Skeleton_GMA_Rater judgement

2025-08-06

Rater Judgements of Fidgety Movements (FM+/FM-) with Skeleton Videos

Key Results

Expertise Effect

Accuracy significantly higher among Tutors compared to others.

Accuracy compared to chance level (bonferroni corrected):

No experience: One-sample t-test statistic: 0.791, p-value: 1.0
Certified: One-sample t-test statistic: 0.937, p-value: 1.0
Tutor: One-sample t-test statistic: 10.19, p-value: 0.00016

Difference in accuracy between groups:

Kruskal-Wallis H-statistic = 11.051, p-value = 0.0040**

Pairwise comparison (bonferroni corrected):

No experience to Certified: Mann-Whitney U statistic: 20.5, p-value: 1.0
No experience to Tutor: Mann-Whitney U statistic: 0.5, p-value: 0.0104*
Certified to Tutor: Mann-Whitney U statistic: 5.0, p-value: 0.0239*

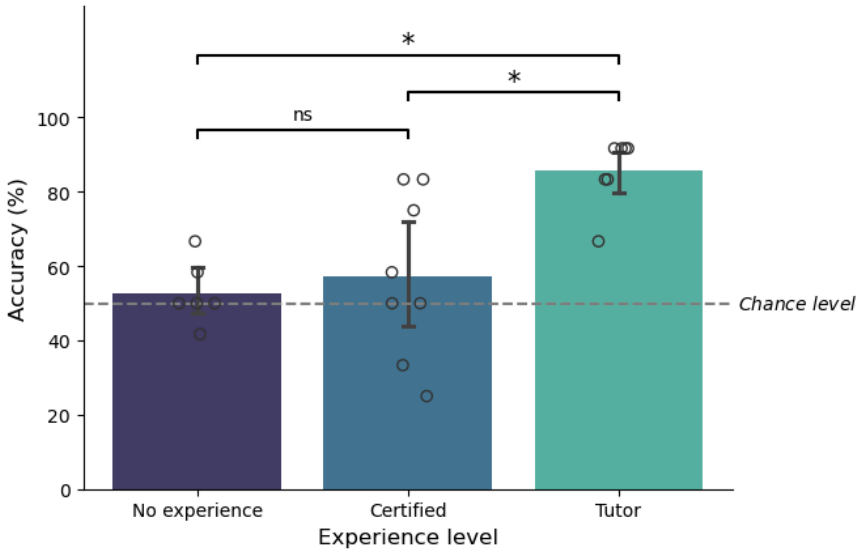


Table 1: Accuracy statistics for each experience level

GExp	mean	median	min	max
0	52.8	50	41.7	66.7
1	57.3	54.2	25	83.3
2	85.7	91.7	66.7	91.7

Table 2: Accuracy of each participant

ID	Accuracy (%)	GExp	Round
1	50	1	1
2	50	0	1
3	25	1	1
4	58.3	1	1
5	41.7	0	1
6	66.7	0	1
7	58.3	0	1
8	50	0	1
9	50	0	1
10	33.3	1	1
11	83.3	1	1
12	50	1	1
13	75	1	1
14	83.3	1	1
15	83.3	2	1
16	91.7	2	1
17	91.7	2	1
17	75	2	2
18	83.3	2	1
18	100	2	2
19	91.7	2	1
19	83.3	2	2
20	66.7	2	1
20	66.7	2	2
21	91.7	2	1
21	83.3	2	2

Accuracy: proportion of videos labelled correctly by participant

Motor Pattern

No significant accuracy difference between FM+ and FM− videos.

Difference in accuracy between motor patterns:

Mann-Whitney U test statistic: 20.0, p-value: 0.80853

Accuracy compared to chance level:

FM-: One-sample t-test statistic: 1.334, p-value: 0.47947
FM+: One-sample t-test statistic: 3.143, p-value: 0.05117

