875	Low
9	22	M	1	1.25	2.125	Very low
10	23	M	1	1.25	1.625	Very low
	Mean age:	26.1		VLFI mean:	4.9	Low
				VLFI std:	3.86	

Table 5: Participant demographics and VLFI scores for Story 1, continuous scale

4 Continuous scale annotation task

4.1 Story 1 agreement

The heatmap in Figure 5 indicates mild mixed agreement between annotator pairs, as Pearson's r scores typically range from 0.3-0.7 with several cases of higher or lower agreement. Participant 10 appears to at least mildly agree with all other annotators, while no participant consistently strongly agrees or disagrees with all other participants.

The mean VLFI of 4.9 indicates low visual language fluency as shown in Table 5. However, there does not appear to be a strong relationship between individual VLFI scores and agreement amount. Participant 10, who agreed the most with all other annotators, has the second lowest VLFI score. Participants 1 and 6 were the only two to have a VLFI score in the average range, yet neither shows consistent strong agreement with all other annotators - in fact, participant 1 exhibits mostly mid-range Pearson's r values and one particularly low score or 0.11. While the overall group low visual language fluency may contribute to disagreements, the VLFI scores do not point out any clear potential for annotator unreliability or incompetence.

The weak agreement is best explained by the annotation task not being well-formed. The use of a continuous scale to describe semantic information in a panel may open up too much room for interpretation, and too fine-grained a scale when compared with the ordinal scale.

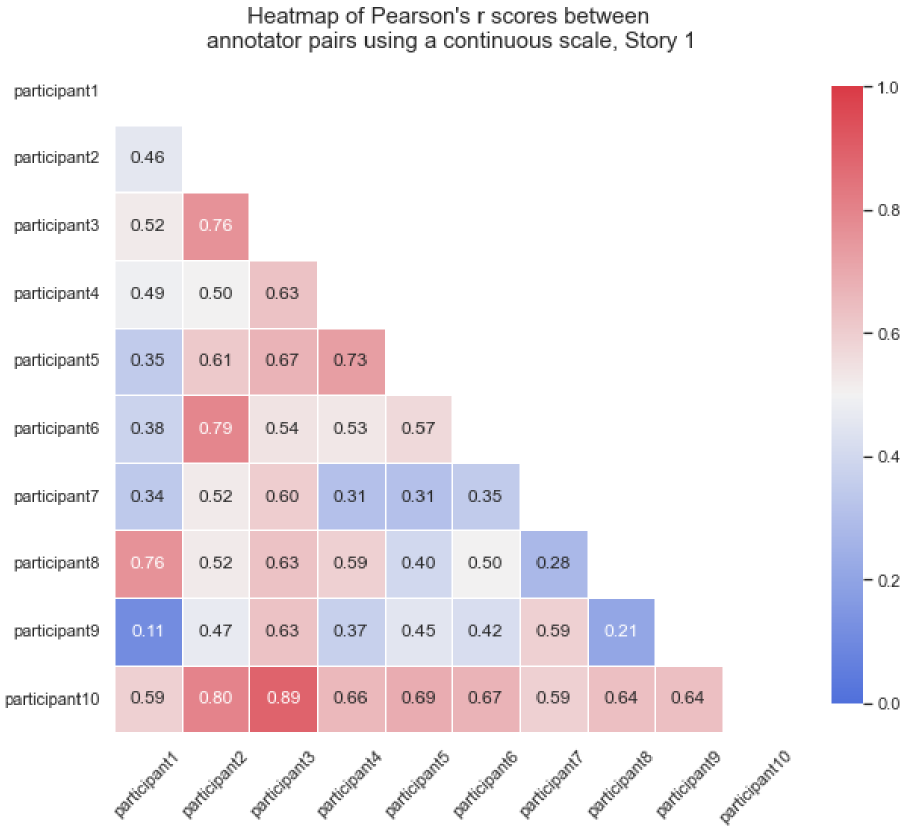


Figure 5: Story 1 Continuous scale heatmap

4.2 Story 2 agreement

Unlike the mixed agreement of story 1, agreement amount is split between annotators that tend to agree more with all others, and annotators that tend to agree less with all others. Participants 3 and 5 show consistent disagreement, while participant 2 exhibits a mild range of agreement with Pearson's r scores between 0.29-0.59, with one case of higher agreement at 0.63. Other annotator pairs tend to achieve a score of over 0.60.

