Comparing the Online and Paper-Based Versions of the TOEIC L&R

Jean-Pierre J. Richard

PanSIG 2022, The University of Nagano
July 8 - July 10 2022

長野県立大学
THE UNIVERSITY OF NAGANO

Outline

- Background, Literature Review, Research Question
- Methodology
- Results
- Discussion & Limitations
- Conclusion & Future Work

Background

Standardized Testing at Shōzan U.

- **Class Placement**:
  - since 2018: CASEC *(Computerized Assessment for English Communication)*
- **Program Evaluation**:
  - 2018 - 2019: TOEIC L&R (paper-based)
  - since 2020: TOEIC L&R (online)

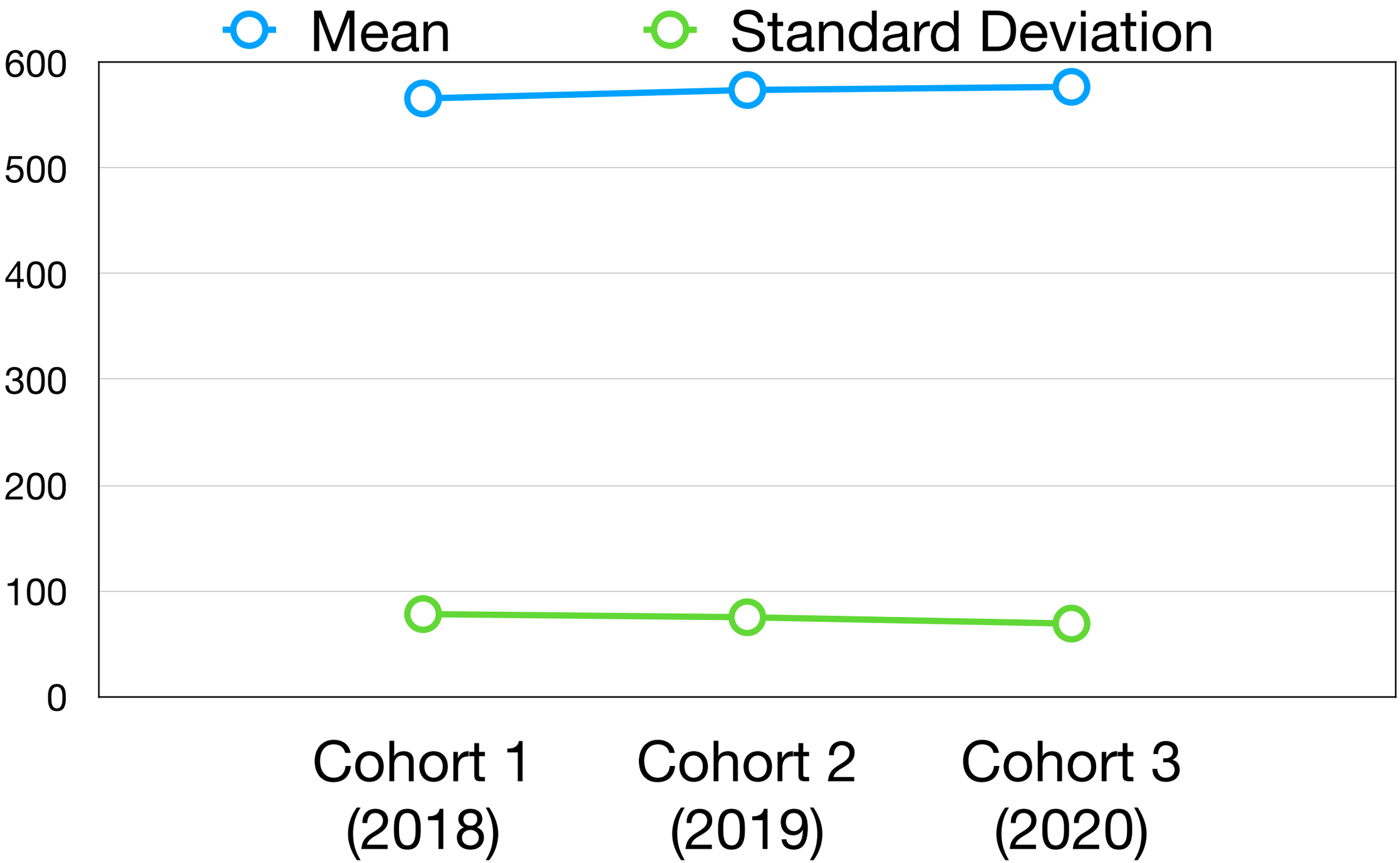
長野県立大学
THE UNIVERSITY OF NAGANO

The *Computer-Adaptive* CASEC

Section	Qs (Sum)	Max Time
1 Vocabulary	16 (16)	60 sec / Q
2 Phrasal Expression and Usage	16 (32)	90 sec / Q
3 Listening for the Main Idea	17 (49)	60 sec / Q
4 Listening for Specific Information	11 (60)	120 sec / Q

