

Decision Making by Examiner Pairs in Clinical Assessments:

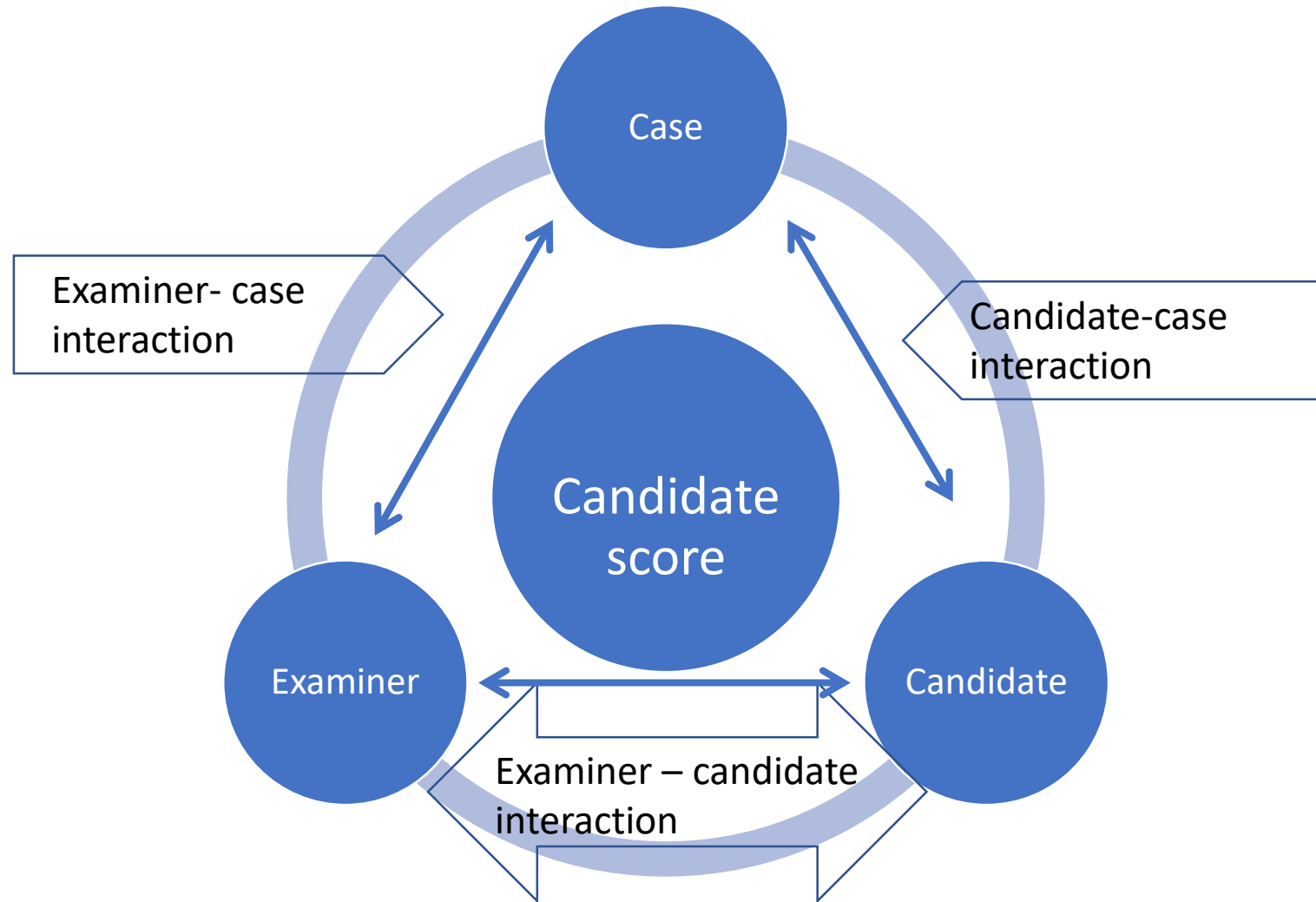
An exploration of factors which may influence candidate ratings