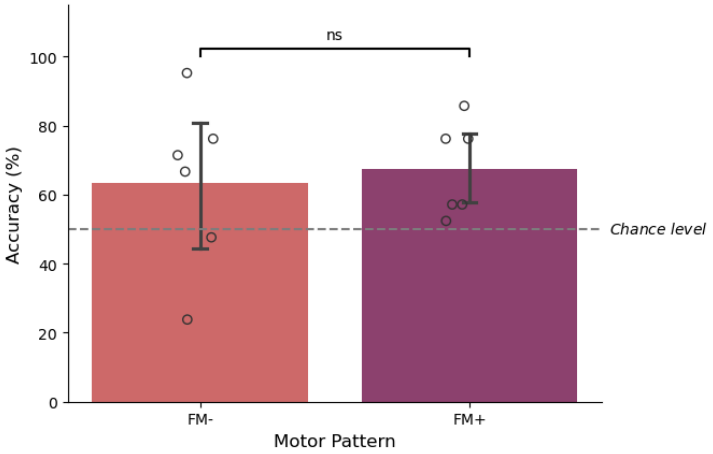


Table 3: Accuracy statistics for each Motor Pattern

Motor Pattern	mean	median	min	max
FM-	63.5	69	23.8	95.2
FM+	67.5	66.7	52.4	85.7

Table 4: Accuracy of each video

Video	Accuracy (%)
1	66.7
2	23.8
3	85.7
4	57.1
5	57.1
6	47.6
7	52.4
8	95.2
9	76.2
10	76.2
11	76.2
12	71.4

Accuracy: proportion of participants that labelled video correctly

Confidence Level (per video)

Confidence level did not predict accuracy.

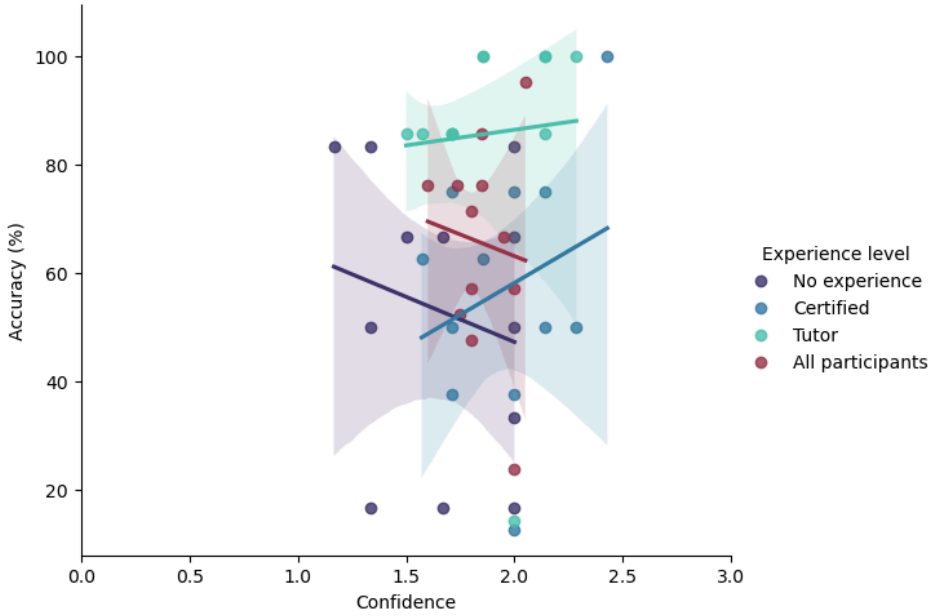
Correlation between confidence and accuracy (bonferroni corrected):

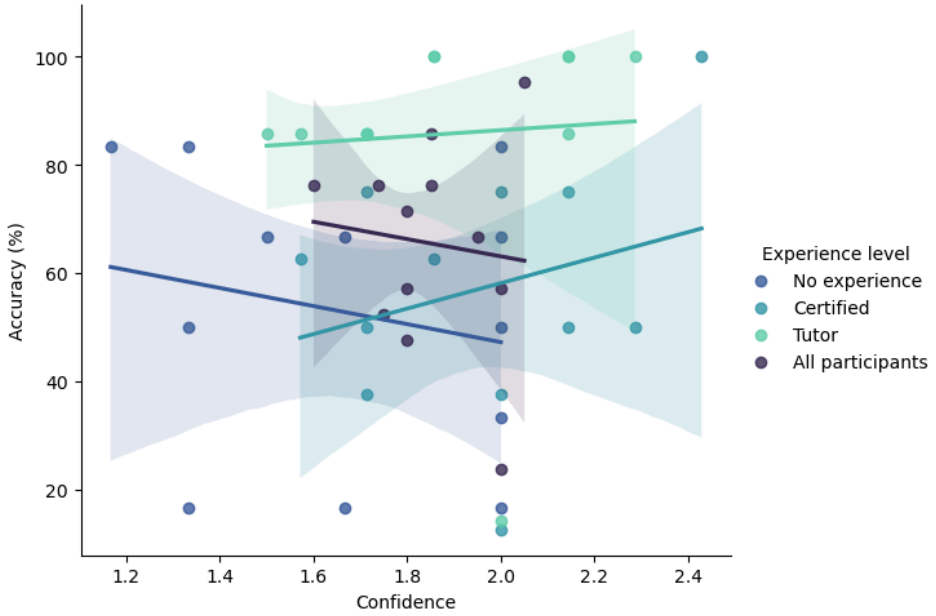
No experience: Pearson`s R correlation coefficient: -0.205, p-value: 1.0

