FACTORS INFLUENCING ASSESSOR'S CHECKLIST AND GLOBAL SCORES AT OSCE

Matic Mihevc[1], Klara Masnik[1], Tadej Petreski[1], Nejc Pulko[1], Assoc Prof Sebastjan Bevc[1,2]

[1] University of Maribor, Faculty of Medicine, Centre for Medical Education, Clinical Skills Laboratory, Maribor, Slovenia

[2]University Medical Centre Maribor, Clinic for Internal Medicine, Maribor, Slovenia

Maribor, 5th of April 2018

BACKGROUND

OSCE has become a leading method for assessing clinical skills.

- *What is the examiner‘s effect on the candidates outcome?*
- *How the examiner‘s effect can be reduced and avoided?*

EXAMINER‘S BIAS

- **Stringency-leniency effect**
- **Gender effect**
- **Time effect**

- Inconsistency effect
- Halo effect
- Restriction of range
- Contrast error
- Logical error
- Proximity error

STRINGENCY – LENIENCY EFFECT

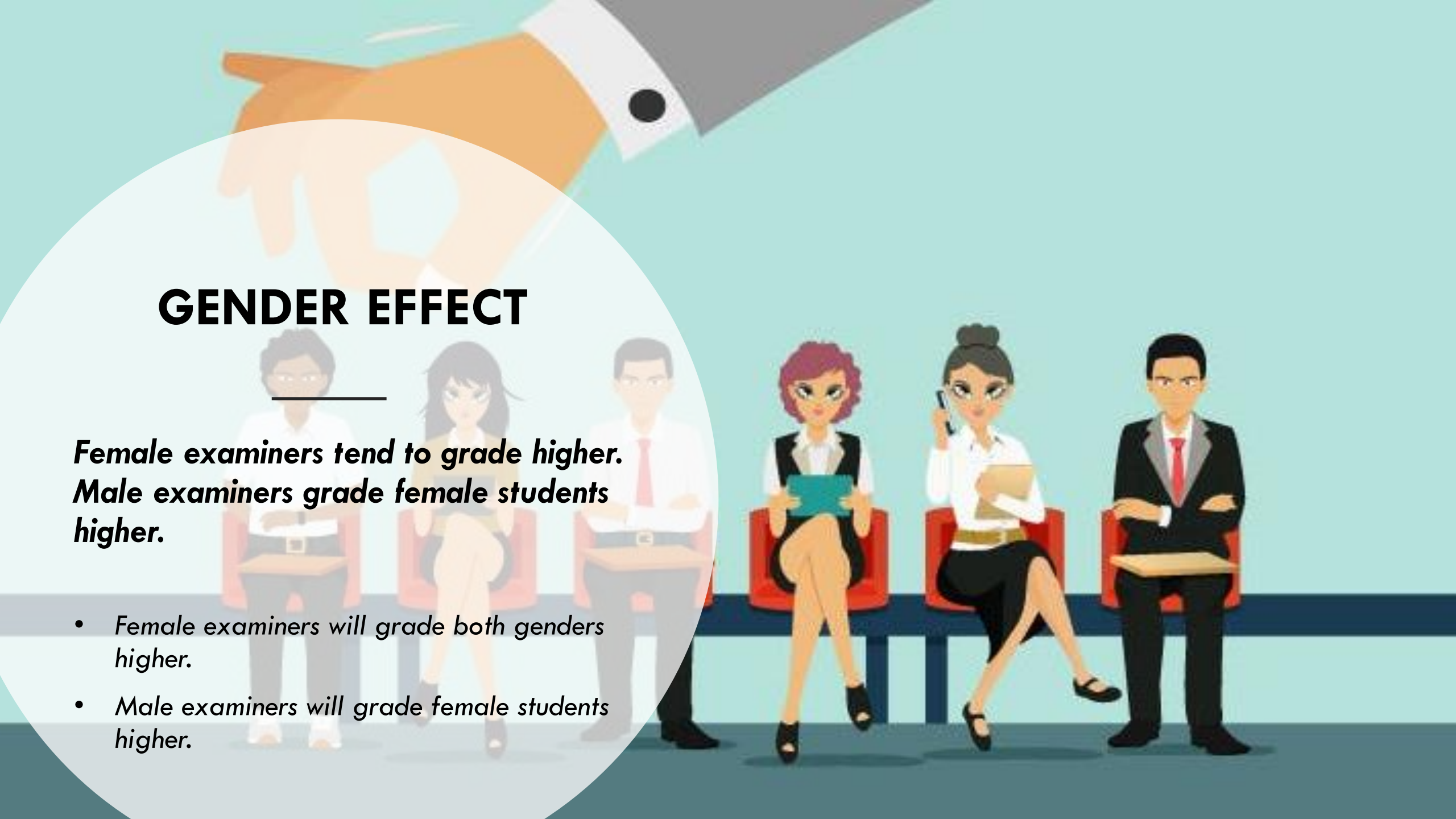
A consistent tendency of a rater to give examinees higher/lower ratings than what they should receive.