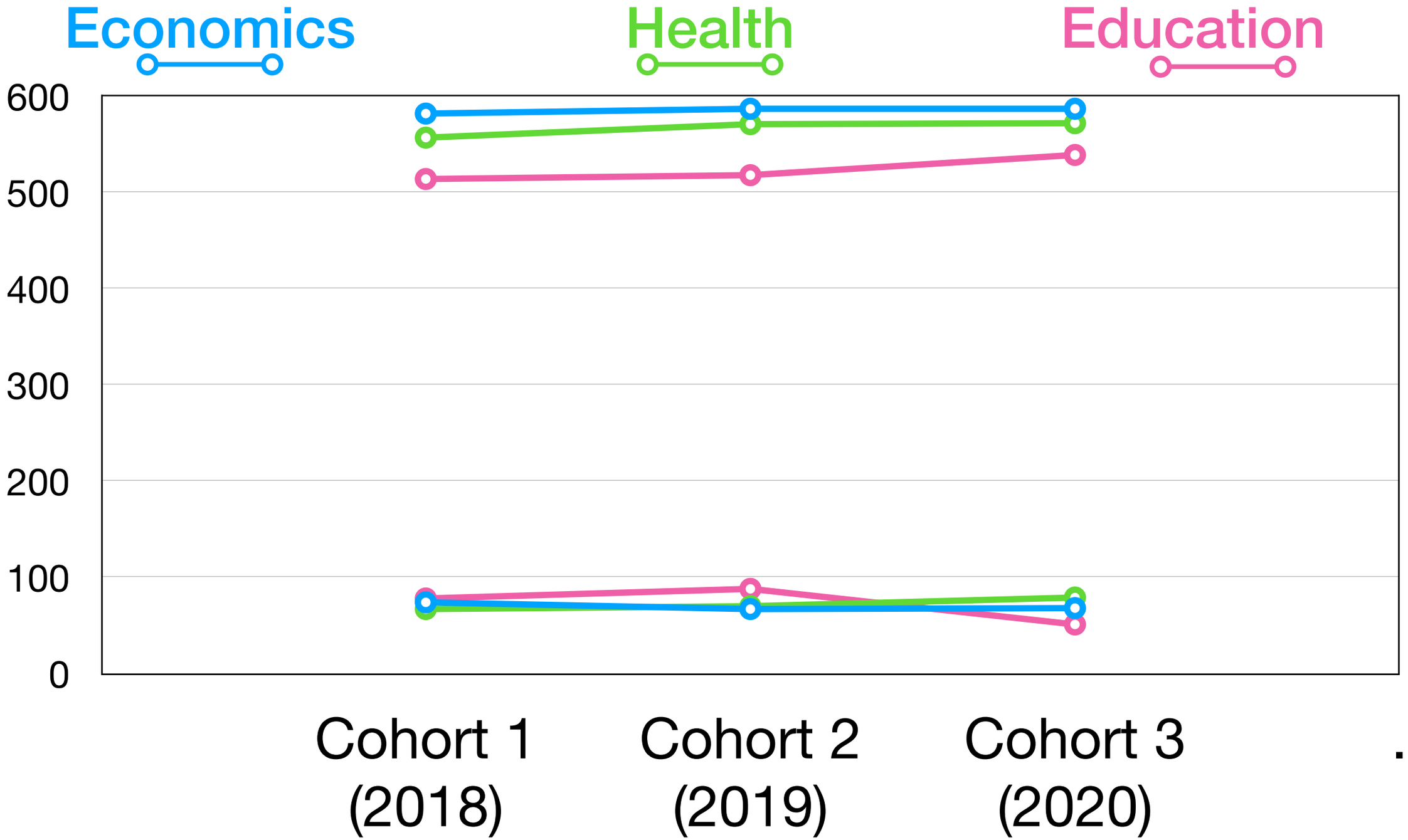
* According to CASEC, most students finish in 40-50 minutes.

CASEC Scores by Cohort



長野県立大学
THE UNIVERSITY OF NAGANO

CASEC Scores by Department



.

The TOEIC L&R (Paper-Based)

Skill	Type	No. of Qs (Total)	Minutes (Total)
Listening	Photos	6 (6)	45 (45)
	Question-Response	25 (31)	
	Conversations	39 (70)	
	Short Talks	30 (100)	
Reading	Incomplete Sentences	30 (130)	75 (120)
	Text Completion	16 (146)	
	Reading Comprehension	29 (175)	
		25 (200)	