Certified: Pearson`s R correlation coefficient: 0.267, p-value: 1.0

Tutor: Pearson`s R correlation coefficient: 0.062, p-value: 1.0

All participants: Pearson`s R correlation coefficient: -0.109, p-value: 1.0





Difference in confidence between groups:

Kruskal-Wallis H-statistic = 5.592, p: 0.0610

Table 5: Confidence and Accuracy statistics broken down by experience level (per video)

	Accuracy				**Confidence**		
Experience level	**mean**	**median**	**min**	**max**	**mean**	**median**	**min**
No experience	52.8	58.3	16.7	83.3	1.7	1.7	1.2
Certified	57.3	56.2	12.5	100	2	2	1.6
Tutor	85.7	85.7	14.3	100	1.9	1.9	1.5
All participants	65.5	69	23.8	95.2	1.8	1.8	1.6

Table 6: Accuracy and Confidence scores broken down by experience level (per video)

	Accuracy				Confidence		
Video	**All participants**	**No experience**	**Certified**	**Tutor**	**All participants**	**No experience**	**Certifiec**
1	66.7	66.7	37.5	100	2	1.7	2
2	23.8	50	12.5	14.3	2	2	2
3	85.7	83.3	75	100	1.8	2	1.7
4	57.1	16.7	62.5	85.7	1.8	1.7	1.6
5	57.1	33.3	50	85.7	2	2	2.3
6	47.6	16.7	37.5	85.7	1.8	2	1.7
7	52.4	16.7	50	85.7	1.8	1.3	2.1
8	95.2	83.3	100	100	2	1.3	2.4
9	76.2	66.7	75	85.7	1.8	2	2
10	76.2	83.3	50	100	1.6	1.2	1.7
11	76.2	66.7	75	85.7	1.7	1.5	2.1
12	71.4	50	62.5	100	1.8	1.3	1.9

Accuracy: proportion of participants that labelled video correctly

Confidence level (per participant)

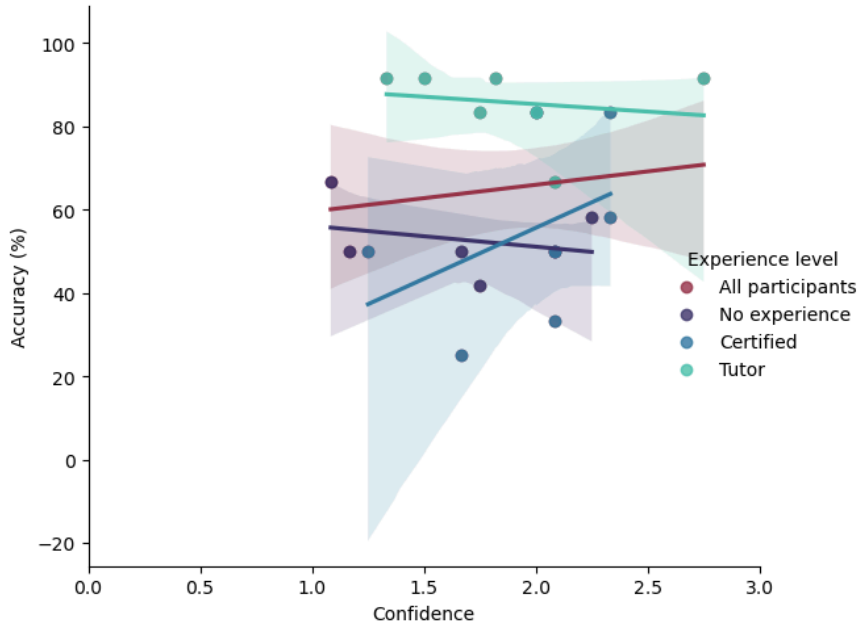
Just realised you may have intended to asses the relationship of confidence and accuracy per participant not per video. Assessing the relationship per participant or per video are both appropriate, just depends what you prefer.