Dr. Aileen Faherty, Discipline of General Practice, NUIG

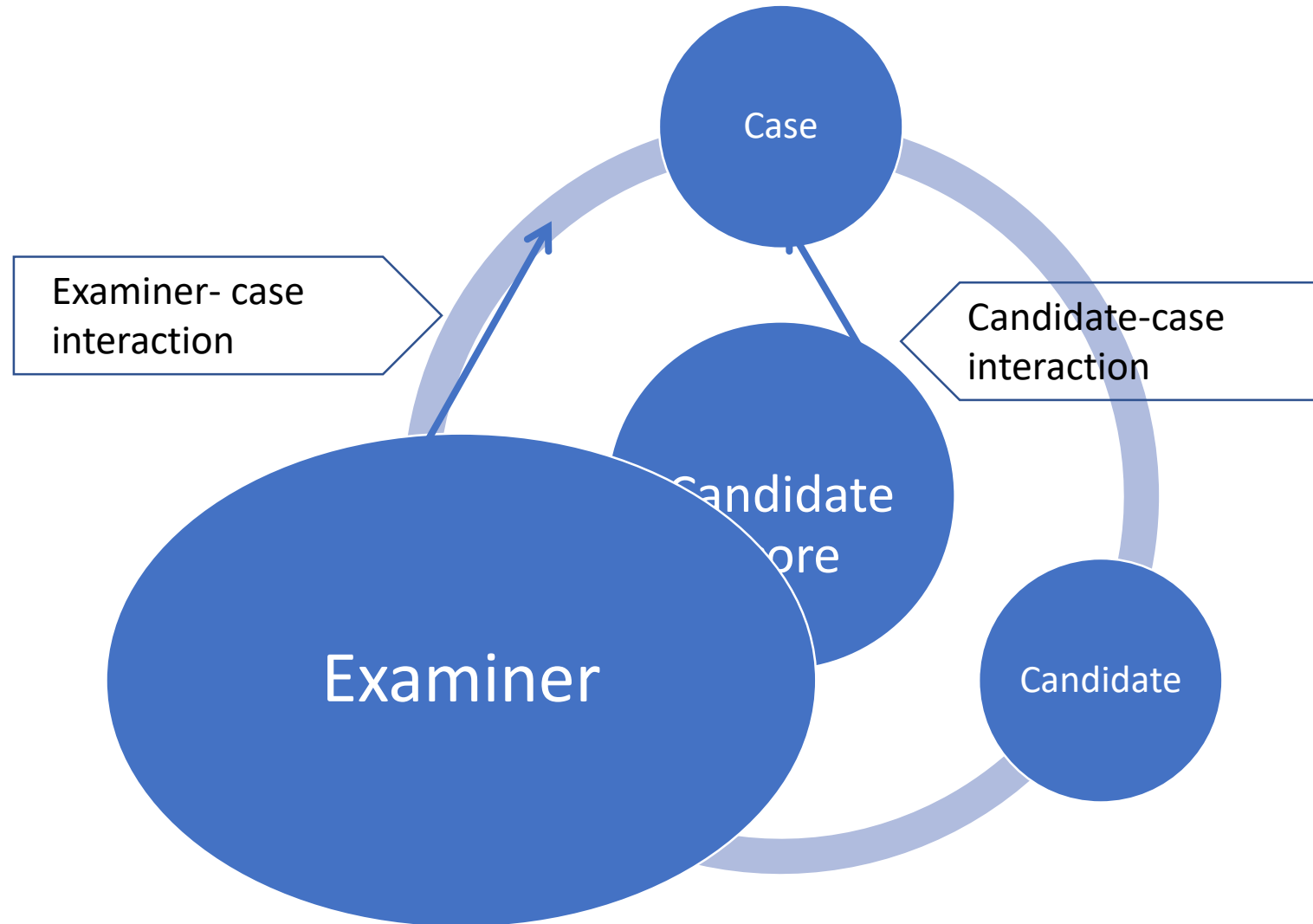
Background:

- Assessments must be **RELIABLE**
- Different results on repeated testing = **VARIABILITY**
- Comparable candidates – different results - undesirable
- **CLINICAL ASSESSMENTS** - More challenging – complex skills and behaviour being tested, case variation¹.

Variability in clinical examinations



Variability in clinical examinations



Examiner Variability

- **HAWK-DOVE** effect - can be adjusted for e.g. G-theory².
- **HALO** effect³
- **FAMILIARITY** with candidates⁴
- Gingerich et al⁵: Examiner **MOOD**
- Govaerts: Individual examiners' **PERCULIARITIES, IDIOSYNCRATIC** judgements⁶.

Gap in the literature

- Conclusion of many studies- **the MECHANISMS that contribute to Examiner variability remain unexplained and unclear** ⁷.
- Can be **overcome** by involving many examiners in the observation of many performances
- Many schools have adopted the approach of using examiner pairs
- BUT little is known about what occurs when these **paired examiners interact** to generate a score.
- Not well explored in the literature.

Aims and Objectives

- Compare inter-examiner variability of paired vs individual examiners
- To explore how ratings of examiner pairs differ from those of individual (independent) examiners
- To explore any relationship between examiners' rating and personality factors.
- Identify the cognitive processes involved in reaching an agreed score between examiner pairs.

Methods

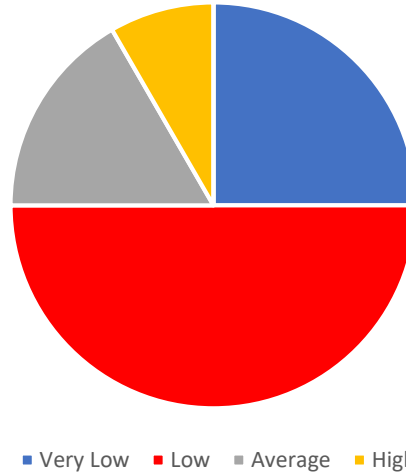
- Mixed-methods
- Quantitative arm: Quasi-experimental research design
- 12 independent examiners watching 3 videos
- Versus 6 pairs of examiners
- Physician and surgeon
- Qualitative arm: Content analysis of transcribed examiner discussions
- Convenience sample of examiners at our school.
- Demographic and personality data collected by questionnaire.