The TOEIC L&R Online Test

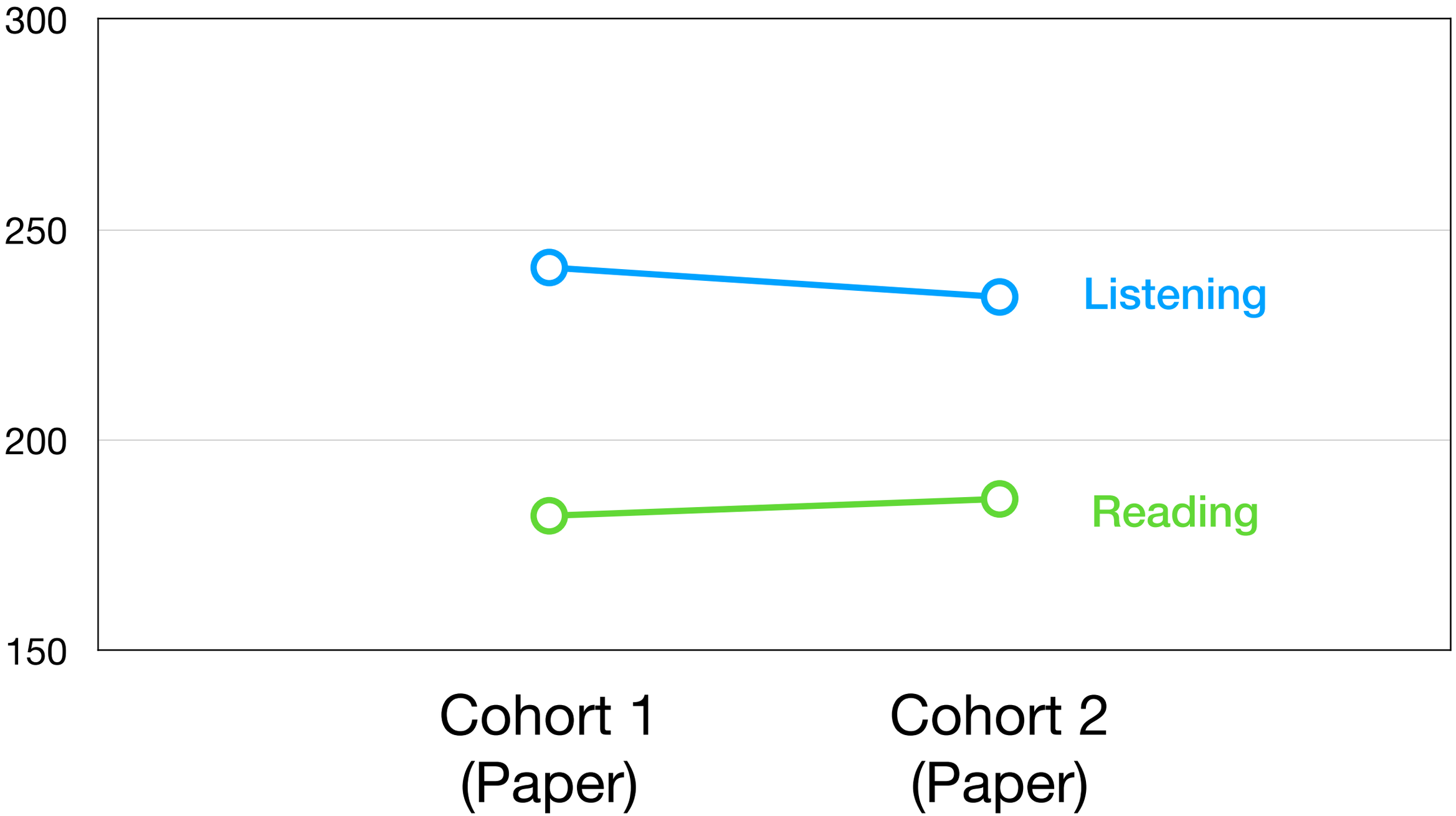
Skill	CAT	Type	No. of Qs (Total)	Minutes (Total)
Listening	No	Photos	3 (3)	
		Question-Response	4 (7)	
		Conversations	9 (16)	
		Short Talks	9 (25)	
	Yes	Question-Response	5 (30)	
		Conversations	9 (39)	
		Short Talks	6 (45)	25 (25)
Reading	No	Incomplete Sentences	5 (50)	
		Text Completion	4 (54)	
		Reading Comprehension	16 (70)	
	Yes	Incomplete Sentences	7 (77)	
		Text Completion	4 (81)	
		Reading Comprehension	9 (90)	37 (62)