- *STRINGENCY EFFECT will appear with less students assessed (<3) and with higher year of the assessor.*
- *LENIENCY EFFECT will appear with more students assessed (>10) and with lower year of the assessor.*

GENDER EFFECT

Female examiners tend to grade higher. Male examiners grade female students higher.

- *Female examiners will grade both genders higher.*
- *Male examiners will grade female students higher.*



TIME EFFECT

Time used influence on the overall scores.

- *100% of allocated time used will result in lower overall scores.*
- *85-95% of allocated time used will result in higher overall scores.*

AIM OF THE STUDY

- To identify assessors biases,
- to compare its influence on global and checklist scores and
- to propose solutions for reduction of identified biases.

STUDY

1

The study was carried out during regular OSCE for third year medical students.

2

55 third year MS

4 researchers

10 assessors

13 models

3

Examinations of the CVS were evaluated with checklist and global rating scales.

GLOBAL SCORE

1

Communication and interpersonal skills (15%)

2

Theory and fluency of protocol performance (25%)

3

Technique (40%)

4

Report of findings (20%)

STATISTICAL ANALYSIS

- Mann-Whitney U test
- Factorial ANOVA
- Model of multiple linear regression
- Statistical significance was set at $p<0.05$.

STRINGENCY-LENIENCY EFFECT

VARIABLE	GLOBAL SCORE T vs. NT; mean±SD (n)	p	CHECKLIST SCORE T vs. NT; mean±SD (n)	p
Stringency-leniency effect				
>10 PRIOR CANDIDATES	9.16±0.71 vs. 8.97±0.75 (n=10 vs. n=45)	0.540*	**9.59±0.38 vs. 9.18±0.61 (n=10 vs. n=45)**	**0.043***
<3 PRIOR CANDIDATES	**8.57±0.92 vs. 9.16±0.59 (n=15 vs. n=40)**	**0.035***	9.02±0.73 vs. 9.34 ±0.51 (n=15 vs. n=40)	0.170*
AFTERNOON	**9.34±0.57 vs. 8.89±0.76 (n=14 vs. n=41)**	**0.041***	9.48±0.42 vs. 9.18±0.63 (n=14 vs. n=41)	0.160*
ASSESSOR IN THE 6TH YEAR	**8.36±0.88 vs. 9.11±0.66 (n=8 vs. n=47)**	**0.021***	**8.60±0.58 vs. 9.37±0.52 (n=8 vs. n=47)**	**0.002***

n=number of students; T=true; VS=versus; NT=not true; SD=standard deviation; *Mann-Whitney U test; **=factorial ANOVA