Results: Quantitative

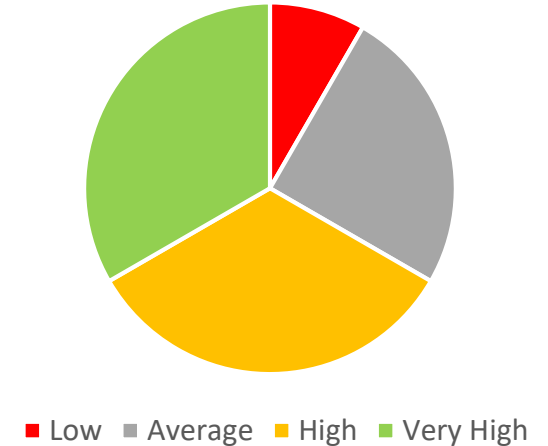
Personality

- 12 participants
- Neuroticism = 75% **below average**
- Conscientiousness = Two thirds scored **high or very high**.
- Extroversion = 75% scored **high or very high**

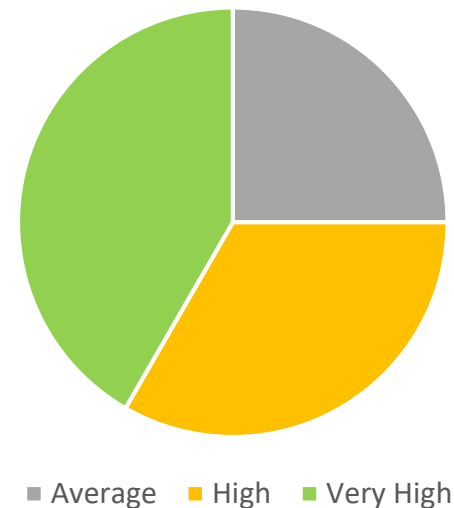
Neuroticism Score - All Examiners



Conscientiousness - All Examiners

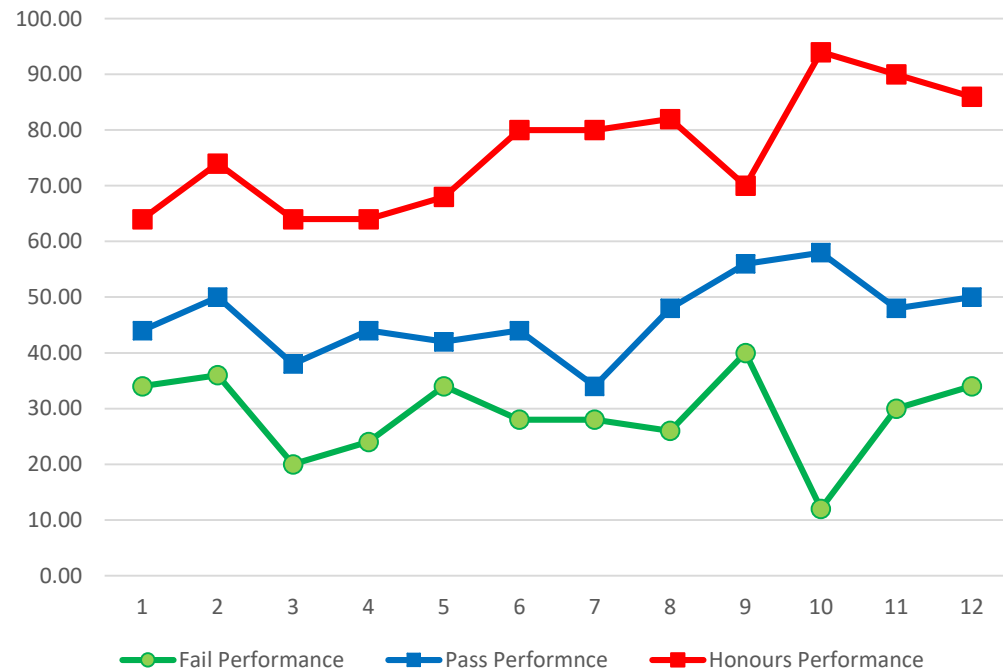


Extroversion - All Examiners

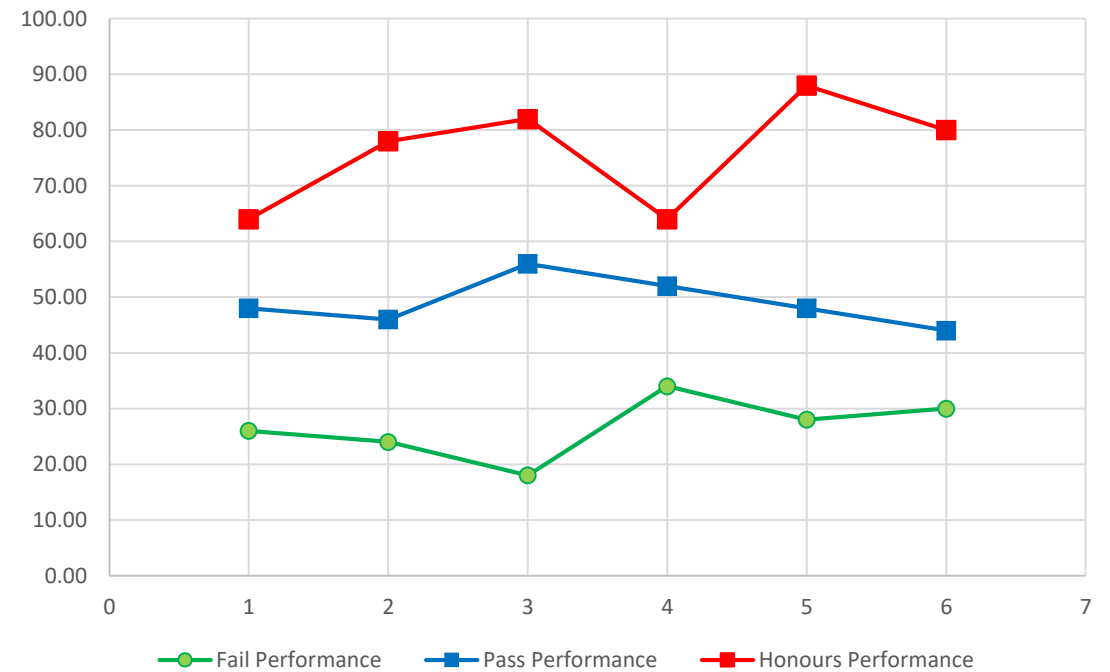


Variability of Overall Scores

Individual Examiners



Examiner Pairs



Variability of Overall Scores

	Mean Overall Score		Range	
