EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School & Research Network

Leibniz-Institut für
Wissensmedien

Towards reflecting student heterogeneity in adaptive systems for authentic school contexts

Detmar Meurers
Leibniz-Institut für Wissensmedien (IWM)

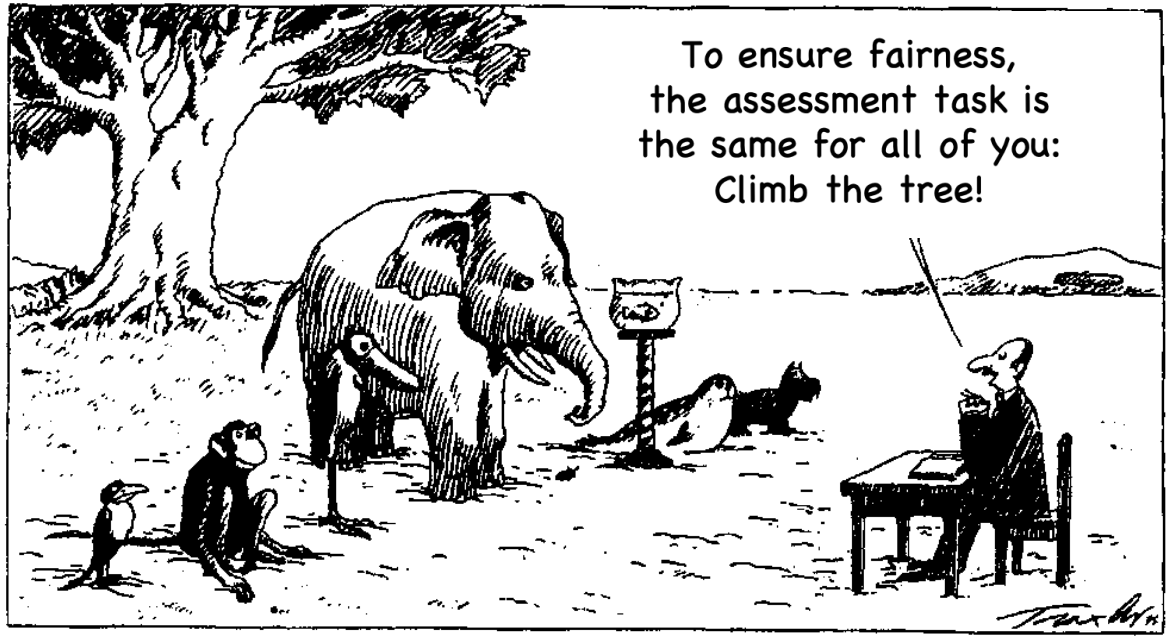
based on the collaboration in my
Language & AI in Education lab

19th European Conference on Technology Enhanced Learning (ECTEL 2024)
Krems, Austria, 16–20 September 2024

Heterogeneity

- Learners differ substantially (cf. Höhmann 2009; Trautmann & Wischer 2011) in their
  - domain knowledge and competencies
  - language competence
  - cognitive characteristics
  - interests and motivation
  - socio-cultural background
- Relevant from several perspectives:
  - appreciation of diversity
  - fairness of tests

  ⇒ What is individually conducive to learning? (aptitude-treatment interactions, . . .)



Classic by Hans Traxler (*betrifft: erziehung, 6/1975*, our translation)

Heterogeneity is an important challenge for teachers

- "Improving the way we deal with heterogeneity is probably the central challenge for modernizing the education system."
  Jürgen Baumert (2002), head of the German Pisa study 2000 (our translation)
- Teachers perceive differences in the talents of students as central occupational challenge: Germany 55%, Japan 63%, USA 44% (TIMSS study, Baumert et al. 1997)
- Supporting each student is named as the most important goal to strive for by practically all teachers, but 90% consider this impossible in practice and do not feel up to it.
  (Study at 18 secondary schools, Kunze & Solzbacher 2008)
- Internal differentiation is not frequently used in school practice.
  (Study with 295 teachers from different types of schools, Letzel 2021)

Adaptive assessment methods as a solution?

- Computer-adaptive testing (CAT) offers individualized optimization of test procedures
- Aim of CAT is to provide the most precise diagnostics with as few test items as possible
  - CAT offers an efficient selective measurement instrument of one-dimensional competencies
  - provides no support for learning processes and multidimensional learner differences
  - good test items discriminate strongly (= small difference in competence makes solving impossible)

↯ Adaptive learning systems need incremental activity sequences that learners are mostly able to complete with support (= differences in competence still allow solving).

→ Adaptive learning systems cannot be reduced to adaptive test procedures.

- Adaptive learning also should not be reduced to testing a student for what they already know, followed by teaching them only the parts they have not yet mastered.
  - This would reduce the multi-dimensional heterogeneity to differences in subject-domain knowledge.

How can adaptive learning be supported in real-life schools?

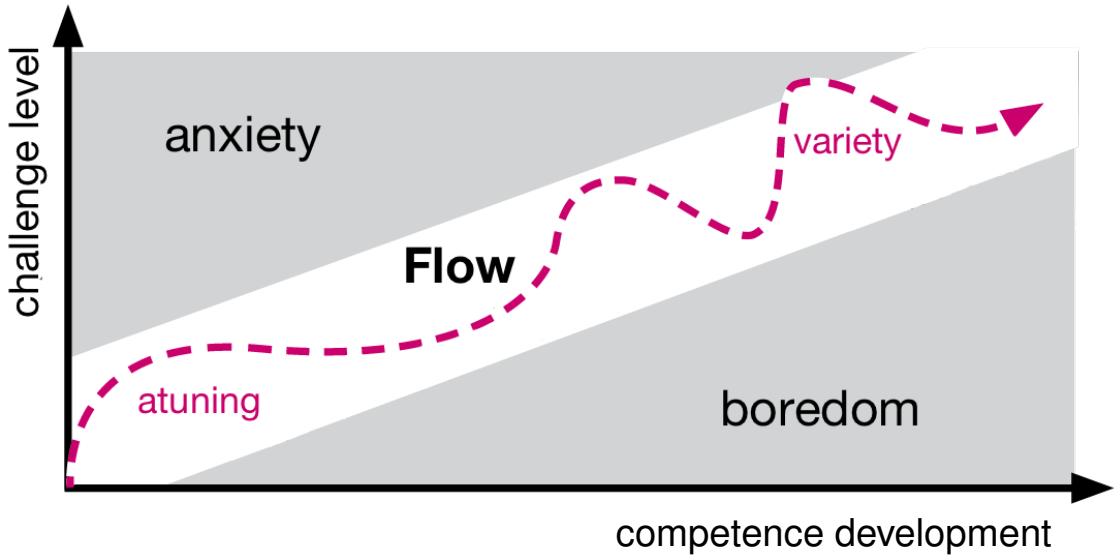
- It cannot be achieved by teachers all by themselves:
  - There is a lack of adaptive materials, time and competence for **diagnosis**, **learning task selection** and **individual scaffolding** of the learning process.
- Potential of digital media:

  *Digital media hold great potential for shaping new teaching and learning processes, especially when considering the possibilities for **individual support for pupils**.*
  (German KMK strategy paper “Education in the Digital World” 2017 & SWK update 2021)

$\Rightarrow$ goal: digitally support individuals in their learning trajectory adaptively

Conceptualizing adaptivity

- Learning as Flow (Csikszentmihalyi 2000) in the Zone of Proximal Development (Vygotsky 1986)



- What do we need to realize such adaptive learning paths (in a multi-dimensional space)?

What is needed to adaptively support learners?

- Individual digital support needs AI methods:
  - automatic analysis of **language** (in learning activities and learner responses)
  - **learner models** of individual competencies
  - **activity models** reflecting their properties and relation towards curricular goals
- Adaptive activity sequencing requires explicit **operationalization of adaptivity**: What is developmentally proximal given learner & activity models and which support?
- We need more research towards answering this question! As a starting point, to adaptively foster learning, we need varied activities that match learners in terms of
  - cognitive,
  - language, and
  - learning-domain complexity.

Relevance of language and domain complexity

- Mathematical word problems differ in mathematical task complexity (operation, carry, . . .):

