JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

SUOMEN AKATEMIA

EDUCA
EDUCATION FOR THE FUTURE
FLAGSHIP

AI as a mediational means in L2 development: Developing conceptions and practices

Vera Vesala, Dmitri Leontjev, Ari Huhta

AFinLA, November 14, 2026, Oulu

Outline of the presentation

- Starting points & conceptual basis of our DD-LANG project
- Illustration of online dynamic-diagnostic assessment of reading in FL English
- Research findings from spring 2025: **comparing students' views on two different kinds of support during reading exercises**
  - Knowledge of response + mediation based on dynamic assessment + chatbot
  - Knowledge of response + chatbot
- (A classroom intervention for supporting writing with AI)

Aim of the DD-LANG project: to address two issues in assessment that supports L2 learning

- **Lack of systematicity** and **basis in theory** in the (formative) assessment supporting learning & teaching of language skills

→ We address this by bringing together diagnostic and dynamic assessment

- **Lack of individualisation** of assessment and support due to large groups and limited teacher time

→ We address this by implementing computerised, AI-supported dynamic-diagnostic assessment

Diagnostic assessment & Dynamic Assessment

Diagnostic assessment (DiagA)

- Based on theories and research in SLA and assessment (e.g. Alderson 2005; Lee 2015; Alderson, Harding & Brunfaut 2015; Alderson et al. 2015; Huhta et al. 2024)
- Careful definition of language constructs; focus on learner weaknesses
- Main weakness of DiagA: how to use assessment results to support learning/teaching (action)

Dynamic assessment (DA)

- based in Vygotsky's Sociocultural psychology (e.g. Vygotsky 1978, 2012; Lantolf & Poehner 2004, 2014; Pohner 2005, 2008; Poehner & Leontjev 2020)
- Full diagnosis includes not only what learners can do independently but also what they can do with support (= mediation) → emerging abilities / Zone of Proximal Development
- **Mediation** [intervening in cognitive processes] through interaction (hints, leading questions, feedback, model, example)
- Main weakness of DA: how to define constructs

DD-LANG (Dynamic Diagnostic Language Assessment) Project

Research to advance the **theoretical, empirical and practical basis** of foreign language (L2) assessment supporting learning

- funded by the **Research Council of Finland** and University of Jyväskylä Sept.1. 2022 – Aug. 31. 2026
- focuses on **reading** and writing in English as a foreign language in upper secondary schools in Finland (16–18 -year-olds)

2 major goals:

1) Elaborating assessment framework to support teaching & learning by integrating **dynamic** and **diagnostic L2 assessment**
2) **Combining online dynamic-diagnostic assessment and classroom assessment, teaching, and learning**

→ **today's focus is on the online activities**

Online reading exercises in DD-LANG were presented on the Revita platform (U. of Helsinki)

Focus on developing reading in English

1) with **mediation** whenever learner fails to reply correctly and

2) **with AI-based chatbot** (additional support AFTER learner works through mediation, completes item; further learning opportunity)

Online dynamic assessment and mediation in the DD-LANG English reading comprehension tasks

Choose the best fitting answer to the question.