	<u>Single Examiners</u>	<u>Examiner Pairs</u>	<u>Single Examiners</u>	<u>Examiner Pairs</u>
Honours Candidate	76.33 (10.54)	76 (9.87)	30	24
Pass Candidate	46.33 (6.86)	49 (4.33)	24	12
Fail Candidate	28.83 (7.69)	34 (5.46)	28	16

Accuracy:

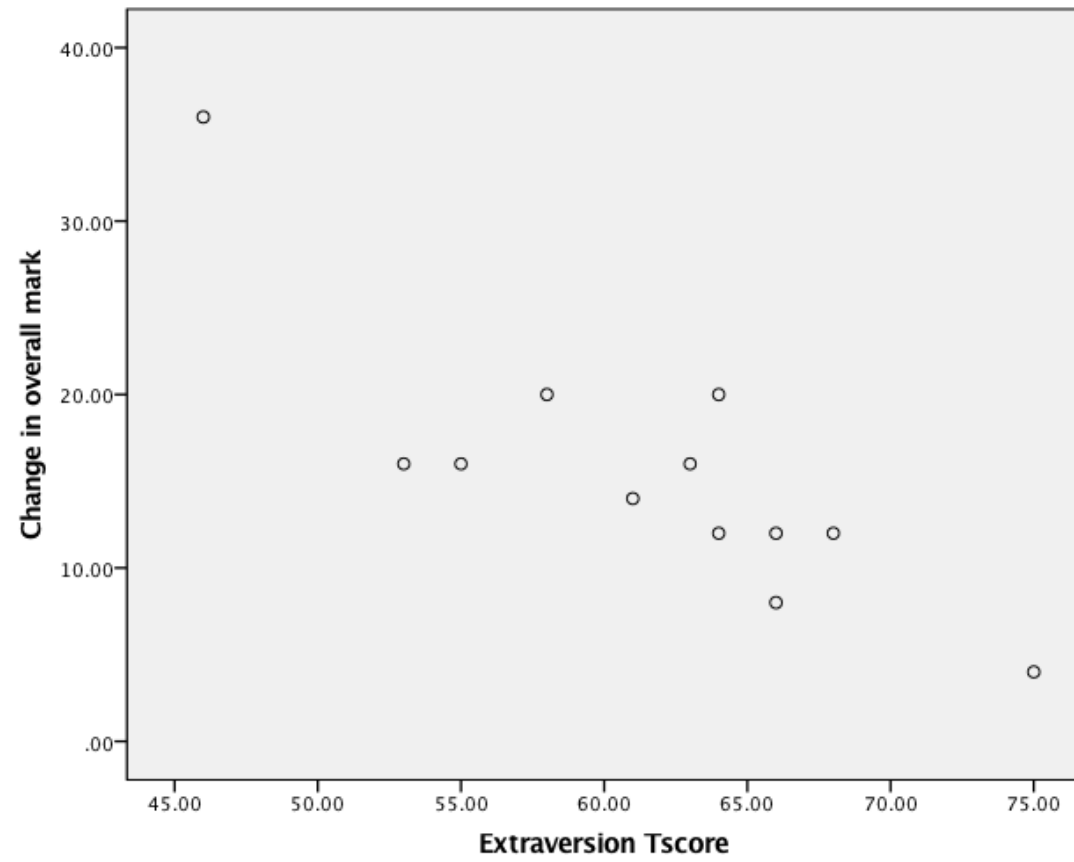
- Improved by Using Examiner Pairs
- Pass performance was **failed** by two examiners and awarded 6 borderline results when examined by **individual examiners**
- When assessed by examiner pairs the pass performance was not failed on any occasion but received 4 borderline marks.
- Statistically significant using Wilcoxon signed rank test ($p=0.0430$).