CAT = Computer Adaptive Test

TOEIC Paper vs Online

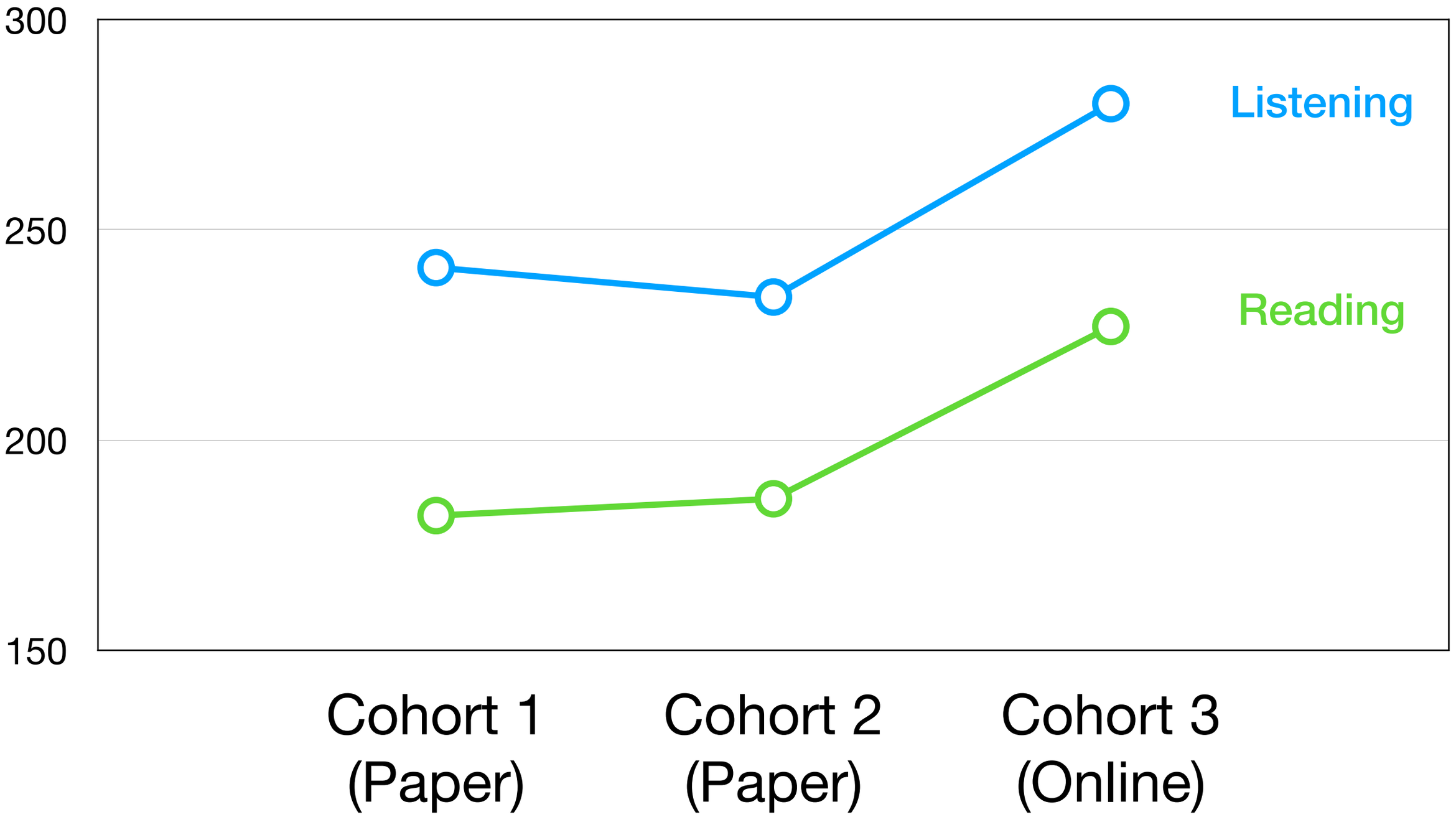
Skill	No. of Qs (Total)		Minutes (Total)		CAT	
	Paper	Online	Paper	Online	Paper	Online
Listening	100 (100)	45 (45)	45 (45)	25 (25)	✖	○
Reading	100 (200)	45 (90)	75 (120)	37 (62)	✖	○

Mean TOEIC L&R Scores By Cohort



長野県立大学
THE UNIVERSITY OF NAGANO

Mean TOEIC L&R Scores By Cohort



長野県立大学
THE UNIVERSITY OF NAGANO

ANOVAs - Comparing Cohorts

	CASEC (late March)	TOEIC Listening (April - May)	TOEIC Reading (April - May)
Parametric or Non-parametric	Non-Parametric		
Test	Kruskal-Wallis 1-way non-parametric ANOVA		
Result	$H(2) = 2.43$, $p = .297$, $\varepsilon^2 = .003$.	$H(2) = 84.30$, $p < .001$, $\varepsilon^2 = .121$	$H(2) = 100.22$, $p < .001$, $\varepsilon^2 = .144$
Pairwise companisons	NA	2018 & 2019 ($z = 1.31$, $p = .10$) 2018 & 2020 ($z = 7.11$, $p < .001$ 2019 & 2020 ($z = 8.55$, $p < .001$)	2018 & 2019 ($z = 1.35$, $p = .09$) 2018 & 2020 ($z = 9.19$, $p < .001$ 2019 & 2020 ($z = 7.97$, $p < .001$)
Summary	2018 = 2019 = 2020	**2020 > 2018 = 2019**	**2020 > 2018 = 2019**

Literature Review

Comparing Pre- and Post-Updated Paper-Based TOEIC L&R[1]

Mean Differences in Scaled Scores

Researcher	*N*	Site	Listening	Reading
Cid et al (2017)	3673	Japan & Korea	3.11	1.39
Kanzaki (2017)	141	Japan	0.96	11.46[2]

Notes:
1. ETS updated the paper-based L&R test in May 2016.
2. The difference in Reading scores observed by Kanzaki was statistically significant but with a negligible effect size (d = 0.16).

Introducing the Online TOEIC L&R

「本物を! ETS開発の正式なテスト従来のスコアと意味は変わらない」

> “The real thing! An ETS formal test where the interpretation is the same as a traditional test score” (IIBC, 2020a)

「評価やスコアの意味合いは、公開テストや従来のIPテ ストと同様で、スコアが同じであれば、英語力も同等です」

> “The meaning of evaluations and scores is the same as in public tests and conventional IP tests, and if the scores are the same, the English proficiency is also the same” (IIBC, 2020b)

Note: Author’s translations.

Research Question

Do the paper-based and online TOEIC L&R tests result in parallel scores?