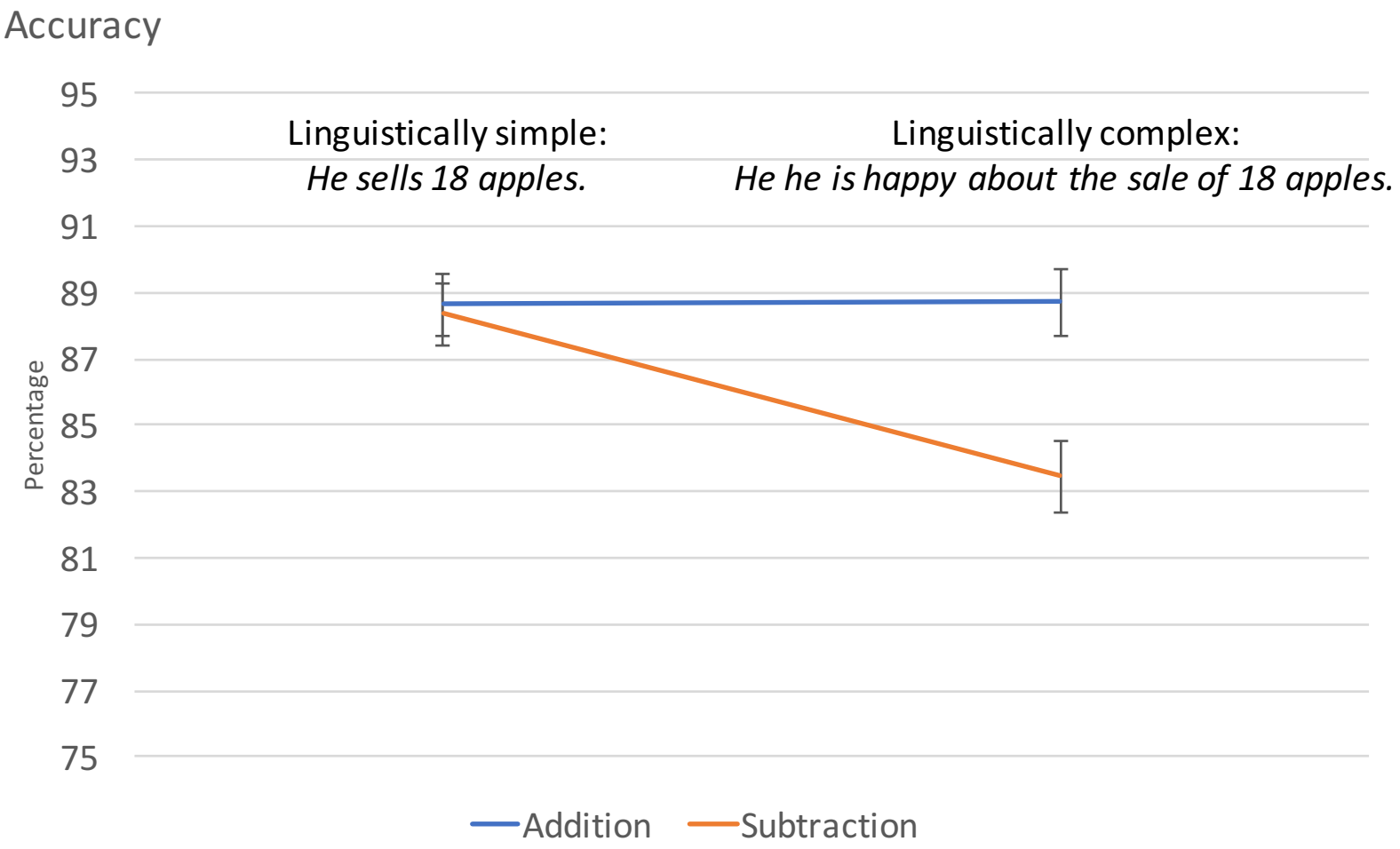
  (1) *A farmer went to market. He arrived with 47 apples.*
  a. *He sells 5 apples.*
  b. *He sells 18 apples.*
  How many apples does he have left?

- They also differ in their language complexity (here: nominal style):

  (2) *a.* *He sells 18 apples.*
  b. *He is happy about the sale of 18 apples.*

⇒ What effect does the combination of domain and language complexity have? (Daróczy, Wolska, Meurers & Nuerk 2015; Daróczy 2021)

There is a need for multidimensional modeling of activities and learners



N=331, school children (mean age=10,91, SD=1,26)

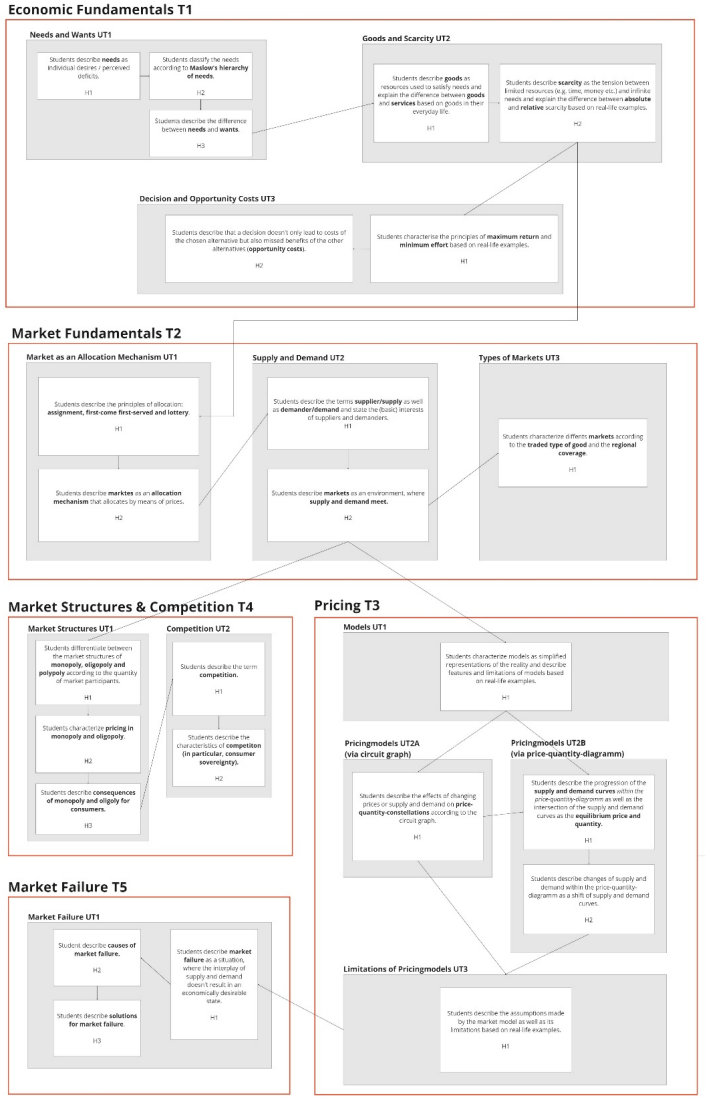
Realizing adaptive learning trajectories in school

- Project ALEE: Adaptive Learning in Economics Education (http://alee.schule)
- interdisciplinary collaboration of
  - subject domain didactics: Institute for Economic Education (IÖB), University of Oldenburg
    Felix Dietrich, Michael Koch, Dirk Loerwald
  - language in education & computational linguistics: University of Tübingen
    Kordula De Kuthy, Detmar Meurers
  - machine learning & software engineering: University of Lüneburg
    Kai Neubauer, Ulf Brefeld

$\Rightarrow$ Develop adaptive activity sequencing complementing regular teaching of economics subject in German secondary schools (grade 8–10).

Foundation: Domain model

- Developed domain model based on school curricula.
  - Curricular topics have subtopics containing learning goals.
  - Each learning goal is a node in the curriculum network.
- For each learning goal, there are
  - an explanatory text and
  - at least ten base activities.
- Per base activity, a harder and an easier variant were created using several difficulty parameters.

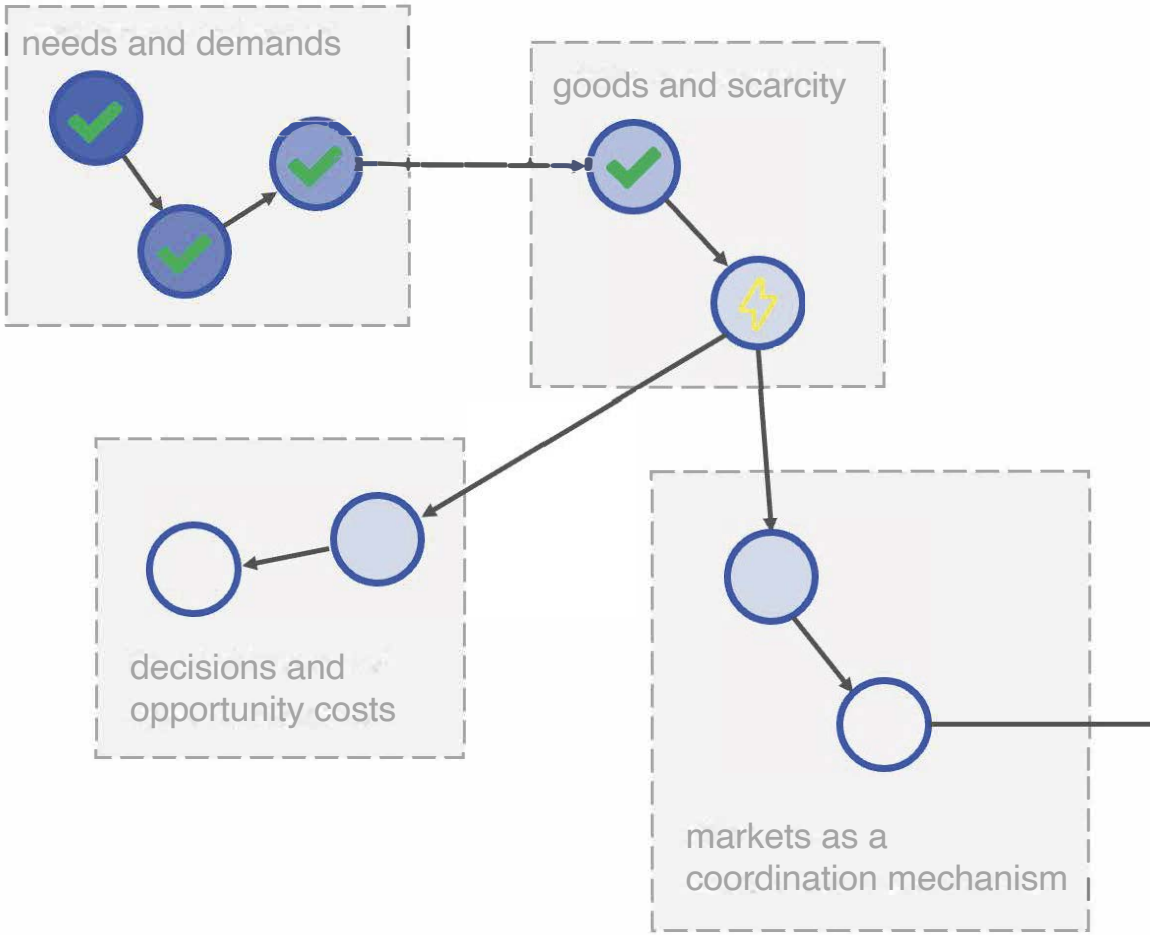


Difficulty parameters defining task complexity space

- **Subject-domain parameters**
  - domain-specific complexity (e.g., *goods* vs. *scarcity*)
  - didactic operator level (*describe/define* vs. *apply/transfer*)
- **Language parameters**
  - explicitness of instruction
  - linguistic complexity (e.g., nominalization, passive, negation)
  - relation of answer to text (verbatim form vs. underlying meaning)
- **Cognitive parameters**
  - activity type (multiple choice, mapping-tasks, fill-in-the-blank, . . .)
  - number of correct answer options
  - number of distractors
  - how distractors are falsifiable (in text, world knowledge, domain knowledge)