Change in marks: Independent vs paired

- Each pair tended to have a 'dominant' examiner?
- In 5/6 pairs this 'dominant' examiner was a physician.
- All of the physicians scored high or very high for extroversion
- Statistically significant correlation between change in examiner score and extroversion - **the higher an examiners score for extroversion the lower the amount of change in his or her score when paired up ($p=0.001$).**

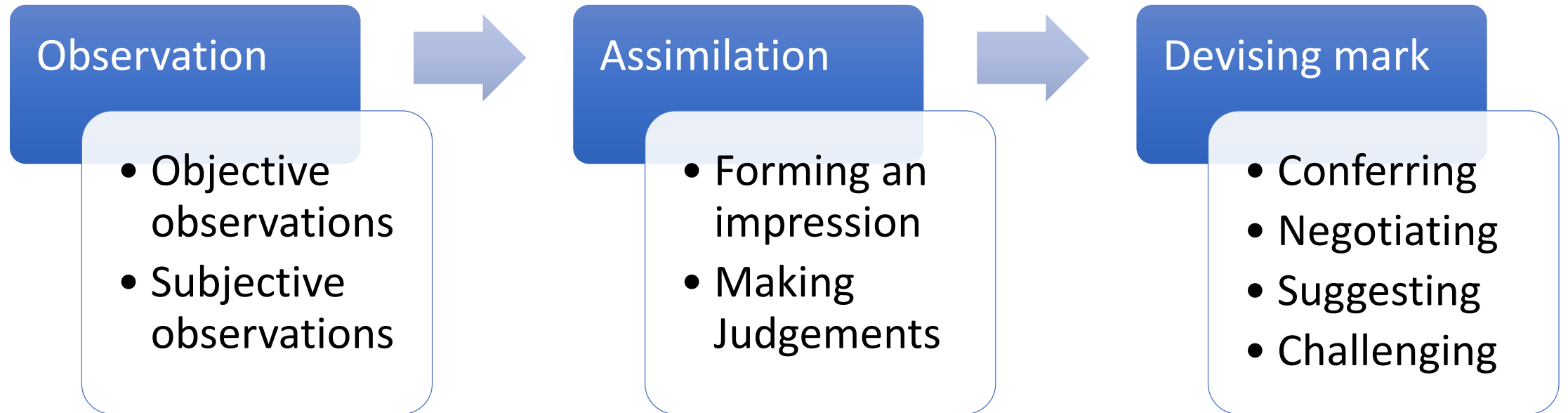
Relationship between the amount of change in examiners scores and personality

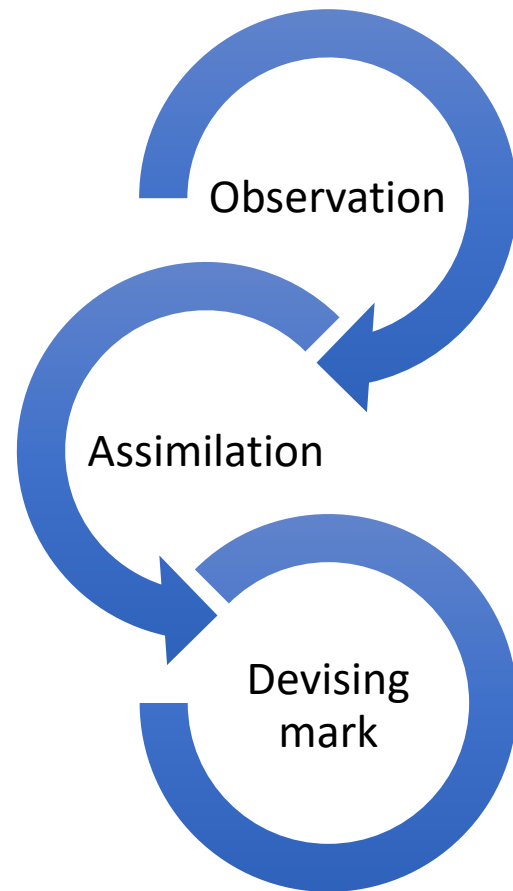
	Spearman's Correlation co-efficient rho	P value
Neuroticism	0.352	0.262
Extraversion	-0.808**	0.001
Openness to Experience	-0.185	0.565
Agreeableness	-0.501	0.097
Conscientiousness	-0.451	0.141