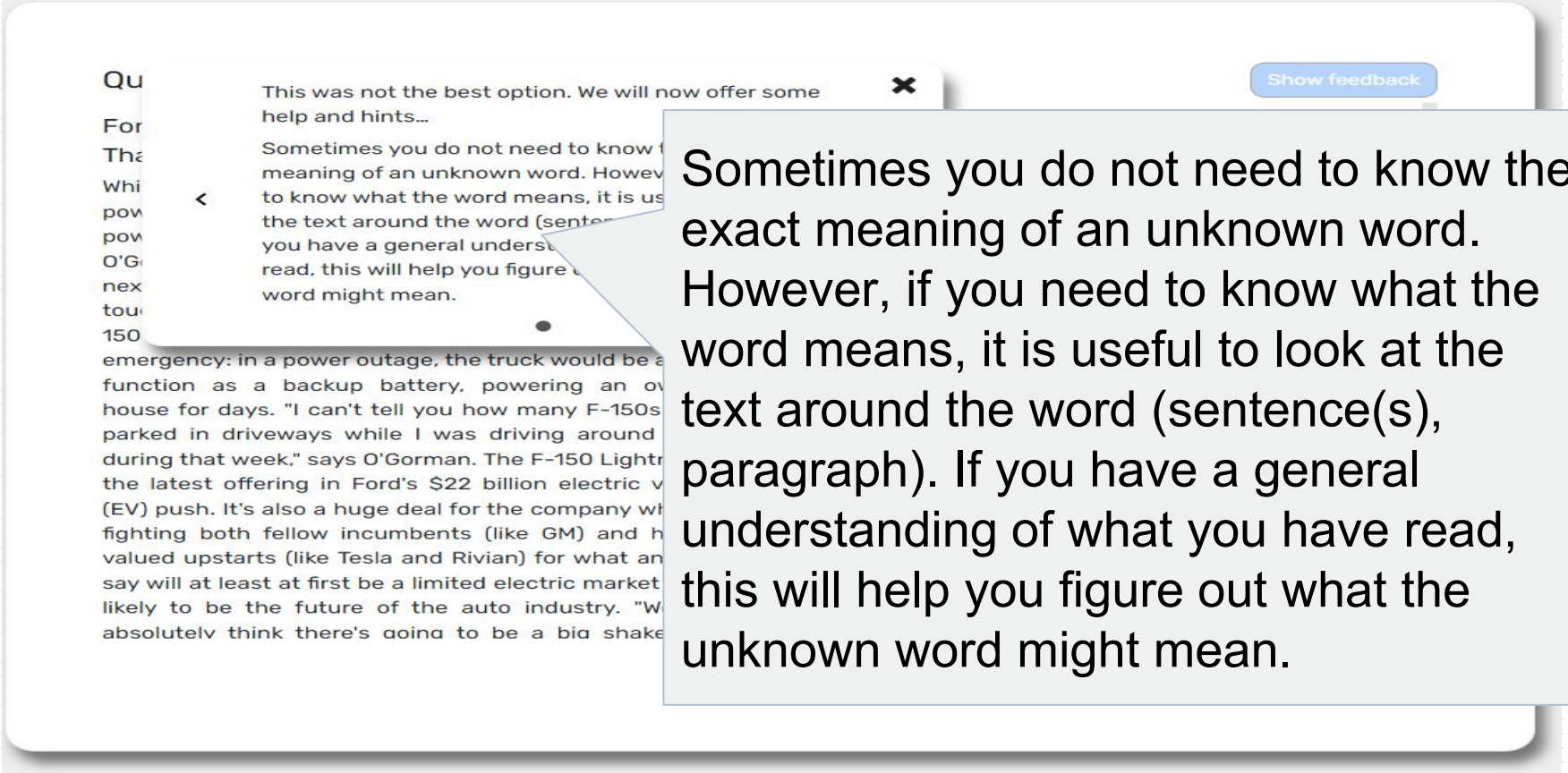
the next one. Ford at the time was putting the finishing touches on an all-electric version of its best-selling F-150 pickup, with a nifty feature for just such an emergency: in a power outage, the truck would be able to function as a backup battery, powering an owner's house for days. "I can't tell you how many F-150s I saw parked in driveways while I was driving around Texas during that week," says O'Gorman.

The F-150 Lightning is the latest offering in Ford's $22 billion electric vehicle (EV) push. It's also a huge deal for the company which is fighting both fellow incumbents (like GM) and highly-valued upstarts (like Tesla and Rivian) for what analysts say will at least at first be a limited electric market but is likely to be the future of the auto industry. "We just absolutely think there's going to be a big shake out," Ryan O'Gorman told TIME of the EV market. "There're going to be clear winners and losers, and the pecking order of the industry is going to change."

What is Ryan O'Gorman's opinion about the future?

- New manufacturers will gain leading positions
- Old manufacturers will suffer bankruptcies
- The future of the car industry is worrisome
- The whole car industry is going to be altered
- There will be new car manufacturers in the US

Mediation in computerised DD-LANG - level 1 mediation for the reading construct 'deducing the meaning of unknown words'



Sometimes you do not need to know the exact meaning of an unknown word. However, if you need to know what the word means, it is useful to look at the text around the word (sentence(s), paragraph). If you have a general understanding of what you have read, this will help you figure out what the unknown word might mean.

Full 3-level example of standardised mediation: Reading for specific details

Level 1: Look at the key parts of the question and the options. Think about what the question is asking. Then try to find this information in the text, thinking about how you search for it.

Level 2: Once you think you found the place with the details you need, **look around this part** and think about the **meanings of words and phrases** you see. Try to connect them to the question and the options. Remember that the details in the text are often said in a different way than in the question.

Level 3: You can also look for:

- The order of details, paying attention to **verb tenses** and words like *first* and *then*.
- How these details are connected, like **comparing** using *likewise*, or **contrasting** using *but*.
- Sometimes you need to choose the details that best match the question and the option.
- You can also use your knowledge about the topic to find the answer.

What is the relationship between learners' independent and mediated performance in FL English reading?

Why is information about learners' mediated score potentially useful?