Participants

2021:

n = 56: Year 1 = 56 (100%)

2022:

n = 54: Year 1 = 14 (26%); Year 2 = 40 (74%)

Participants & Test Procedures

Year	Month	Day 1	Day 3	Day 5
2021	Feb	*n* = 28 (online)	*N* = 56 (paper)	*n* = 28 (online)
2022	Feb	*n* = 27 (paper)	*N* = 54 (online)	*n* = 27 (paper)

T-tests

- used to compare differences between two groups
  - paired sample t-test: 1 group did 2 tests
  - independent sample t-test: 2 groups did 1 test
- **paired sample t-tests were used in this study**

Standard Error of Difference (*SE Diff*)

- *SE Diff*: the error of measurement associated with the difference between scores from two test administrations
- ETS estimates that, for both the Listening and Reading sections, the SE *Diff* is ±35 scaled score points
  - If 2 scores (of the same section of the test) ARE greater than ±35 scaled score points, the scores are statistically different
  - If 2 scores (of the same section of the test) are NOT greater than ±35 scaled score points, the scores are NOT statistically different

Standard Error of Difference (*SE Diff*)

- *SE Diff*: the error of measurement associated with the difference between score[illegible] administrations
- ETS estimates that, for b[illegible] sections, the SE *Diff* is ±35 scaled score points
  - If 2 scores (of the same section of the test) ARE greater than ±35 scaled score points, the scores are statistically different
  - If 2 scores (of the same section of the test) are NOT greater than ±35 scaled score points, the scores are NOT statistically different

a useful statistic with regards to TOEIC L&R

Example of *SE Diff*

	TOEIC Listening (or Reading) SE *diff* ±35 Points			Has this student's score changed
Student	Test 1	Test 2	Score Diff	
Junko	285	300	15	×
Keiji	225	200	-25	×
Taro	270	305	35	×
Mari	250	290	40	○
Naomi	285	235	-50	○
Shin	185	250	65	○

長野県立大学
THE UNIVERSITY OF NAGANO

Results

Mean (*SD*) TOEIC Scores

2021 (*n* = 56)	*Online* *M (SD)*	*Paper-Based* *M (SD)*	*Δ*
Listening	337 (57)	300 (55)	37
Reading	270 (67)	254 (64)	16
2022 (*n* = 54)			
Listening	355 (47)	348 (44)	7
Reading	314 (54)	281 (51)	33

T-test Results

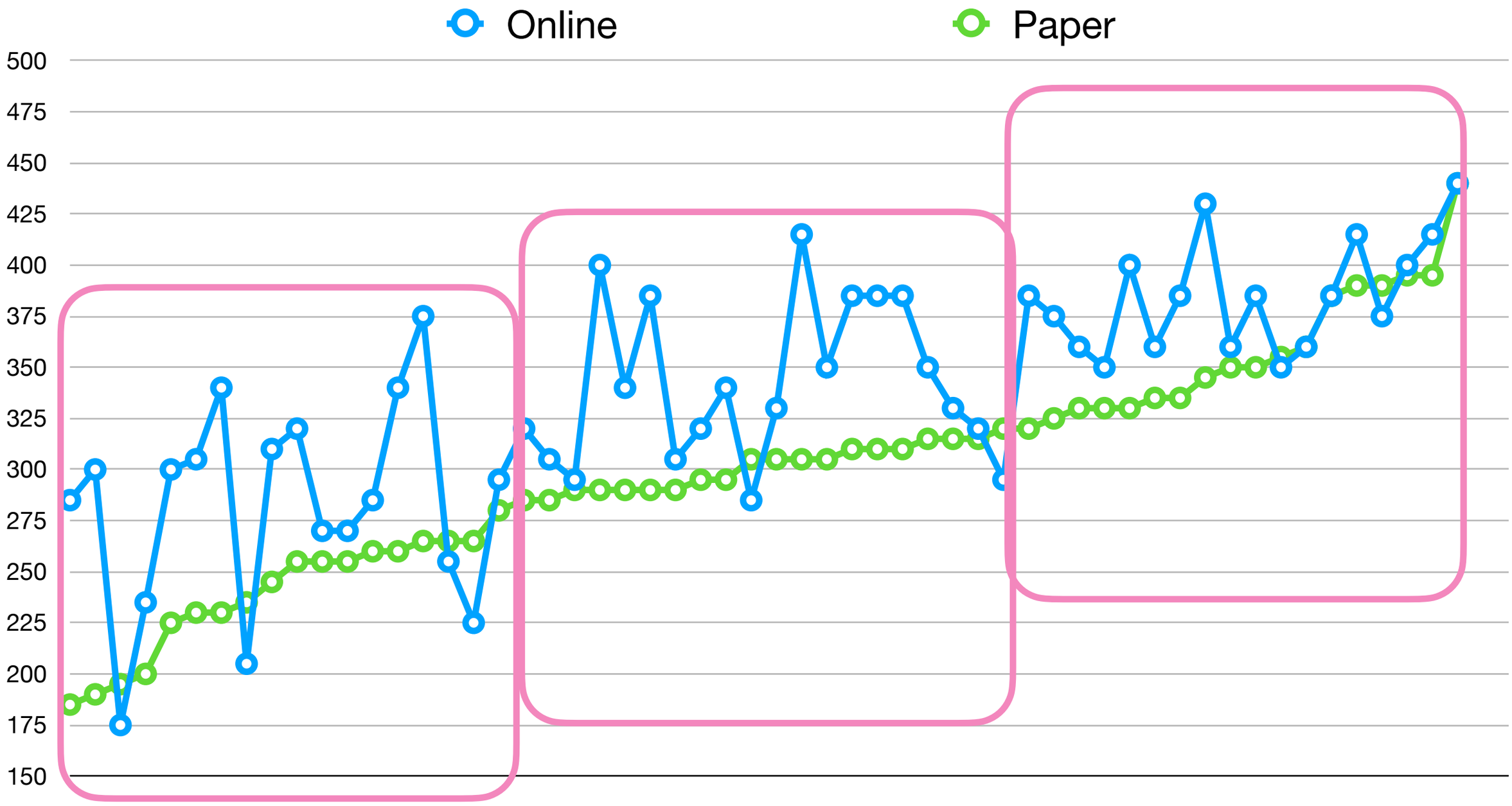
Medium-sized Effect Size

	2021	2022
Listening	○ (Cohen's d = 0.9)	×
Reading	○ (Cohen's d = 0.3)	○ (Cohen's d = 0.6)