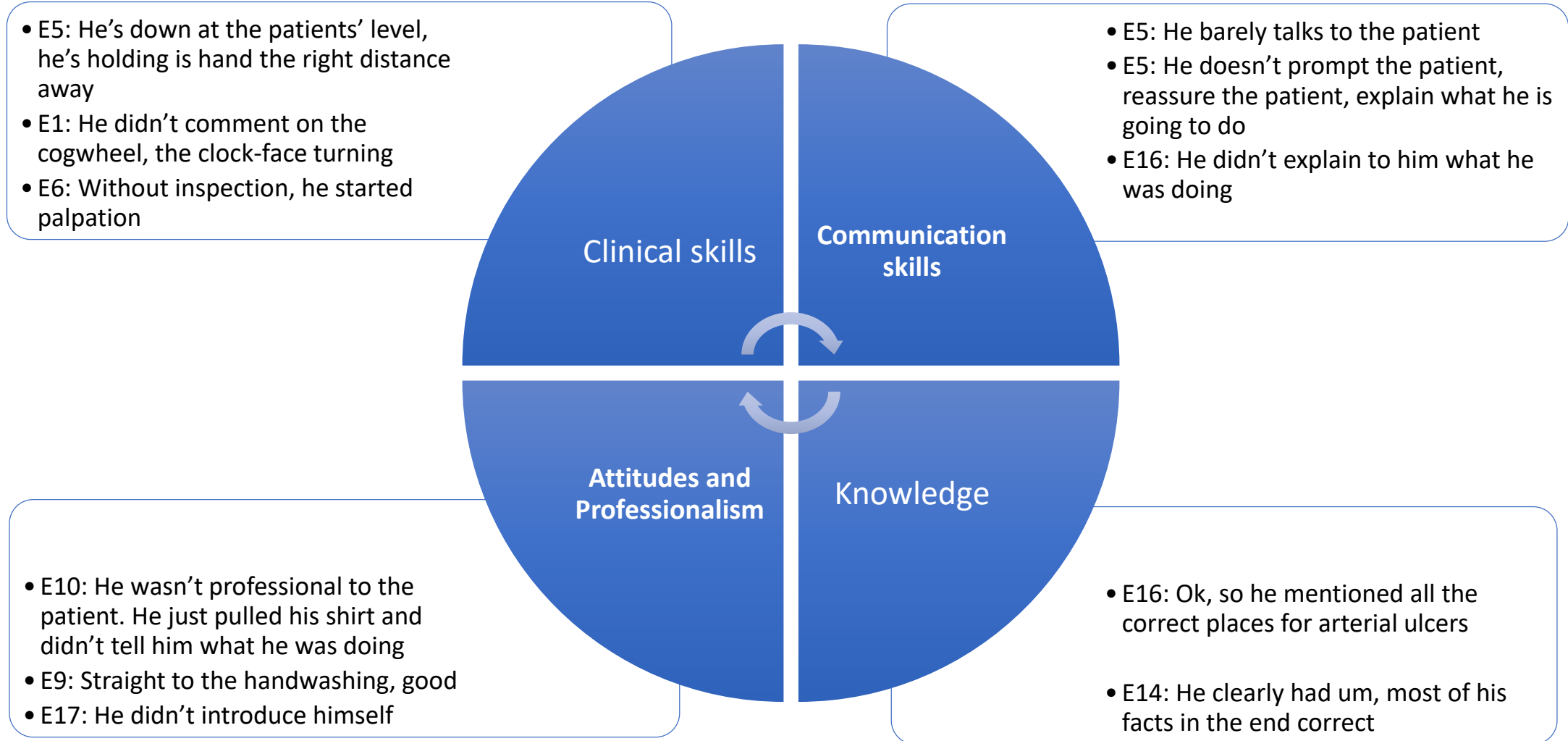
Results: Qualitative

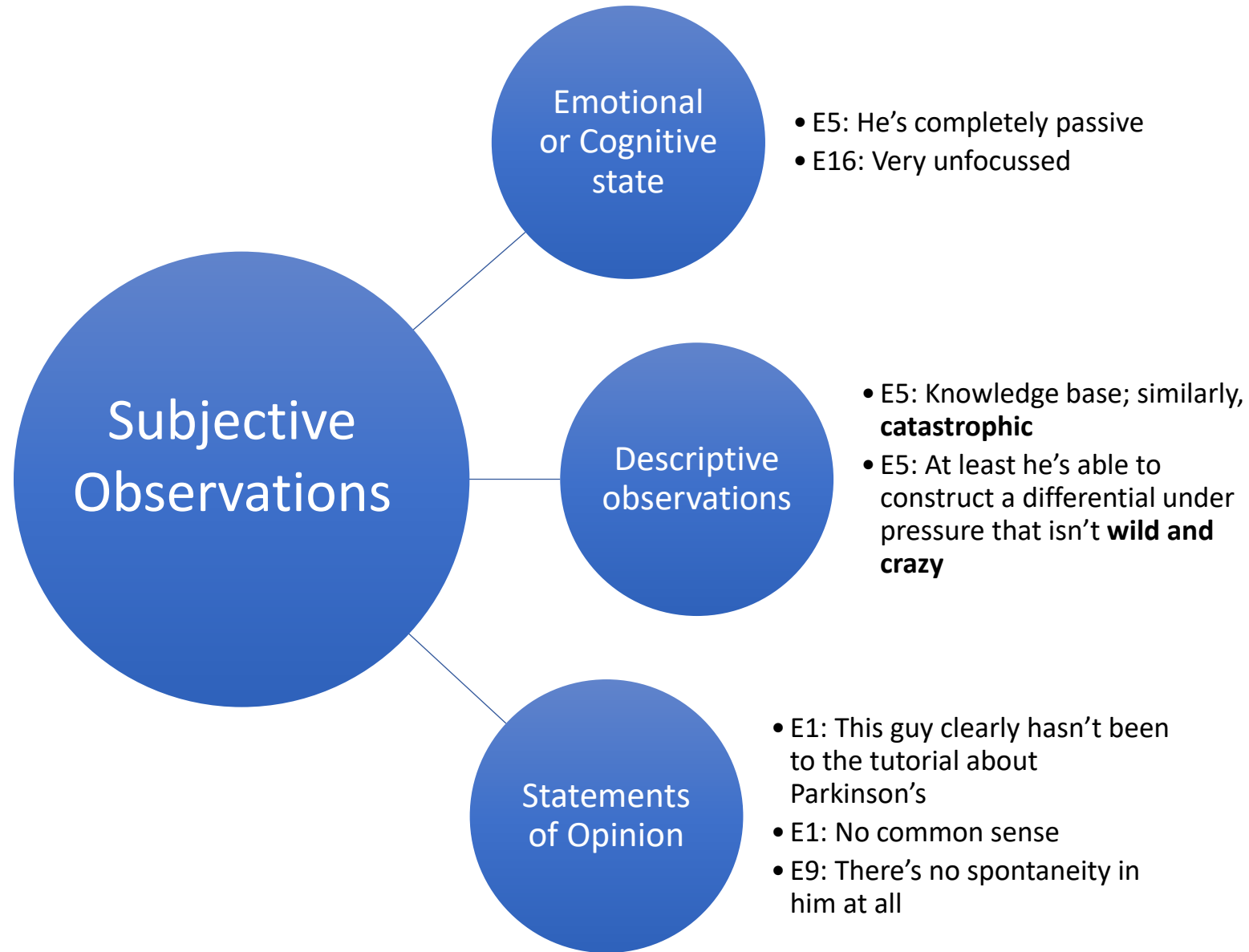
Three Main Processes





Objective Observations

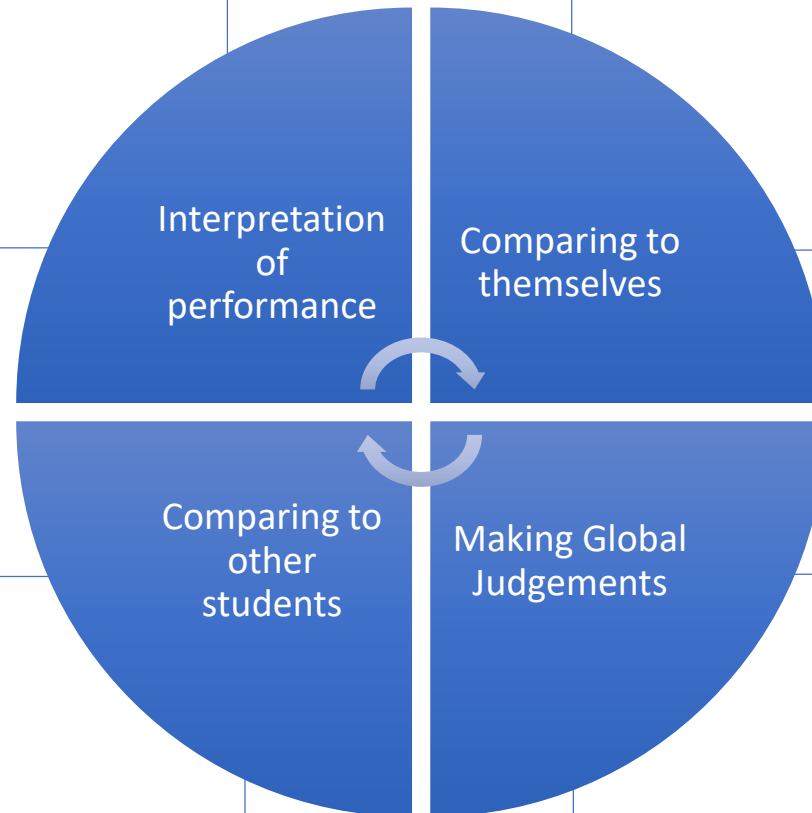




Assimilation

- E1: He needs heavy prompting but he seems to know some of the stuff
- E1: The examiner is dragging him through the exam but he's not making any fundamentally dangerous errors