Adaptive activity assignment

- estimation of learner competence and activity difficulty using probabilistic ELO rating model
- assignment of activities within a learning goal based on these ratings
- transition to next learning goal based on performance on reference activities



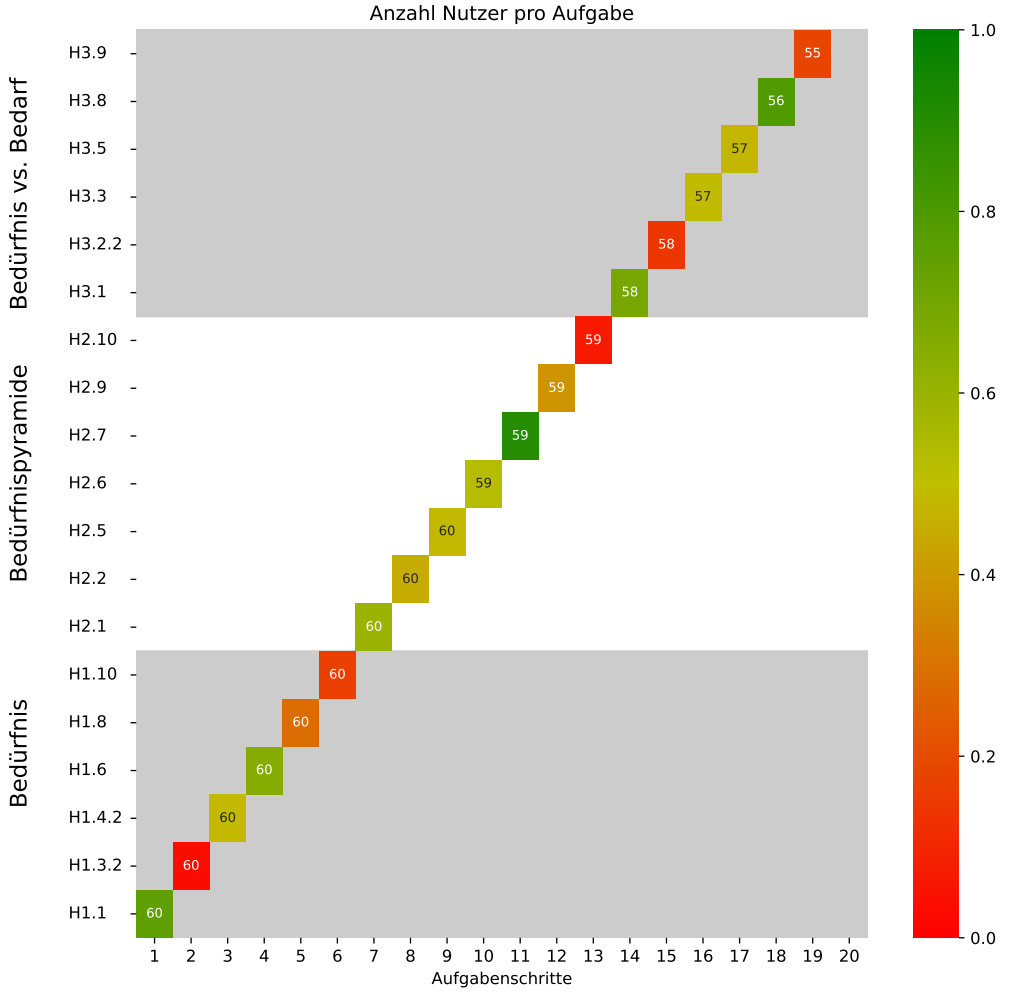
ALEE field study

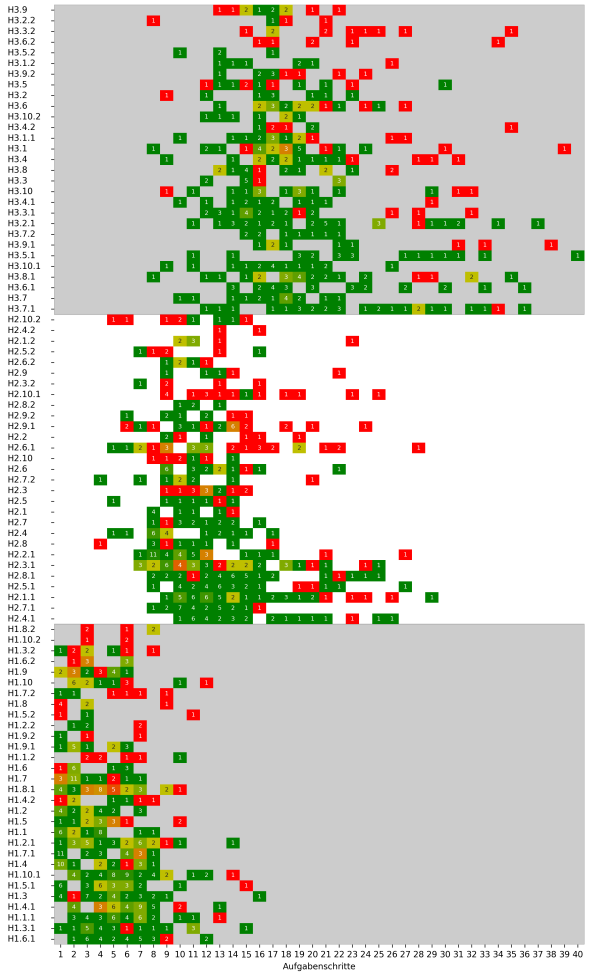
- randomized controlled field study in authentic secondary school context:
  - five week intervention as part of regular economics subject classes
  - eleven classes, grade 7–10 in academic and non-academic track schools in Lower-Saxony and Bavaria
- random student assignment within each class to
  - adaptive activity selection group
  - static activity sequence control group
- n = 156 students (80 adaptive, 76 static)

  $\rightarrow$ learning process data: $>$ 15k activity completions for the $\approx$ 700 activities for the 16 learning topics

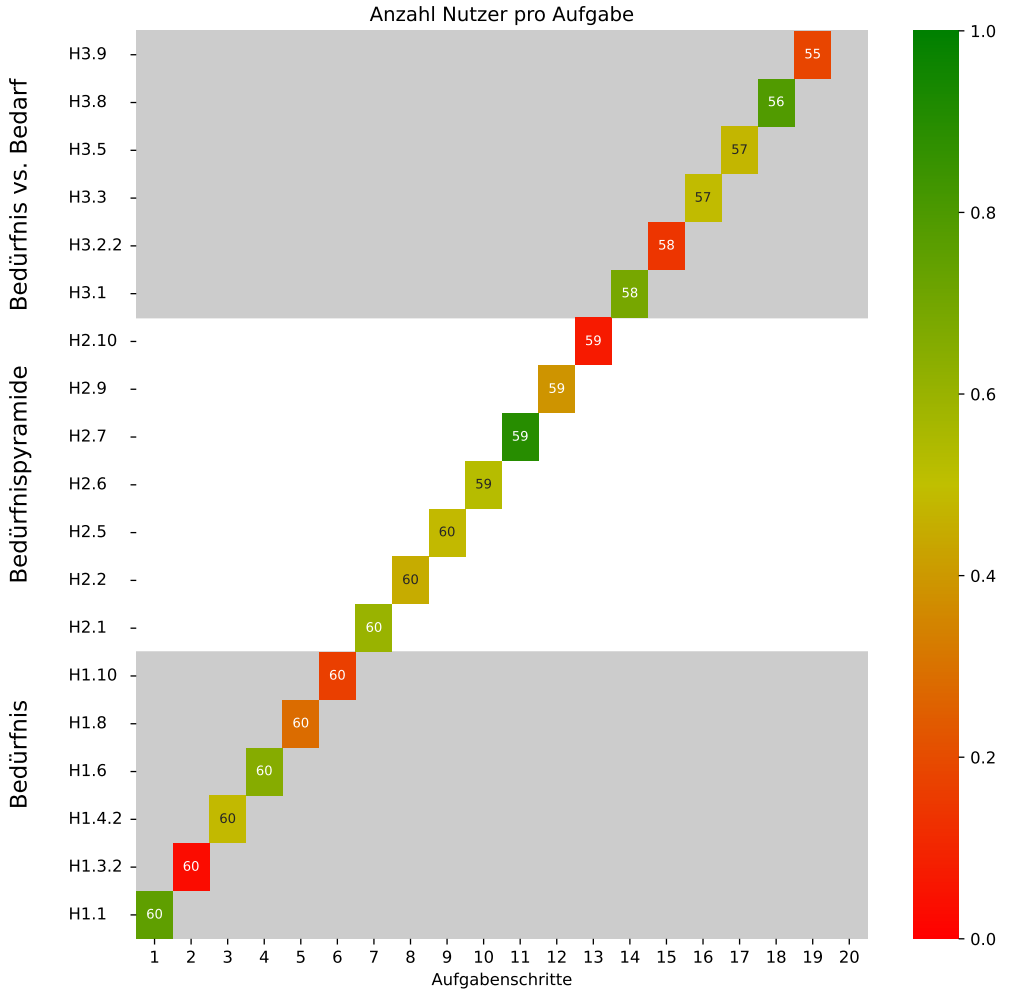
  $\rightarrow$ questionnaires at beginning and end of study (student interests, motivation, self-efficacy, grades, . . .)