Small effect size

長野県立大学
THE UNIVERSITY OF NAGANO

Individual Scores - Listening (2021)



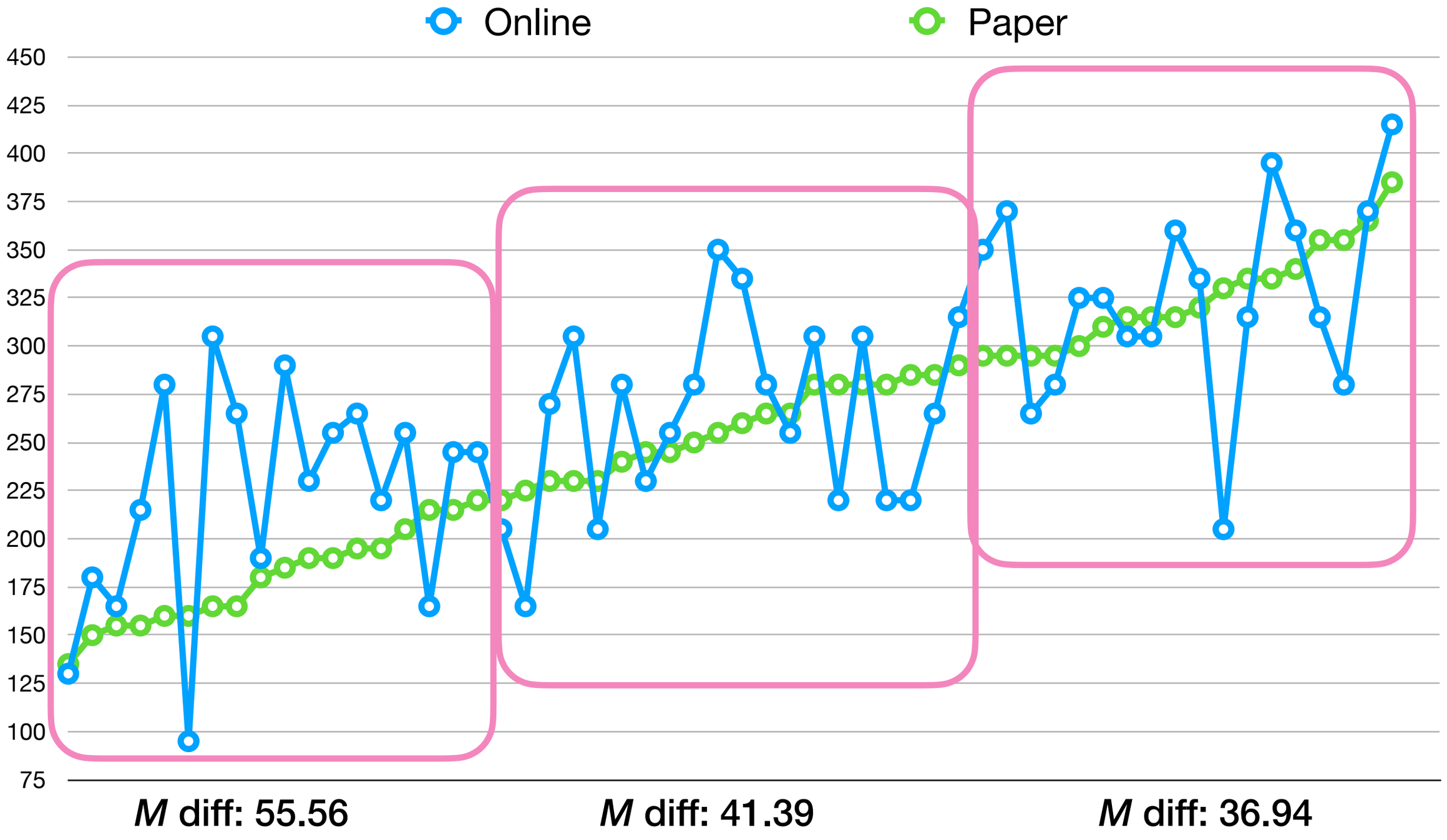
M diff: 55.28