1) It is diagnostic information about learners' skills that are maturing (not yet mastered without support; Zone of Proximal Development)
2) Teachers can pair / group students for classroom exercises not only based on their independent skill level but also on their ZPD (mixing or matching learners depending on the pedagogical goal of the exercise)

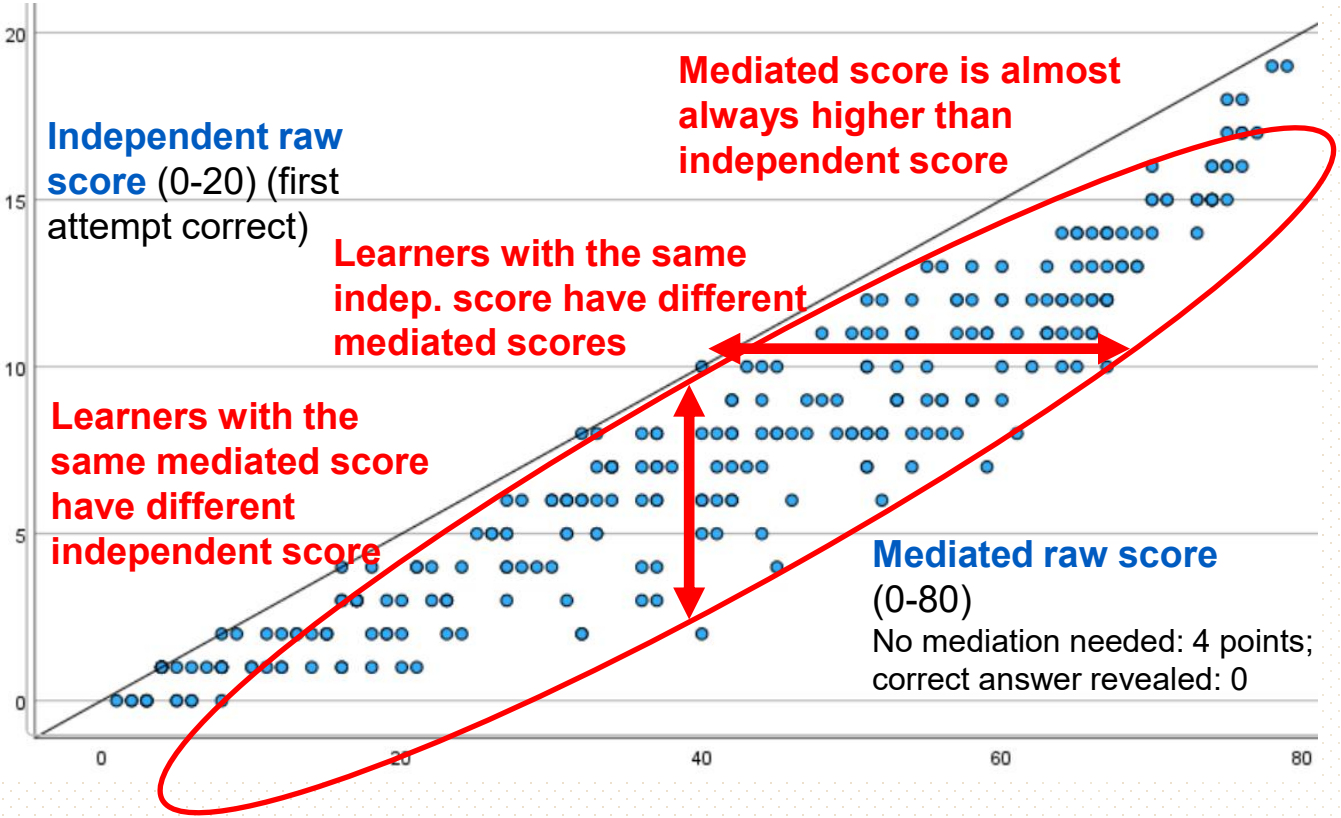
What do we know about the relationship between learners' independent and mediated performance in FL English reading

Spring 2025 DD-LANG data from 280 learners

Pearson correlations:

Raw scores (independent x mediated): .952 →90.5% shared variance

Rasch scores (independent x mediated): .923 → 85.2% shared variance

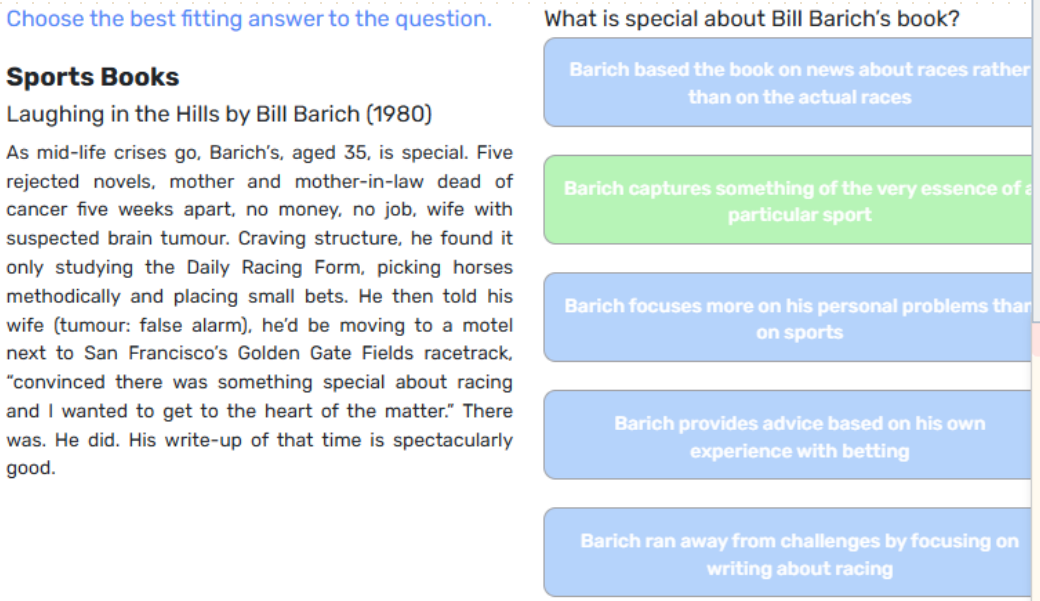


Learner performance reports on their independent and mediated reading skills → basis for classroom activities

Username	Set	item_count	attempt_count	correct_on_first_attemp	correct	mediation_score	avg_mediation_wo_correct
S1	**Specific details, easier**	**4**	**4**	**4**	**4**	**4**	**0**
	Specific details, difficult	5	5	5	5	4	0
	Deducing vocabulary easier	1	1	1	1	4	0
S2	**Specific details, easier**	**5**	**13**	**1**	**5**	**2.4**	**2**
	Specific details, difficult	5	9	3	4	3.2	2
	Deducing vocabulary easier	5	7	3	5	3.6	1
	Deducing vocabulary difficult	5	15	0	5	2	2
S3	**Specific details, easier**	**5**	**16**	**0**	**3**	**1.8**	**2.2**
	Specific details, difficult	5	7	3	5	3.6	1
	Deducing vocabulary easier	2	3	1	2	3.5	1

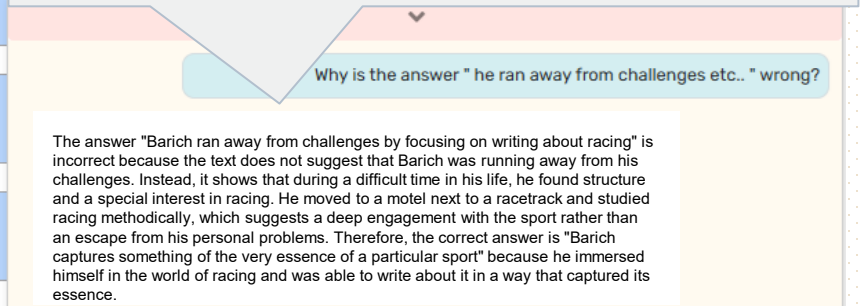
Chatbot in the online DD-LANG English reading comprehension tasks: a supporting / complementary approach to mediation

The chatbot appears after the correct / best answer is selected (an example of interaction between learner & bot)



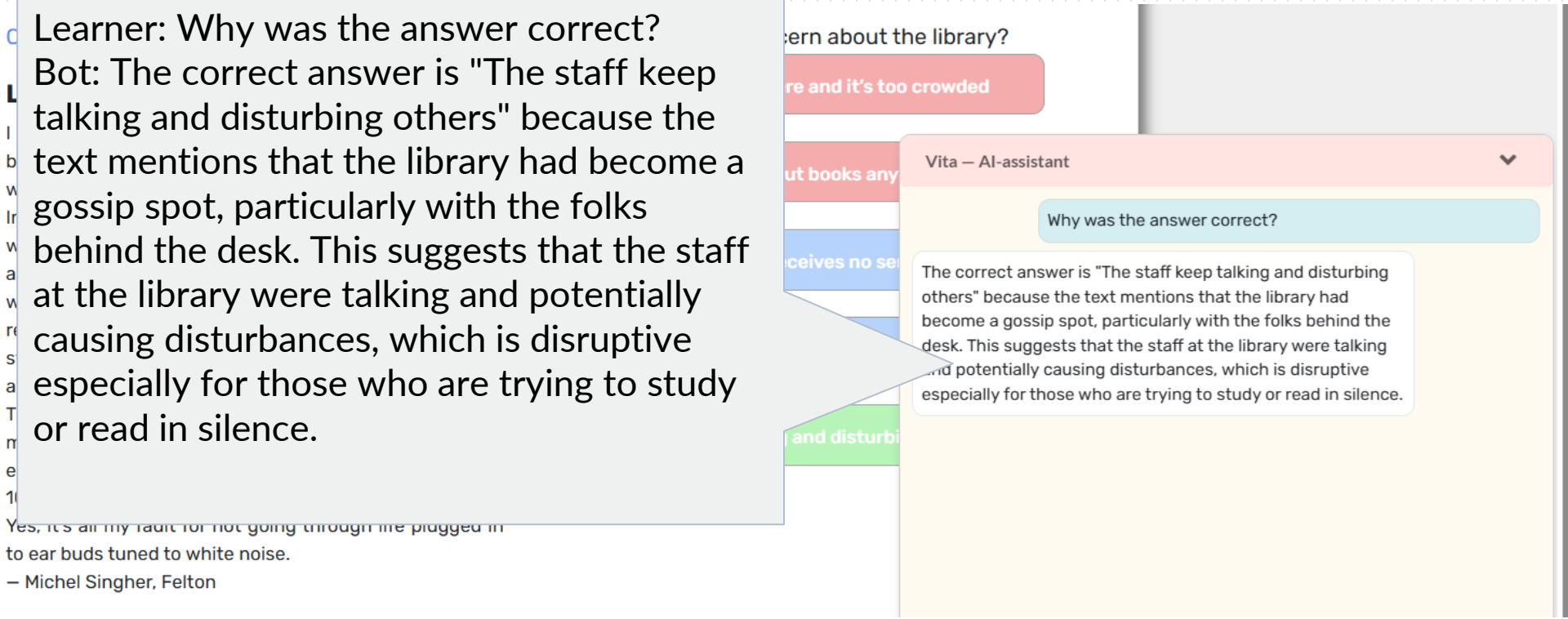
Learner: Why is the answer " he ran away from challenges etc.. " wrong?