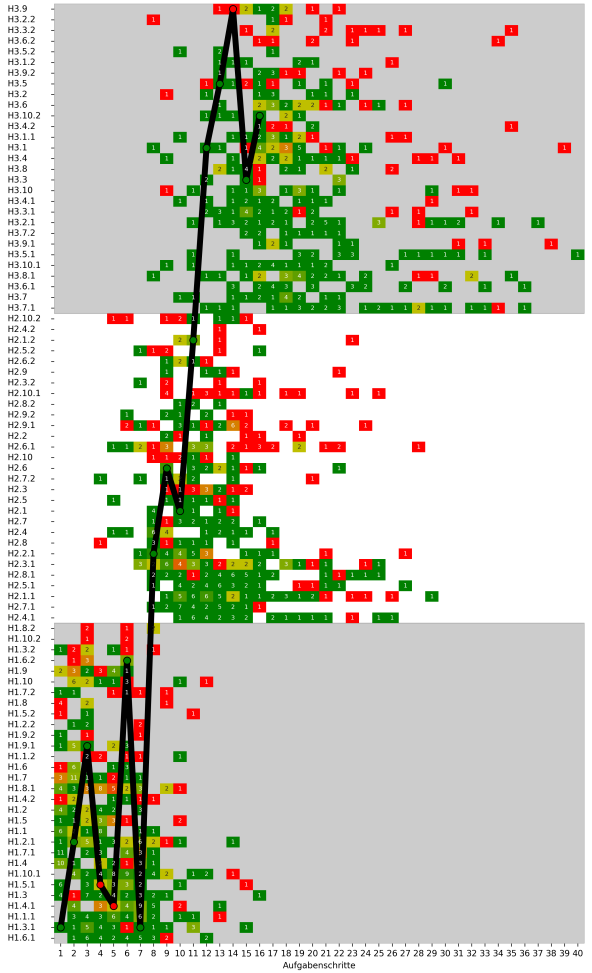
Usage of the standard and the adaptive activity space



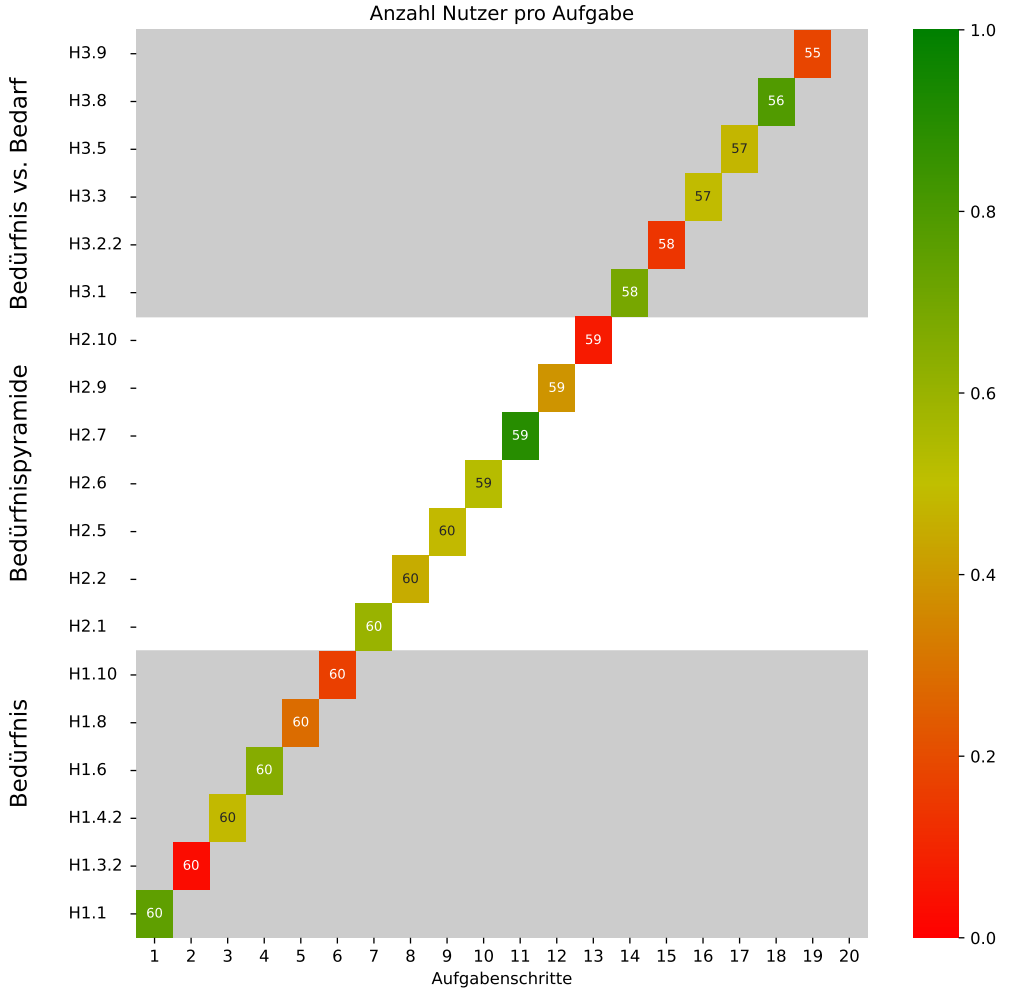


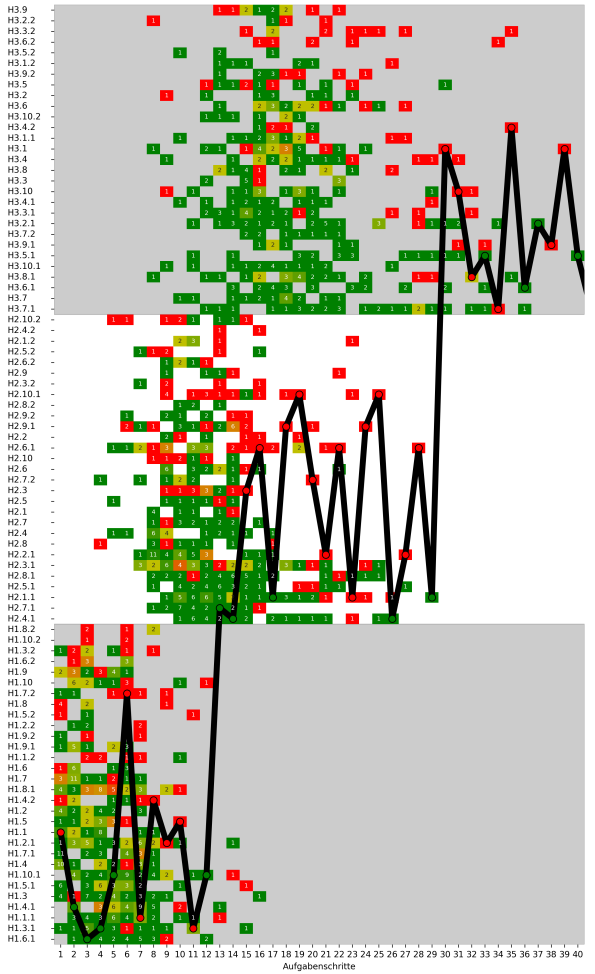
Standard sequence vs. a short learner trajectory in the adaptive space





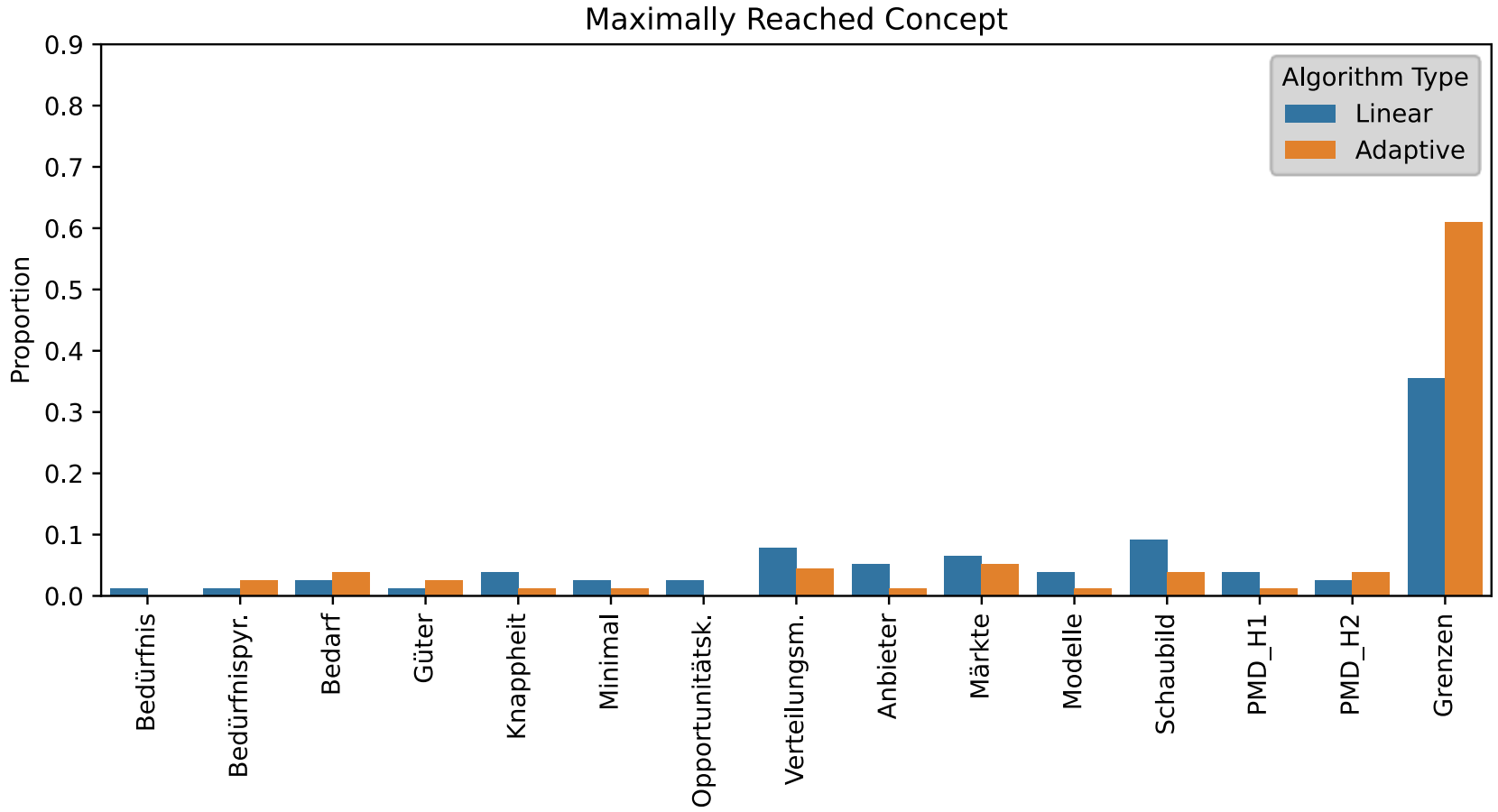
Standard sequence vs. a long learner trajectory in the adaptive space