M diff: 46.94

M diff: 29.72

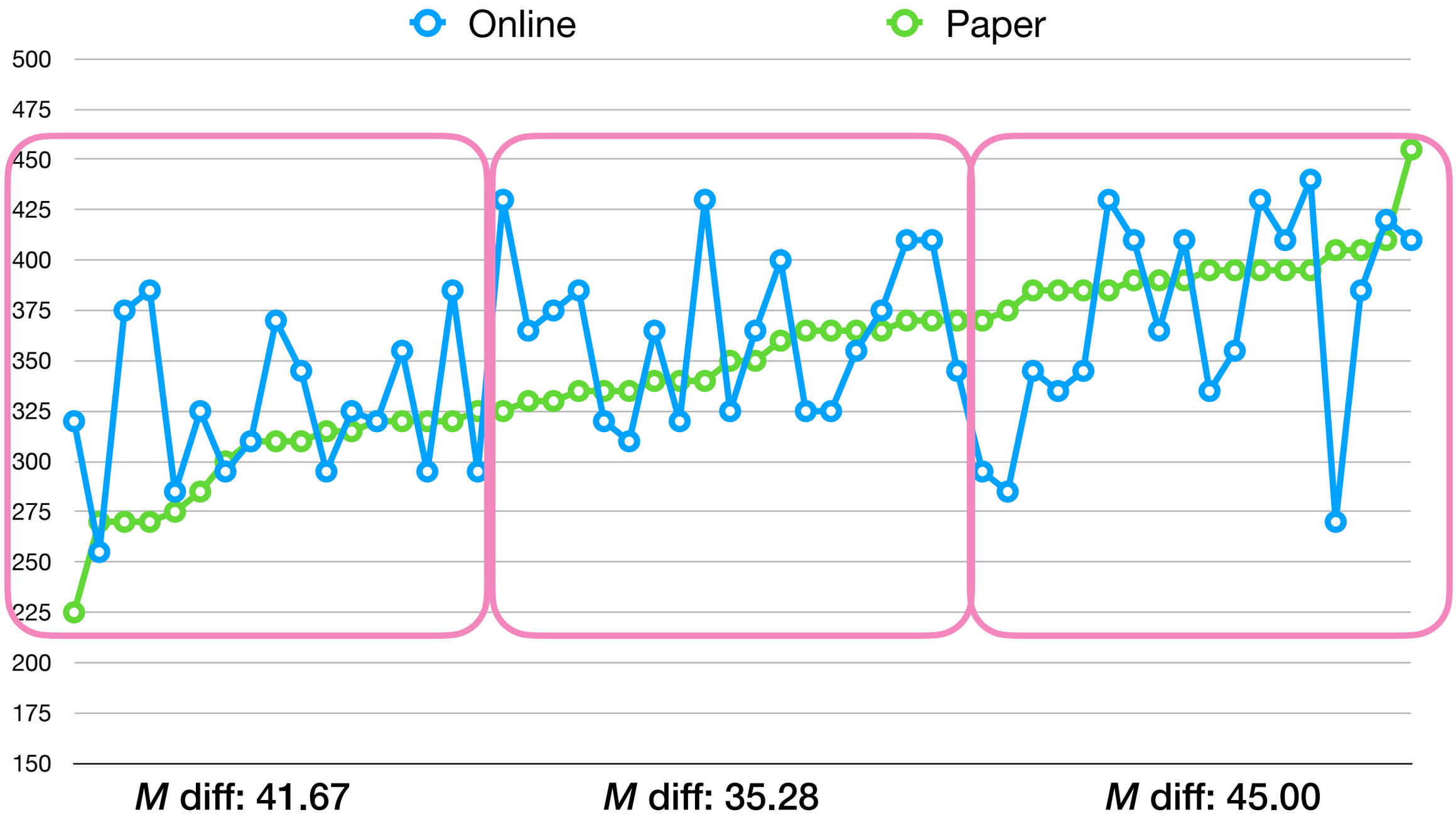
長野県立大学
THE UNIVERSITY OF NAGANO

Individual Scores - Reading (2021)