Chatbot: The answer "Barich ran away from challenges by focusing on writing about racing" is incorrect because the text does not suggest that Barich was running away from his challenges. Instead, it shows that during a difficult time in his life, he found structure and a special interest in racing. He moved to a motel next to a racetrack and studied racing methodically, which suggests a deep engagement with the sport rather than an escape from his personal problems. Therefore, the correct answer is "Barich captures something of the very essence of a particular sport" because he immersed himself in the world of racing and was able to write about it in a way that captured its essence.



II Chatbot interaction

Learner: Why was the answer correct?
Bot: The correct answer is "The staff keep talking and disturbing others" because the text mentions that the library had become a gossip spot, particularly with the folks behind the desk. This suggests that the staff at the library were talking and potentially causing disturbances, which is disruptive especially for those who are trying to study or read in silence.



A study within DD-LANG comparing two types of support for reading

How did the students perceive the comprehensibility and usefulness of the help / feedback provided by mediation and the chatbot? Did they feel they learned anything about reading and reading tasks?

- We compared students who **took the same DD-LANG English reading tasks** under two different conditions
  - **Experimental** groups: Knowledge of response + mediation based on dynamic assessment + chatbot
  - **Control** groups: Knowledge of response + chatbot
- In both groups, the chatbot was activated only after the student had answered correctly or after the system revealed the answer, which happened either after all 3 levels of mediation had been presented (experimental group) or after two unsuccessful attempts (control group)
- Students replied to an **online questionnaire** (in Finnish) immediately after taking the tasks

Comparison of two conditions

- Experimental groups (n = 71) – response correctness + mediation + chatbot
  - School A: 27 students
  - School B: 44 students
  - Mostly second year students
  - Mean grade on the latest course: 7.88

- Control groups (n = 74) – response correctness + chatbot
  - School A: 74 students
  - Both first and second year students
  - Mean grade on the latest course: 8.16
- Groups did not differ significantly in terms of their course grades or performance on the DD-LANG reading tasks

Did the groups differ in how clear and useful they considered item/task level feedback?

*Task specific feedback: How **clear / understandable** did you find the information on the scale of 0 (not clear at all) to 10 (very clear)?*

GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 54)	6.15	2.30	-7.10	**p<.001**
Control (n = 74)	8.73	1.81		effect size 1.27

*Task specific feedback: How **useful** did you find the information on the scale of 0 (not useful at all) to 10 (very useful)?*

GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 53)	5.57	2.32	-5.49	**p<.001**
Control (n = 73)	7.82	2.25		effect size .99

Did the groups differ in how clear and useful they considered feedback on the whole set of reading tasks?

*If you had time to complete all the tasks, you received a short report on the percentage of tasks you got right: How **clear / understandable** did you find it on the scale of 0 (not clear at all) to 10 (very clear)?*

GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 25)	7.52	2.49	-1.69	p = .10
Control (n = 69)	8.43	1.75		

*How **useful** did you find the final report on the scale of 0 (not useful at all) to 10 (very useful)?*

GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 31)	6.16	2.35	-2.85	**p = .005**
Control (n = 68)	7.46	1.97		**effect size .618**

Examples of learners' positive and negative comments on item-specific and overall feedback

EXPERIMENTAL:

Joskus palaute oli täysin samaa, kun vastasi väärin johonkin kohtaan, joka ei juurikaan auta

Mielestäni saatu palaute oli hyvää, mutta vielä parempaa olisi jos jokainen tehtävä saisi oman tehtäväkohtaisen ohjeen. Toki ymmärrän että tämä on valtavan työllistävää mutta heitin vain ajatuksen ilmoille.