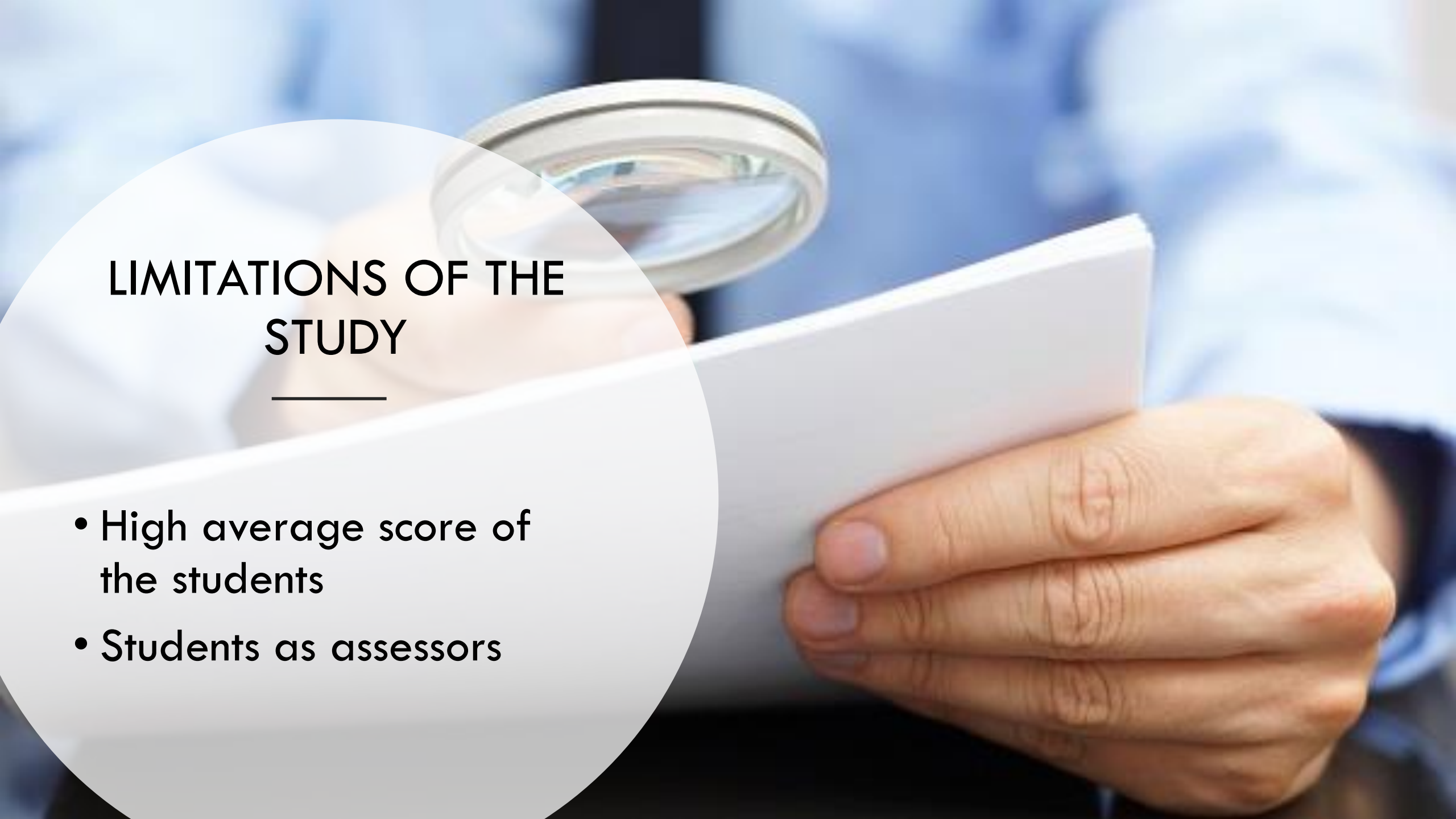
TIME AND GENDER EFFECT

VARIABLE	GLOBAL SCORE T vs. NT; mean±SD (n)	p	CHECKLIST SCORE T vs. NT; mean±SD (n)	p
Time effect				
85-95% ALLOCATED TIME USED	9.05±0.41 vs. 8.99±0.80 (n=10 vs. n=45)	0.623*	9.62±0.33 vs. 9.17±0.61 (n=10 vs. n=45)	0.028*
100% ALLOCATED TIME USED	8.84±0.84 vs. 9.16±0.60 (n=27 vs. n=28)	0.149*	8.91±0.60 vs. 9.59±0.35 (n=27 vs. n=28)	<0.001*
Gender effect				
FEMALE ASSESSOR	9.19±0.68 vs. 8.83±0.76 (n=26 vs. n=29)	0.048*	9.38±0.45 vs. 9.15±0.68 (n=26 vs. n=29)	0.276*
MALE ASSESSOR – FEMALE STUDENT	8.88±0.68 vs. 8.69±1.00 (n=7 vs. n=22)	0.397**	9.19±0.60 vs. 9.00±0.94 (n=7 vs. n=22)	0.661**

n=number of students; T=true; VS=versus; NT=not true; SD=standard deviation; *Mann-Whitney U test; **=factorial ANOVA

MULTIPLE LINEAR REGRESSION

VARIABLE	ASESSORS GLOBAL SCORE		ASESSORS CHECKLIST SCORE	
	β	p	β	p
NUMBER OF PRIOR CANDIDATES			0.288	0.011
MALE STUDENT			- 0.218	0.049
YEAR OF THE ASSESSOR	- 0.392	0.003	- 0.310	0.006
TIME USED			- 0.415	<0.001
OVERALL R^2 /p	0.154	0.003	0.440	<0.001