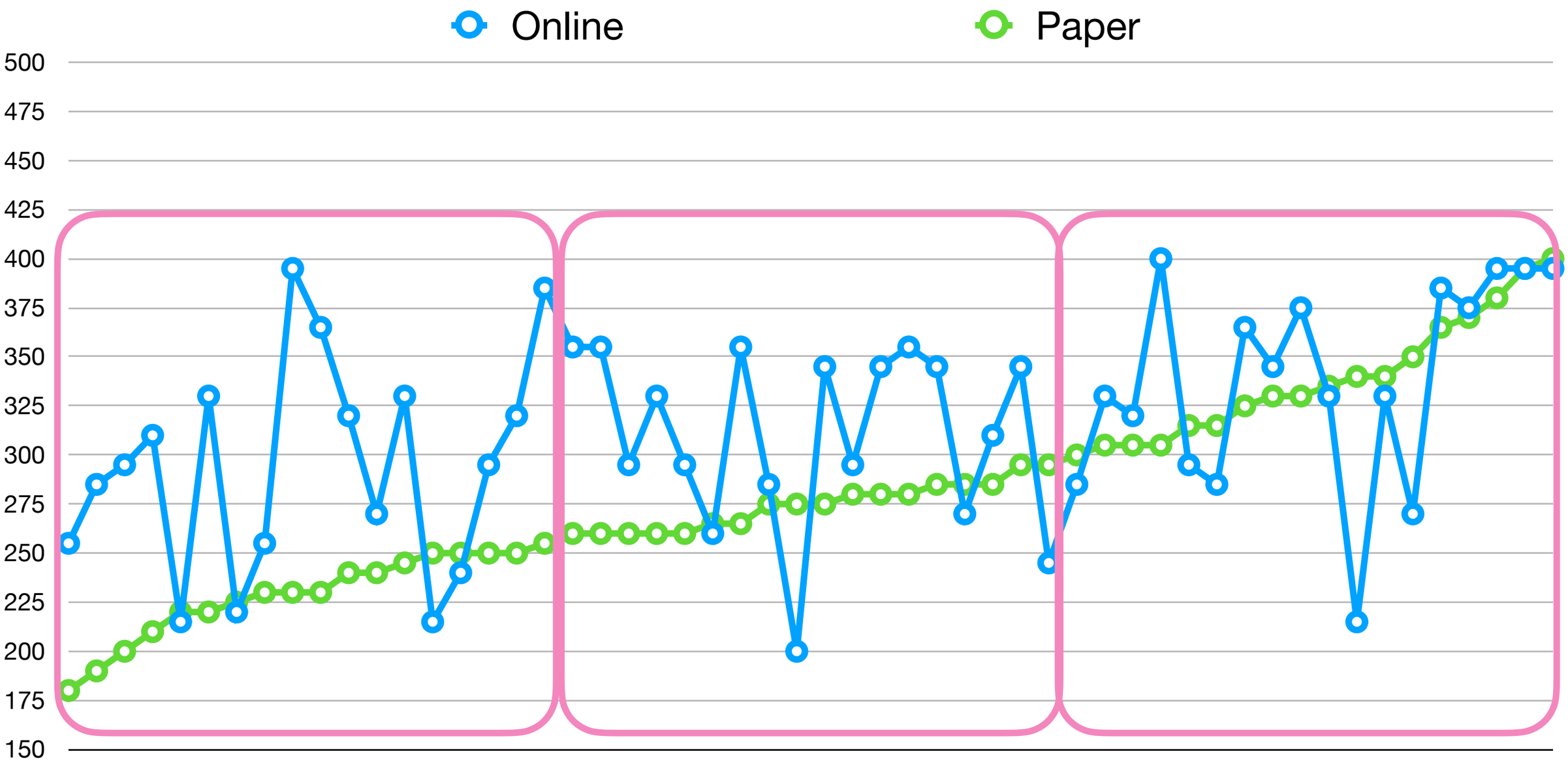
長野県立大学
THE UNIVERSITY OF NAGANO

Individual Scores - Listening (2022)



長野県立大学
THE UNIVERSITY OF NAGANO

Individual Scores - Reading (2022)



M diff: 71.94

M diff: 43.89

M diff: 29.72

長野県立大学
THE UNIVERSITY OF NAGANO

Participants with SE *diff* ±40, ±80

	2021 (*n* = 56)		2022 (*n* = 54)	
Diff	Listening	Reading	Listening	Reading
±40	43%	48%	48%	52%
±80	16%	9%	13%	28%

長野県立大学
THE UNIVERSITY OF NAGANO

Summary

- **Research Question**: Do the paper-based and online TOEIC L&R tests result in parallel scores?
  - **No.**
    - 2021: online listening results > paper-based
    - 2022: online reading results > paper-based
    - Approx. 50% of participants had SE *diff* greater than ±35 points)

長野県立大学
THE UNIVERSITY OF NAGANO

Discussion

- According to IIBC, the two tests are parallel
- However, ETS has yet to publish reports about the online test
- Many questions remain, for example,
  - what is the reliability of the online test?
  - what is the SE *Diff* of the online test?
  - how accurate are the CAT algorithms?

Limitations (1)

- Participants were motivated volunteers.
  - Although not reported, online test mean scores for the participants were higher than the mean scores of all students at Shōzan U.
- Perhaps the results *appear* unstable for these participants but if all students had participated, the results would be less different

Limitations (2)

- In 2021, the online test was completed online, without supervision, and students choose the time to take the test
- In 2022, the online test was completed online, with online supervision, within a set time frame
- For both, it cannot be guaranteed that students did not receive assistance off-camera.

Conclusions & Future Studies

- Conclusion
  - The online and paper versions were found to be statistically different for the participants at Shōzan University
  - Wide variation in students' individual scores were observed
  - This reduces the test validity for these test-takers
- Although burdensome for test-takers, having a large group of students complete the online and paper-based tests upon entering the university would be useful; as would following these students through the 2-year English program at Shōzan University

References

- Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). *Statistical Analyses for the Updated TOEIC® Listening and Reading Test*. Research Memorandum: ETS RM-17-05. ETS. https://www.ets.org/Media/Research/pdf/RM-17-05.pdf
- IIBC. (2020a). TOEIC Program IP テスト(オンライン). https://www.iibc-global.org/toeic/corpo/guide/toeic/online_program.html
- IIBC. (2020b). 特集。場所と時間を合わずに活用できるIIBCのオンラインプログラム。 [Special Feature: IIBC's online program that can be used at any time and place.] https://www.iibc-global.org/iibc/activity/iibc_newsletter/nl141_feature_01.html
- JASP Team (2020). JASP (Version 0.13.1) [Computer software].
- Kanzaki, M. (2017). New and old TOEIC L&R: Score comparison and test-taker views on difficulty level. *PanSIG Journal 2017*, 104-112.

Acknowledgements

- This research project was supported by two grants from the president of *Shōzan University*.

長野県立大学
THE UNIVERSITY OF NAGANO

Thank you.

- If you have any questions, I would be happy to hear from you.
- richard.jean-pierre@u-nagano.ac.jp

長野県立大学
THE UNIVERSITY OF NAGANO