Olin alkuun skeptinen ja hieman negatiivinen, mutta lopussa kiinnitin enemmän huomioita palautteeseen ja ne olivat aika hyviä

CONTROL:

Palaute oli ihan hyvä

Palaute oli ihan hyödyllistä

Did the groups differ in how often they used the chatbot and how useful they perceived it to be?

How often did you ask the chatbot anything on the scale of 0 (not even once) to 10 (in every task)?

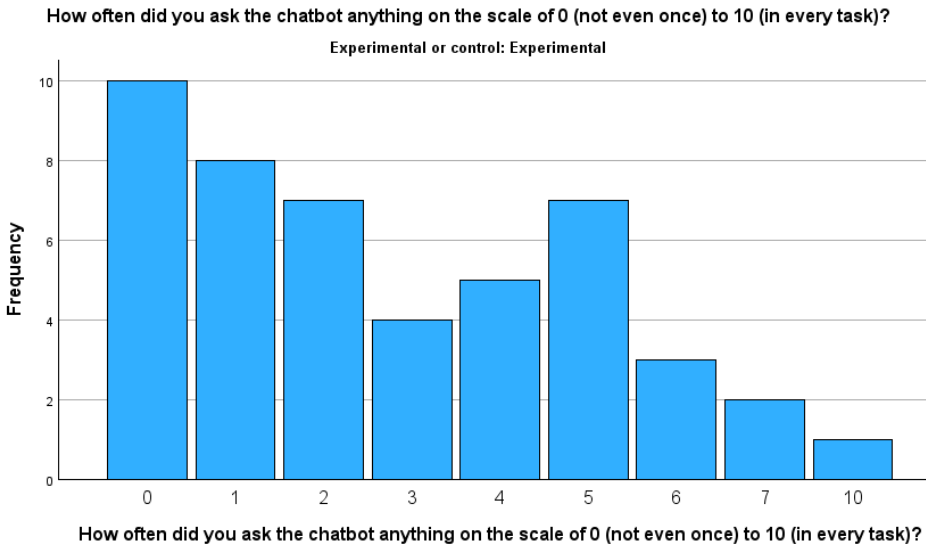
GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 47)	2.79	2.41	-1.80	p =.075
Control (n = 67)	3.58	2.26		

If you used the chatbot, how useful were its responses on the scale of 0 (not useful at all) to 10 (very useful)?

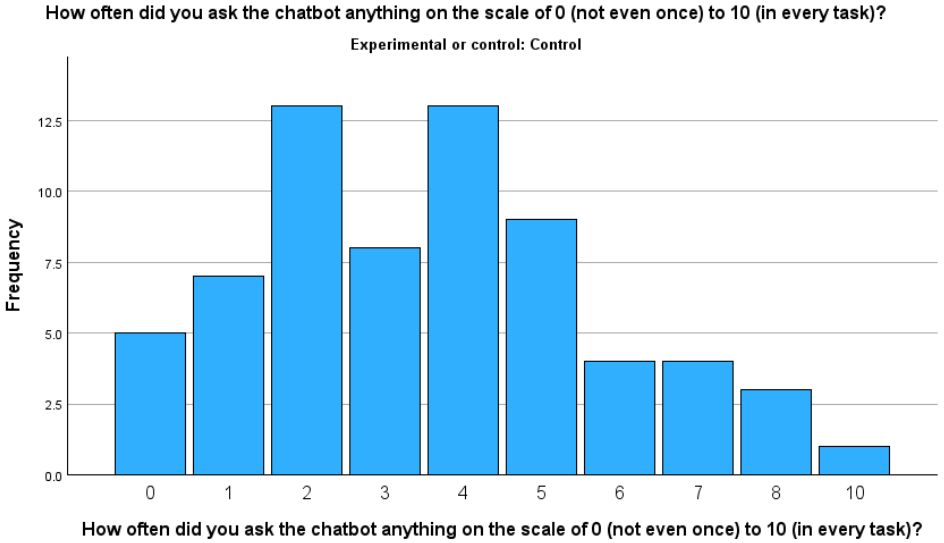
GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 34)	5.38	2.75	-1.35	p =.182
Control (n = 62)	6.10	2.33		

Frequency of using the chatbot

Experimental groups



Control groups



Examples of learners' comments on what was useful vs not useful in the chatbot's responses

EXPERIMENTAL: detailed extensive responses; explaining was helpful; translation

Vastaukset olivat aina pitkiä ja laajoja, joten ne selittivät hyvin tarvittavan vastauksen.

Se auttoi bongaamaan tekstistä oikean vastauksen kannalta oleellisimmat tiedot.