LIMITATIONS OF THE STUDY

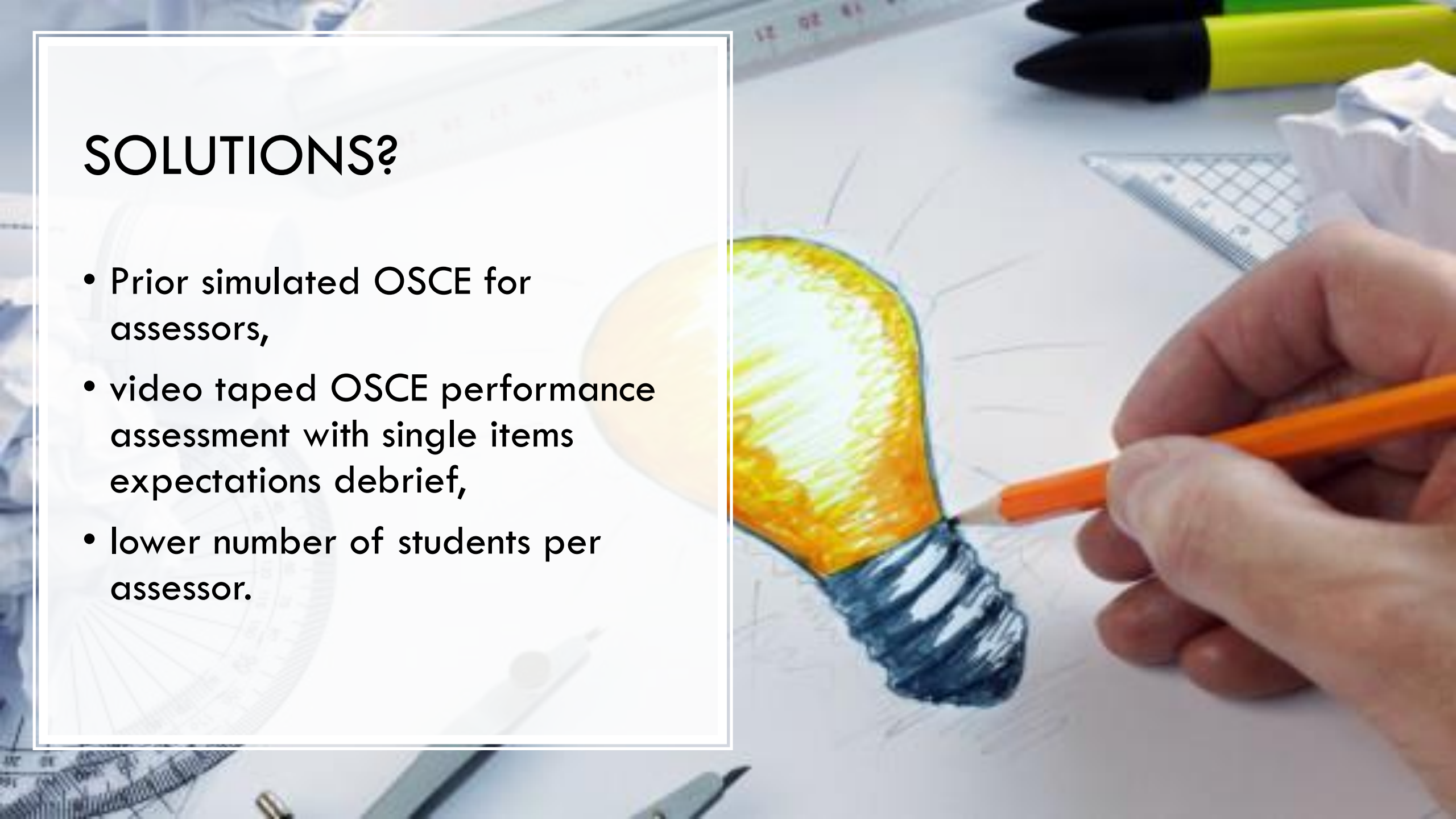
- High average score of the students
- Students as assessors

CONCLUSIONS

- Both assessment methods were prone to **stringency-leniency effect:**
- STRINGENCY FACTORS: 6th year of the assessor, less than 3 prior candidates.
- LENIENCY FACTORS: afternoon, more than 10 prior candidates.
- **Time effect** was evident from the checklist scores only.
- **Female assessors** graded candidates significantly higher when GRSs were used.

SOLUTIONS?

- Prior simulated OSCE for assessors,
- video taped OSCE performance assessment with single items expectations debrief,
- lower number of students per assessor.



REFERENCES

- Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802-11.
- Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 2015;37(12):1106-13.
- Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49(2):161-73.
- Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181-9.
- Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. Eval Health Prof. 2007;30(3):266-83.
- Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, et al. Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ. 2017;17(1):71.

THANK YOU FOR YOUR ATTENTION!