The split between annotators that consistently agree and annotators that consistently disagree may be due to two distinct interpretations of the annotation task. This echoes the low agreement results for the continuous scale in story 1 in the previous section, where the wide range of interpretation plausibly explains general weak agreement. However, the divide between annotators is more stark in this case, which additionally may lend to differences in annotator reliability. While the mean VLFI score for this annotator group is average according to 6, the standard deviation indicates a wide range of visual language fluency. Again, similar to the story 1 in the previous section, there does not appear to be a clear relationship between individual VLFI score and an annotator's overall agreement. Participant 3 and 5, who displayed consistent disagreement with all other annotators, have respective VLFI score in the low to very high average range. Participant 2, who exhibited consistent mild agreement, has an average VLFI score.

The range of a agreement is best explained through the continuous scale not being a well-formed annotation task, especially when taking the results of the first story into account. While participants 3 and 5 may exhibit incompetence, and participant 2 only attains weak agreement, there are 7 other annotators showing agreement.

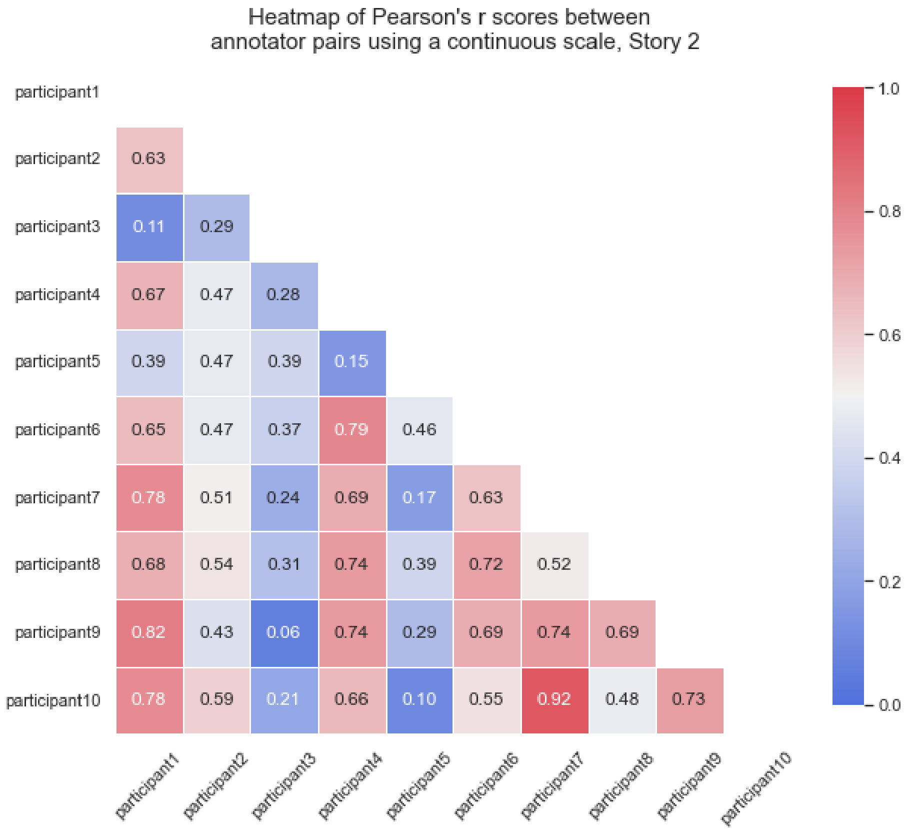


Figure 6: Story 2 Continuous scale heatmap

No.	Age	Gender	Adult VLFI	Kid VLFI	Full VLFI	Category
1	45	F	3.75	9.75	14.75	Average
2	30	F	9.75	7	11.375	Average
3	21	F	2	2	3.25	Low
4	24	NB	1	2.5	2.5	Very low
5	25	F	12	15	19.125	High average
6	38	F	6.75	14.25	17.25	High average
7	35	F	3	1.25	6.5	Low
8	33	F	7.5	9.75	18.625	High average
9	25	F	1	1.5	2.25	Very low
10	23	M	1	1	2.5	Very low
	Mean age:	29.9		VLFI mean:	9.8	Average
				VLFI std:	6.82	

Table 6: Participant demographics and VLFI scores for Story 2, continuous scale

4.3 Continuous scale suitability

The continuous scale is generally not well-formed for assigning a numerical attribute for semantic information in a panel. Overall, the results support that a continuous scale can be widely interpreted even if a trend of weak agreement emerges. The ordinal scale, while also open to interpretation of interval size between integers, may curtail the potentially wider interpretations of the continuous scale. In addition, both ordinal and continuous studies included annotators that consistently disagreed, which may be considered unreliable or incompetent.

References

Neil Cohn. The visual language fluency index: A measure of “comic reading expertise”. *Visual Language Lab: Resources*, 2014.