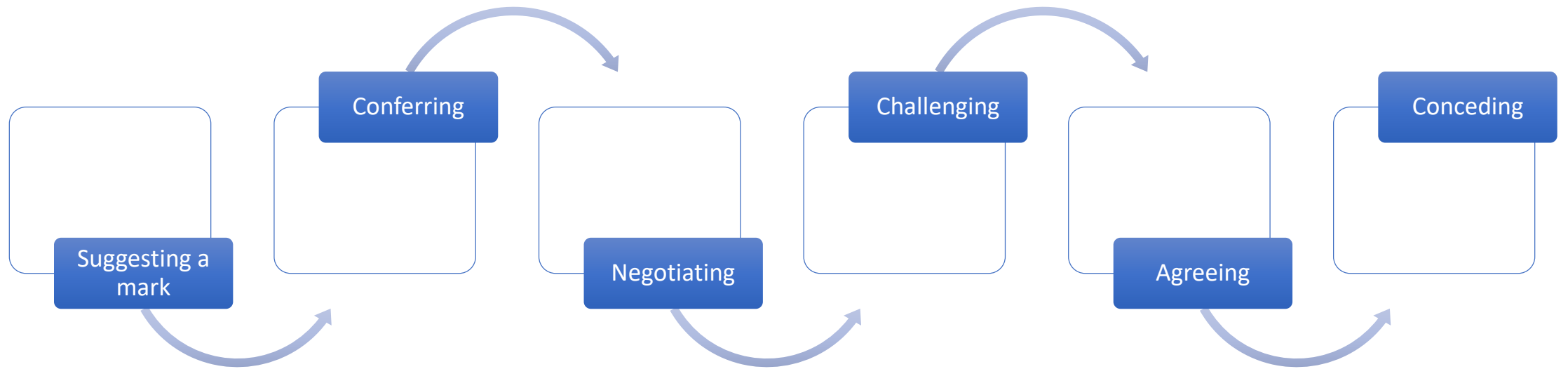
- E17: He reminded me of myself! (laughing)
- E6: **"I just thought he missed those things because I'm doing these things every day, he's a student"**



- E5: **"I guess he is an improvement from the first guy"**
- E9: This is very different to the other two

- E17: There was nothing good
- E16: I wish all students were like this!
- E7: The question you would have to ask is, could you have this guy working as your intern? ..Probably not.

Devising a mark: Potential stages



Conclusions

- Overall, it would appear from our study that the practice of using paired examiners in clinical assessments is to be recommended
- Using paired examiners improved accuracy - when examining alone two examiners would have failed the pass performance ($p=0.0430$)
- Statistically significant correlation between change in examiner score and extroversion - the higher an examiners score for extroversion the lower the amount of change in his or her score when paired up ($p=0.001$).
- Score of examiner pairs more robust score than simply averaging two independent examiners scores

Limitations

- Small sample was small
- Learning or testing effect?
- Recording – “Hawthorne effect”?

Thank you

Data Collection Exercise

- Designed to mimic our Final Medical short-case examination
- 3 video recordings of standardised student performances – 1 fail 1 pass 1 honours.
- Different case types
- First each participant viewed and graded independently the 3 recordings
- Later, examiners paired up with another to view and grade the same three performances again, with the order counterbalanced
- Discussion between examiners was recorded to produce qualitative data.
- Dependent variable: candidate's scores
- Independent variables: Examiner numbers (single or paired), examiner demographics and examiner personality.

Reliability was comparable

	Cronbach's Alpha	Intraclass Correlation Co-efficient							
		Intraclass Correlation		95% Confidence Interval		F Test with True Value 0			
				Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Examiners	0.99	Single Measures	0.887	.648	.997	98.97	2	22	.000
		Average Measures	0.990	.957	1.00	98.97	2	22	.000
Paired Examiners	0.983	Single Measures	0.925	.700	.998	60.533	2	10	.000
		Average Measures	0.987	.933	1.00	60.533	2	10	.000

Changes in examiners marks when they moved from examining alone to examining in a pair.

Examiners	Pair A		Pair B		Pair C		Pair D		Pair E		Pair F	
	<u>1</u>	<u>5</u>	<u>3</u>	<u>11</u>	<u>7</u>	<u>12</u>	<u>6</u>	<u>14</u>	<u>9</u>	<u>16</u>	<u>10</u>	<u>17</u>
Honours	0.0	0.0	4	-4	-4	-6	18	-12	8	-2	0	-6
Pass	4.0	10	-4	-2	10	-4	12	-2	4	0	10	-6
Fail	-8.0	6.0	-12	-2	0	-6	-6	6	0	-2	2	-4
Total	12	16	20	8	14	16	36	20	12	4	12	16