En ymmärtänyt joitain sanoja, jotka oli pakko ymmärtää, niin pyysin chatbotia kääntämään ne.

Mielestäni vastaukset siitä miksi joku vastaus oli oikein olivat aika pintapuoleisia eikä oikein auttanut oikeasti ymmärtämään miksi se oli parempi

CONTROL: explaining was helpful (rather similar as in the experimental groups)

Pyysin, että se selittää miksi joku vastaus oli juuri oikea vaikka muutkin vaikuttivat hyviltä. Se kertoi hyvin miksi juuri se ja oli hyödyllinen

Kertoi miksi jokin ei ollut oikein.

Se ei ynmärtänyt kaikkea, vaikka mielestäni olin selkeä.

Did the groups differ in their views on whether they learned anything about TASKS or about READING?

*Did you learn anything from the **tasks** that could help you, for example, in the English matriculation examination on the scale of 0 (I did not learn anything new) to 10 (I learned a lot of new things)?*

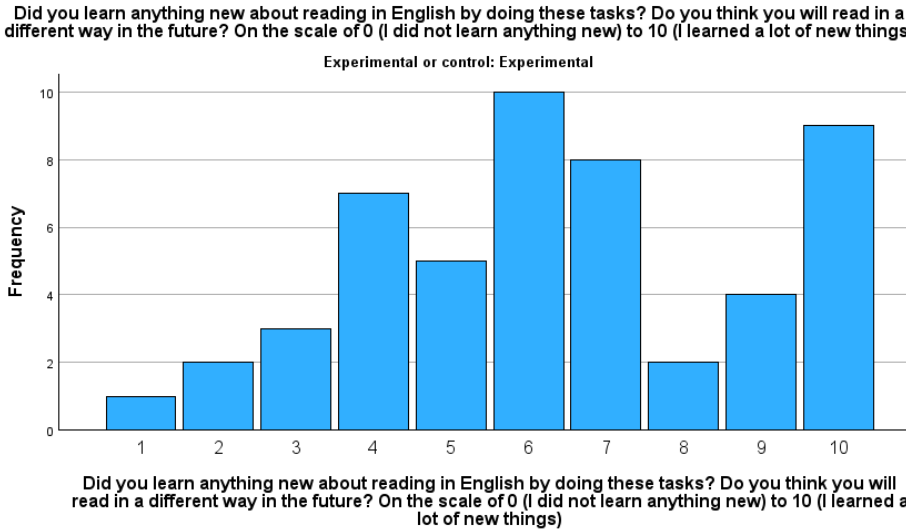
GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 52)	6.56	2.11	1.55	p = .124
Control (n = 63)	5.97	1.97		

*Did you learn anything new about **reading** in English by doing these tasks? Do you think you will read in a different way in the future on the scale of 0 (I did not learn anything new) to 10 (I learned a lot of new things)?*

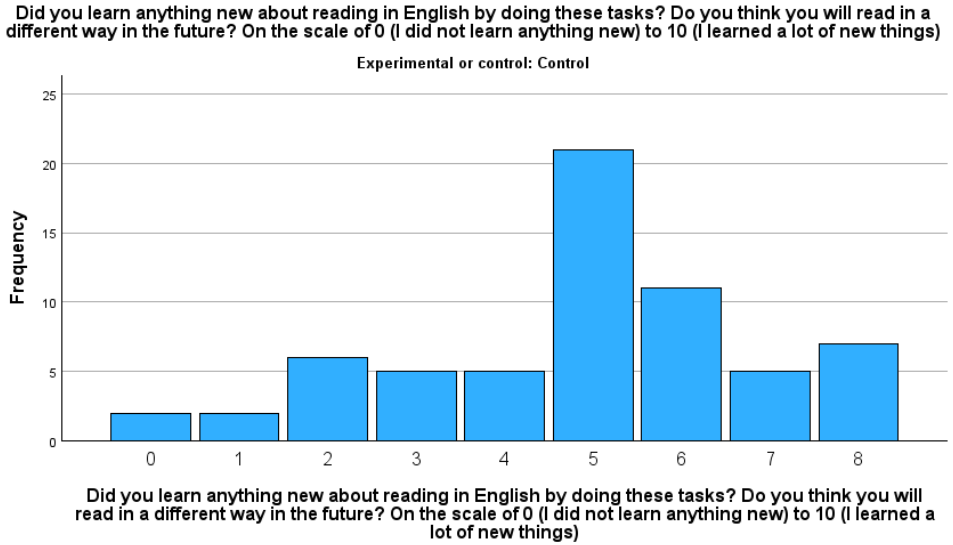
GROUP	MEAN	ST.DEV	T-VALUE	SIGNIFICANCE
Experimental (n = 51)	6.37	2.47	3.54	**p = <.001**
Control (n = 64)	4.86	2.01		effect size .68

Learning something new about reading

Experimental groups



Control groups



Examples of what the students thought they learned about reading and for the Matriculation exam

EXPERIMENTAL: more responses; more detailed examples; about strategies

Opin mistä luetunymmärtäminen kannattaa aloittaa ja mihin tekstissä kannattaa keskittyä.