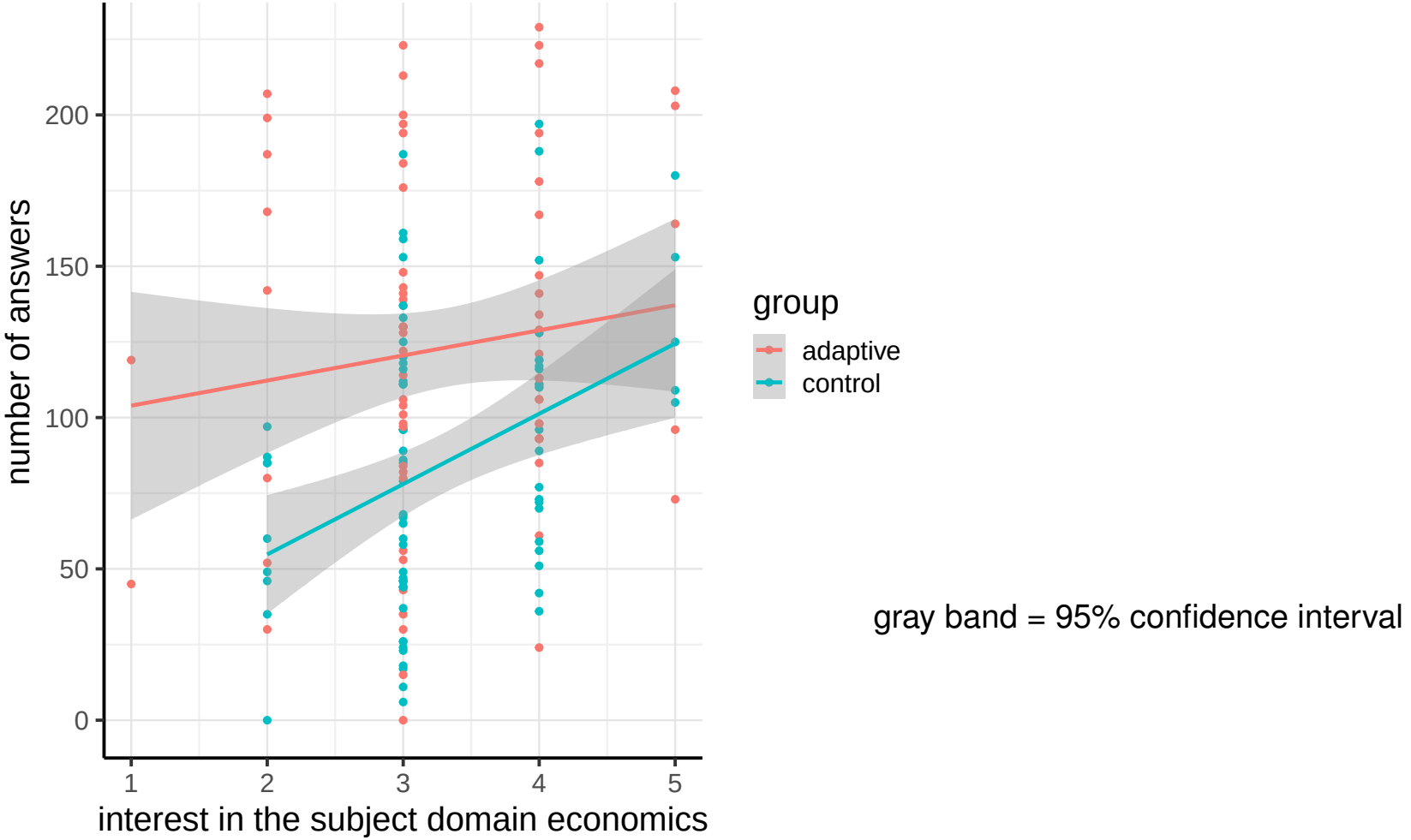


How much did they learn?

→ Students in both groups learned, but students in the adaptive group got significantly further:

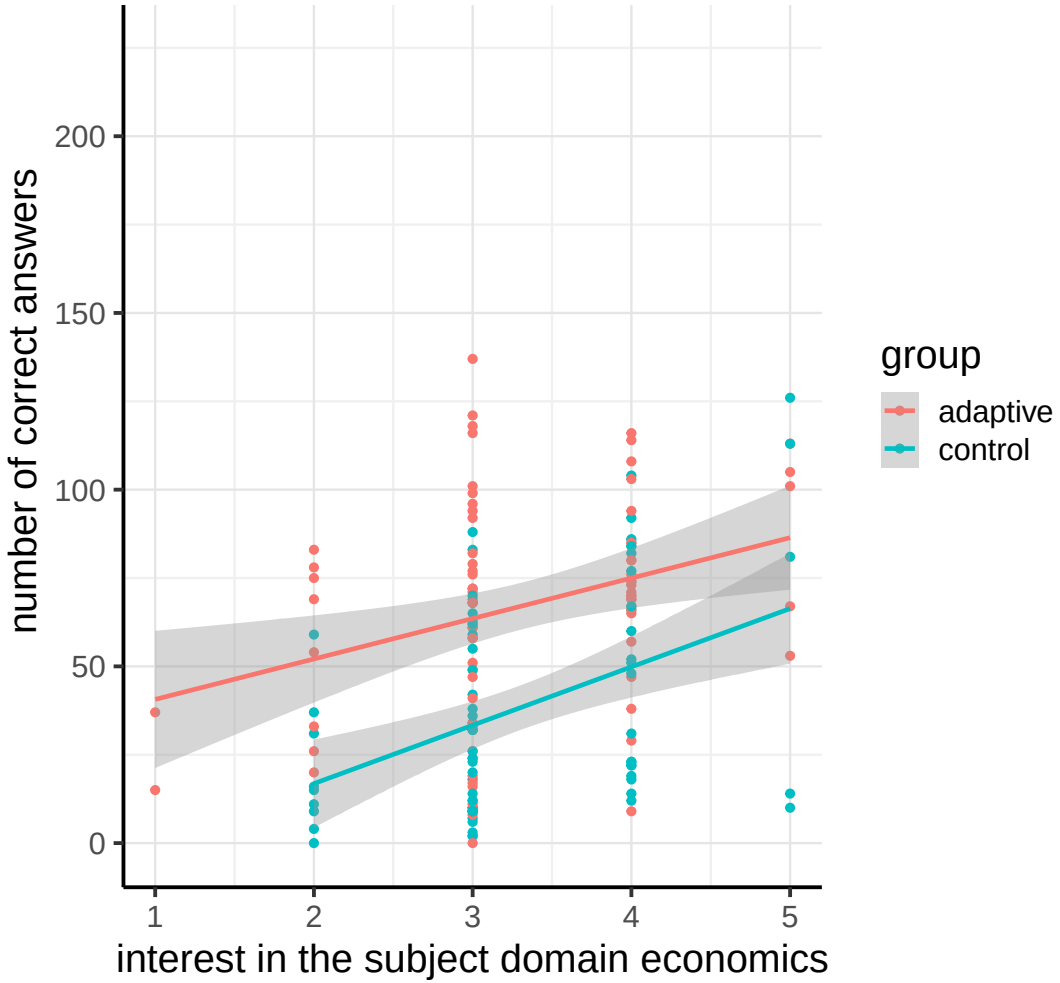


Relating domain interest to engagement (# activities completed)