All participants: Pearsons R correlation coefficient: 0.132, p-value: 1.0

No experience: Pearsons R correlation coefficient: -0.274, p-value: 1.0

Certified: Pearsons R correlation coefficient: 0.421, p-value: 1.0

Tutor: Pearsons R correlation coefficient: -0.178, p-value: 1.0



Difference in confidence between groups:

Kruskal-Wallis H-statistic = 1.629, p: 0.4429

Table 7: Confidence and Accuracy statistics broken down by experience level (per participant)

Experience level	('Accuracy', 'mean')	('Accuracy', 'median')	('Accuracy', 'min')	('Accuracy', 'max')	('Confidence', 'mean')	('Confidence', 'median')	('Confid 'min')
No experience	52.8	50	41.7	66.7	1.7	1.7	1.1
Certified	57.3	54.2	25	83.3	2	2.1	1.2
Tutor	85.7	91.7	66.7	91.7	1.9	1.8	1.3
All participants	65.5	66.7	25	91.7	1.8	1.9	1.1

Table 8: Accuracy and Confidence scores (per participant)

ID	Experience level	Accuracy	Confidence
1	Certified	50	2.1
2	No experience	50	1.2
3	Certified	25	1.7
4	Certified	58.3	2.3
5	No experience	41.7	1.8
6	No experience	66.7	1.1
7	No experience	58.3	2.2
8	No experience	50	1.7
9	No experience	50	2.1
10	Certified	33.3	2.1
11	Certified	83.3	2.3
12	Certified	50	1.2
13	Certified	75	nan
14	Certified	83.3	2
15	Tutor	83.3	1.8
16	Tutor	91.7	1.3
17	Tutor	91.7	1.5
18	Tutor	83.3	2
19	Tutor	91.7	2.8
20	Tutor	66.7	2.1
21	Tutor	91.7	1.8

Accuracy: proportion of videos labelled correctly by participant

Test-retest agreement

Agreement between round 1 and round 2 among the five tutors ranges between fair to almost perfect.

Agreement between test and retest labels:

Rater 17: Cohen's kappa = 0.657
Rater 18: Cohen's kappa = 0.667
Rater 19: Cohen's kappa = 0.833
Rater 20: Cohen's kappa = 0.333
Rater 21: Cohen's kappa = 0.824

Difference in confidence between test and retest:

Wilcoxon signed-rank test statistic: 4.0, p: 0.875

Inter-rater reliability

Inter-rater reliability was substantial among tutors (Fleiss' Kappa: 0.664), and slight among non-tutor raters (Fleiss' Kappa: 0.104).

No experience: Fleiss' Kappa: 0.108
Certified: Fleiss' Kappa: 0.098
Tutor: Fleiss' Kappa: 0.664
All participants: Fleiss' Kappa: 0.191

Extra:

- Video Difficulty Effect: Tutors performed equally well on both easy and hard videos. Other raters showed a significant drop in performance when rating the harder ones. - not sure that this makes sense since “Hard” or “Easy” videos are determined by the rater performance which is mostly non tutors so of course non tutors perform poorly on “Hard” videos.
- *Video Difficulty did not predict confidence level. - difficulty is the same as correctness/accuracy so this is the same as confidence level predicting Mean accuracy?*
- *Warming-up Effect: Accuracy improved significantly from first to second half, despite no feedback being provided. Tutors presented the highest proportional improvement controlling for their initial accuracy level. - is this needed?*

*" *** " : $p < 0.001$ (highly significant)*

*" ** " : $p < 0.01$ (moderately significant)*

*" * " : $p \leq 0.05$ (weakly significant)*

" NS " : Not significant