Opin tulkitsemaan tekstejä enemmän kokonaisuutena enkä yksittäisinä faktoina.

Opin esimerkiksi katsomaan tarkemmin ensimmäistä ja viimeistä virkettä eikä keskittyä turhaan.

Sen että yksityiskohtiin on hyvä kiinnittää paljo huomiota

CONTROL: few responses; rather general examples in nature; task-focused comments

Keskittymään paremmin

Opin, että luetunymmärtämiset ovat vaikeita ja niihin pitää keskittyä.

Opin esimerkiksi miettimään ehkä vähän syvällisemmin ja tarkemmin niitä tekstejä, että mikä olisi oikea vastaus

Conclusions

Type of feedback / information from (reading) tasks:

- Traditional feedback (right / wrong; overall score) is familiar → many learners believe they understand what it means and many think it's useful
- Learners are not sure what to think about very different kinds of feedback such as mediation, and how useful it is (see also feedback on self-assessment in DIALANG; Huhta 2010)

→ Many learners need support to benefit from mediation → teacher support, classroom activities

Chatbot:

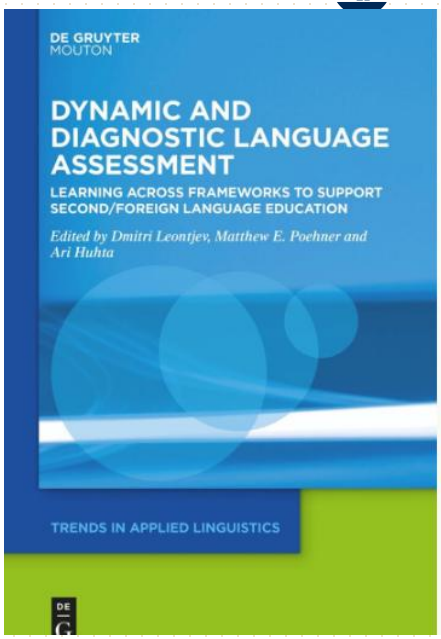
- Most learners used the bot but not all (esp. when they received mediation)
- Mixed reactions but usually somewhat positive (was considered useful)
- Without further guidance, learners ask the bot only about the task (text, options)

What did the students think they learned?

- Mediation seemed to improve (some) learners' **reading** (specific examples were given); only using the chatbot seemed to lead to less systematic re-thinking about e.g. reading strategies

Kiitos! Thank you!

Ari Huhta	ari.huhta@jyu.fi
Dmitri Leontjev	dmitri.leontjev@jyu.fi
Vera Vesala	vera.e.vesala@jyu.fi
WWW:	**http://r.jyu.fi/DDLANG_en**



NEW OPEN ACCESS BOOK: Leontjev, D., Poehner, M. & Huhta, A. (eds.) (2025) *Dynamic and diagnostic language assessment. Learning across frameworks to support second/foreign language education*. De Gruyter Mouton.

Some references

Alderson, J. C. 2005. *Diagnosing Foreign Language Proficiency*. London/New York: Continuum.

Alderson, J. C., T. Brunfaut & L. Harding. 2017. Bridging assessment and learning: A view from second and foreign language assessment. *Assessment in Education: Principles, Policy & Practice* 24(3). 379–387. https://doi.org/10.1080/0969594X.2017.1331201

Huhta, A. (2010). Innovations in diagnostic assessment and feedback: An analysis of the usefulness of the DIALANG language assessment system. Dept. of Applied Linguistics, U. of Jyväskylä. https://urn.fi/URN:ISBN:978-952-86-0718-2

Leontjev, D.; Huhta, A. & Koskela, K. 2024. "I could learn more words": Involving stakeholders in defining L2 English reading constructs in diagnostic and dynamic assessment. *AFinLA Yearbook 2024*. pp 186-210.

Leontjev, D., Poehner, M. & Huhta, A. (eds.) (2025) *Dynamic and diagnostic language assessment. Learning across frameworks to support second/foreign language education*. De Gruyter Mouton.

Leontjev, D., Poehner, M., Huhta, A., & Pollari, P. (2025). Reading with a Chatbot - The added value of Generative AI as a resource in mediating learners in Dynamic Assessment of L2 English reading. *Studies in Language Assessment*, *14*(2), 123-155. https://doi.org/10.58379/LSFJ4641

Poehner, M., & Leontjev, D. (2020). To correct or to cooperate: Mediational processes and L2 development. *Language Teaching Research* 24(3). 295–316. https://doi.org/10.1177/1362168818783212

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.