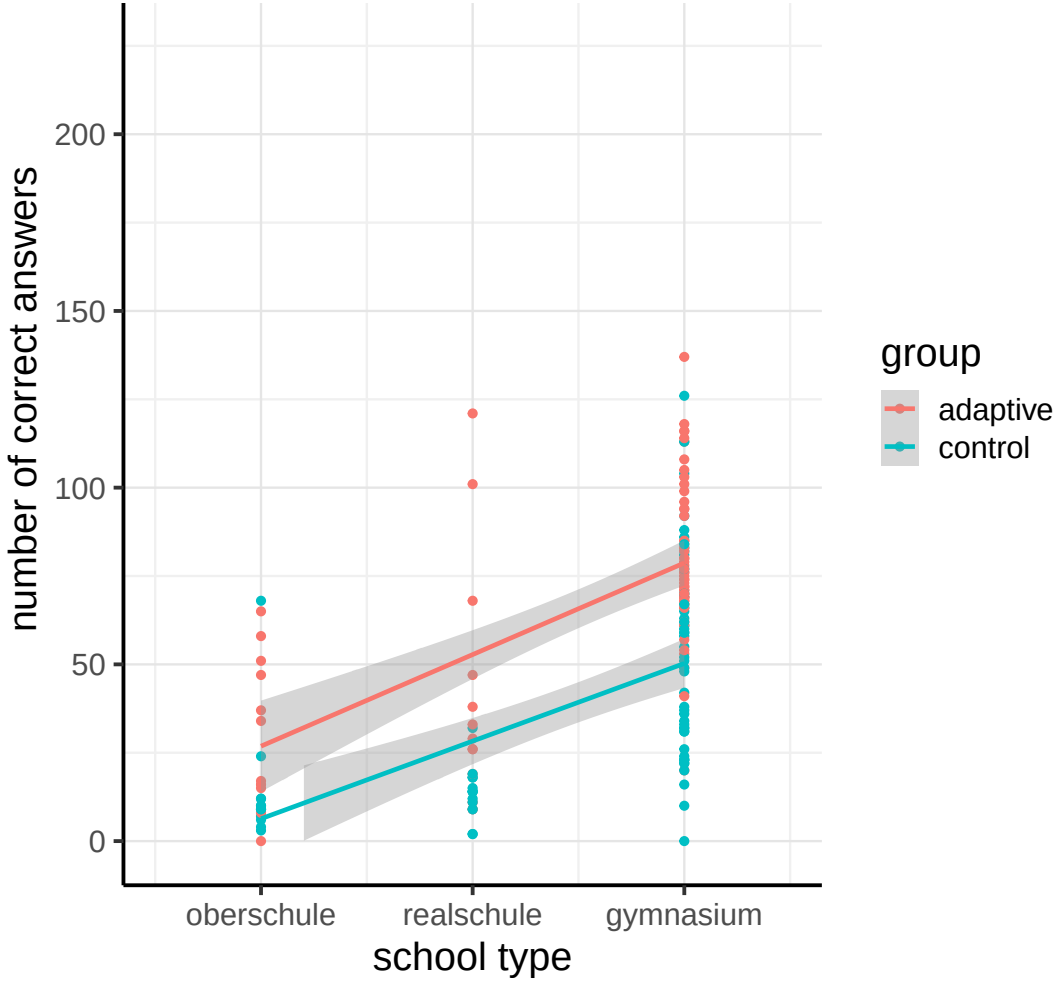


gray band = 95% confidence interval

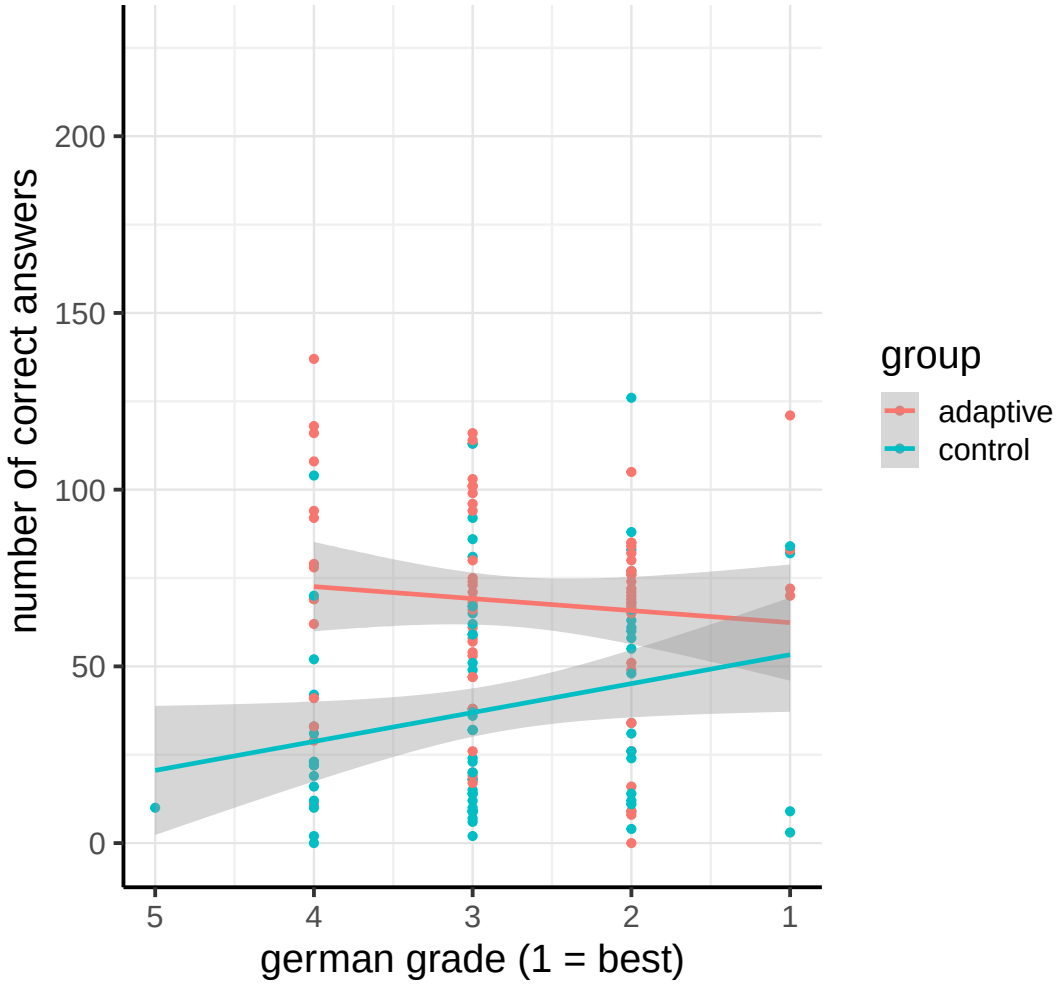
Relating domain interest to learning success (# activities completed correctly)



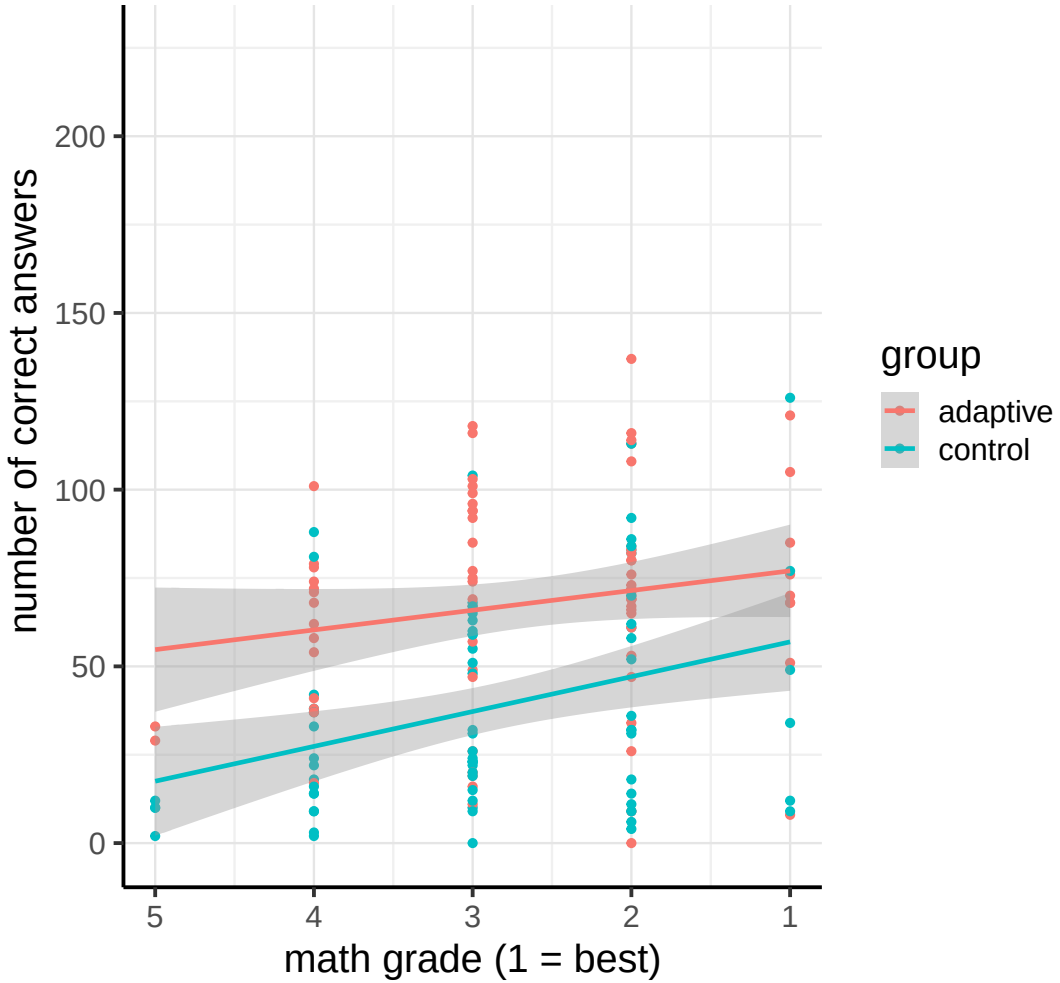
Relating school track type and learning success (# activities completed correctly)



Relating German grade and learning success (# activities completed correctly)



Relating Math grade and learning success (# activities completed correctly)



Adaptive group significantly higher on motivation and self-efficacy ratings

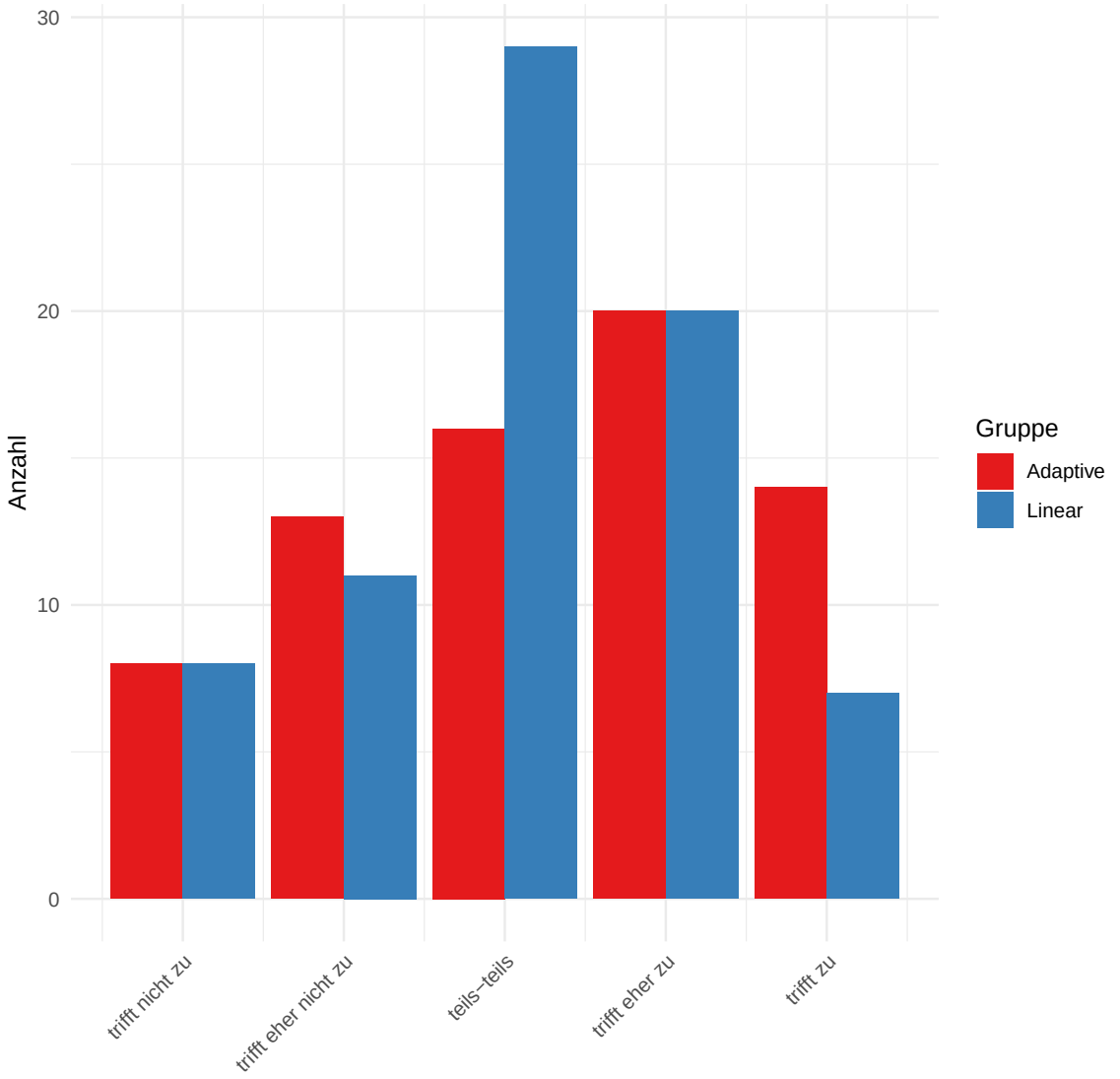
- I think I was pretty good at the tasks of the learning system. ($p = 0.01277$)
- I found it easy to use the learning system. ($p = 0.03927$)
- I am satisfied with my performance in the learning system. ($p = 0.0052$)
- When I was working on the tasks I felt tense.($p = 0.0254$)
- I had concerns about whether I would be able to manage the tasks in the learning system. ($p = 0.0459$)

I found the use of the system very interesting.

$\Rightarrow$ Adaptive learners are more extreme (more positive but also more negative)! ($p = 0.0415$)

For which subgroup is adaptive learning more interesting?

Not: number of activities completed, math or subject grades



Discussion of first ALEE study results

- Multi-dimensional parametrization of activities is feasible and supports some intended effects. The students in the study
  - learn more,
  - show increased engagement or learning success across some dimensions of heterogeneity, and
  - do not seem to mind doing more activities if those are adaptive.
- We are analyzing how the activity parameters and their interaction predict performance
  - for learners overall, and
  - for subgroups of learners

  to improve the adaptive assignment.
- Our field needs a deeper understanding of adaptivity in a multi-dimensional space.
  - Randomized controlled field studies in authentic school settings can support this endeavor with data.

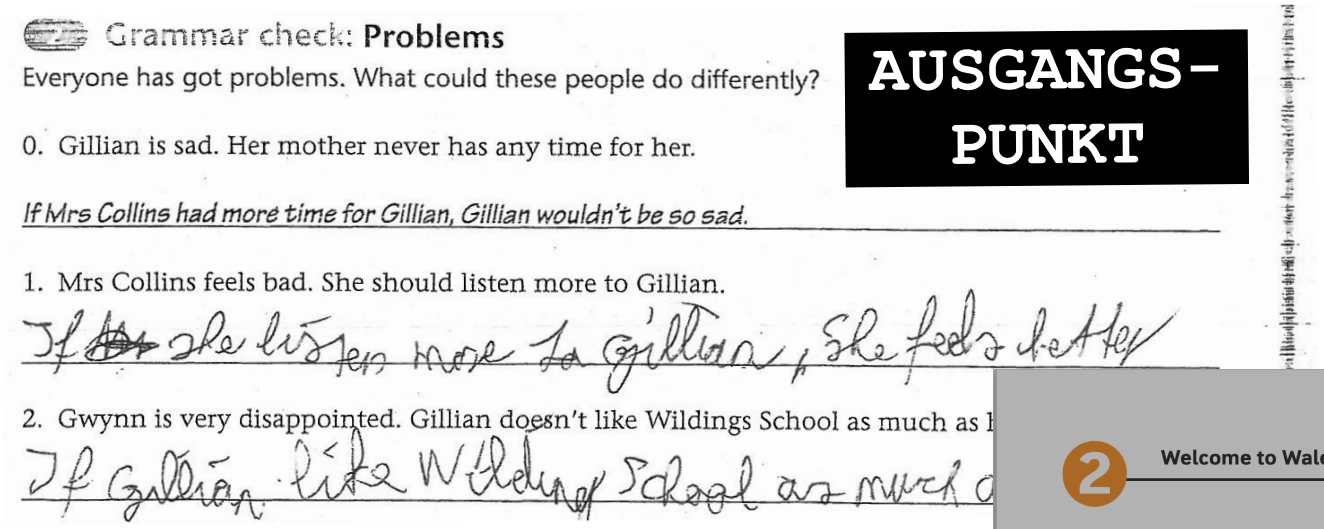
Integrating adaptive systems into real-life school

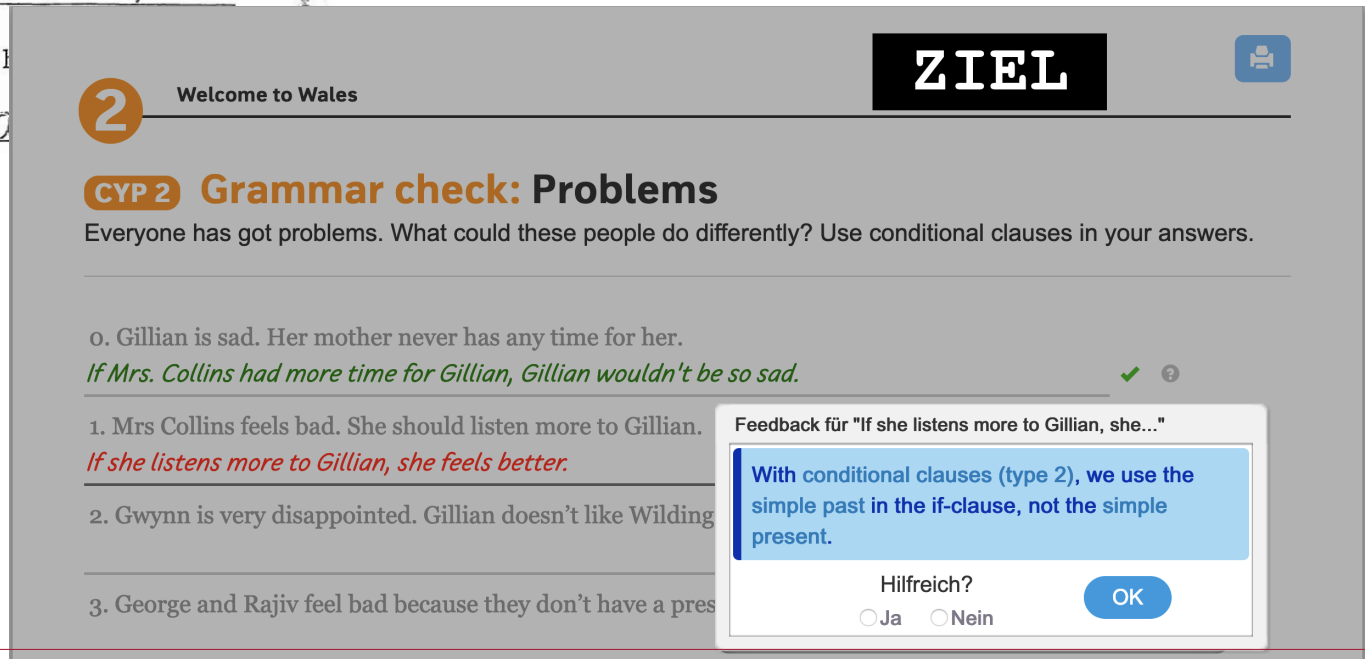
- Schools going digital is a common topic in public discourse, e.g.:

  Computers in all schools, all students at the computers – this is the program the ministers of education want to realize quickly. [Spiegel No. 47 / 18.11.1984]

- Even where individualized learning can successfully be fostered, how can we integrate it in the teacher-orchestrated setting?
- FeedBook supporting English in 7th grade in German schools:
  - (i) micro-adaptivity for English practice (SFB833/T1)
  - (ii) macro-adaptivity integrating ambulatory assessment of cognitive variables (DigBinDiff)
  - (iii) adding student dashboards for a task-based foreign language classroom (Interact4School)
  - (iv) adding teacher dashboards and teacher training (AI2Teach)

From printed workbook to AI-based FeedBook





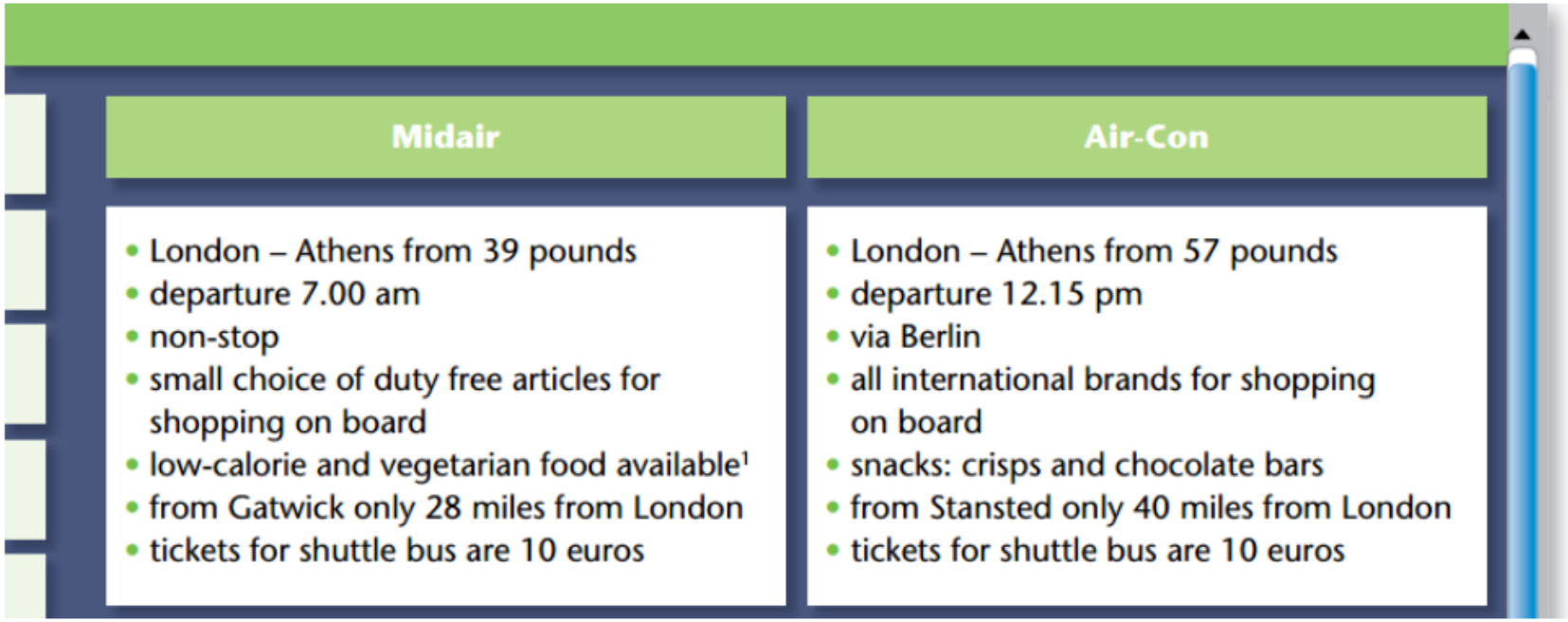
FeedBook: Feedback on language forms

B1 Off to Greece again

Mr Lambraki is checking flights to Greece. Read the information he has found on the two airlines and use the adjectives below to compare them.

- *LiF8R: Comparison of adjectives*

expensive (ticket) · early (departure) · attractive (shopping on board) · good (choice of food offered on board) · healthy (food and drinks) · suitable (airport) · cheap (tickets for shuttle bus) · friendly (service on board) · easy (online booking)



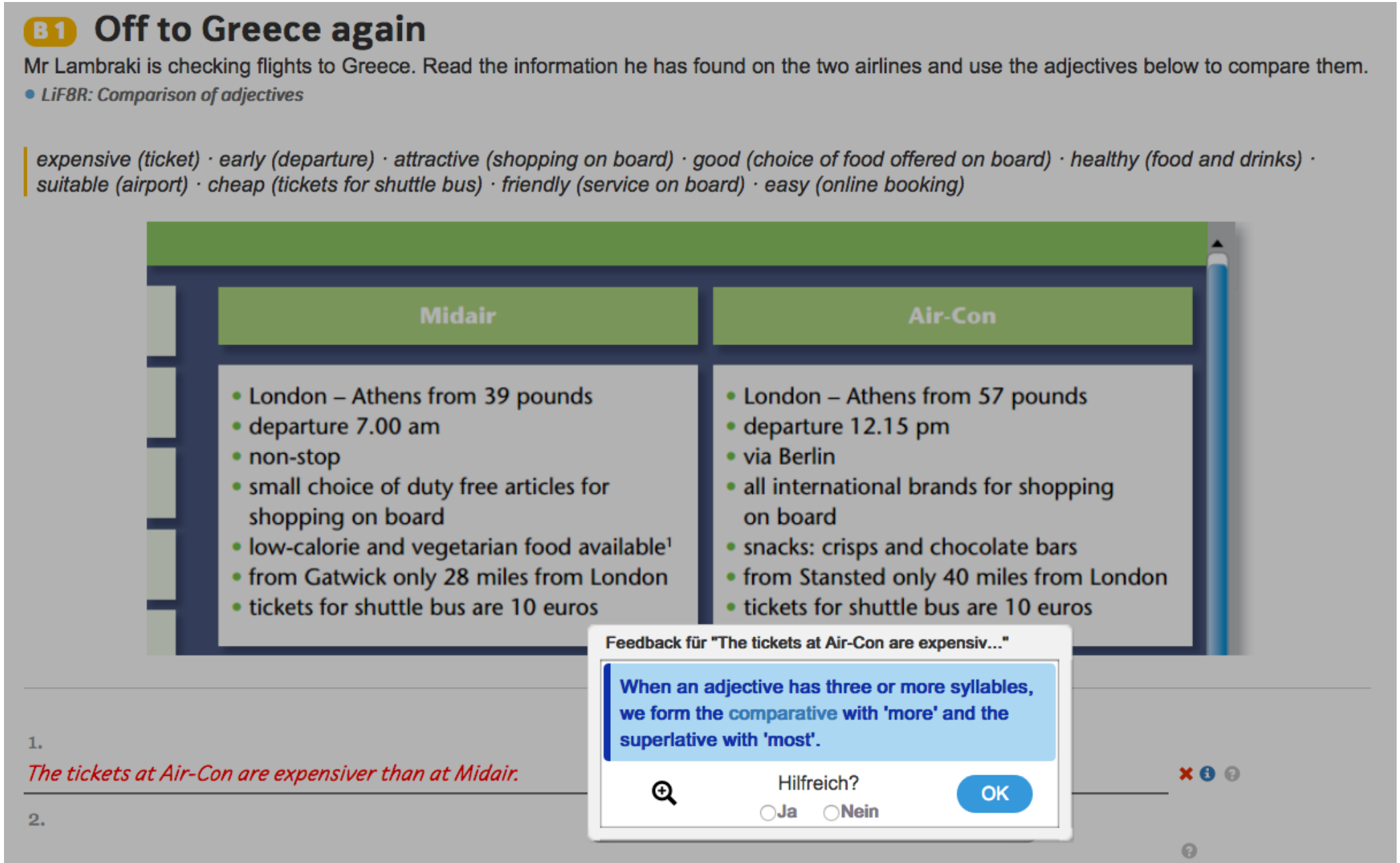
Midair	Air-Con
• London – Athens from 39 pounds • departure 7.00 am • non-stop • small choice of duty free articles for shopping on board • low-calorie and vegetarian food available[1] • from Gatwick only 28 miles from London • tickets for shuttle bus are 10 euros	• London – Athens from 57 pounds • departure 12.15 pm • via Berlin • all international brands for shopping on board • snacks: crisps and chocolate bars • from Stansted only 40 miles from London • tickets for shuttle bus are 10 euros

1.

The tickets at Air-Con are expensiver than at Midair.

2.

FeedBook: Feedback on language forms



FeedBook: Feedback on language meaning and forms

B1 Gillian's diary

Read Gillian's diary entry and complete these sentences.

Friday 23rd September

We'll drive to the north coast tomorrow and have a look at two boarding schools with Gwynn. His sister went to Wildings and he says it would be great for me. If they had a football team, it wouldn't be so bad, but it's all so girly-girly with horses and ballet dancing and everything, yuk! I don't want to go to the school in Llandysul either. It still feels like Gwynn and Mum just want me to go to boarding school because of the new baby. Miss my pals like crazy, miss London and my old school. If my friends were here, boarding school would actually be fun. Being the new girl at school without any friends will be horrible. :(And it's all Gwynn's fault! The village where we live now is the worst. 20 minutes to the nearest supermarket. If there was a shopping centre, I could at least go shopping. But there's nothing, no shops, no cinema, no nothing … only sheep!

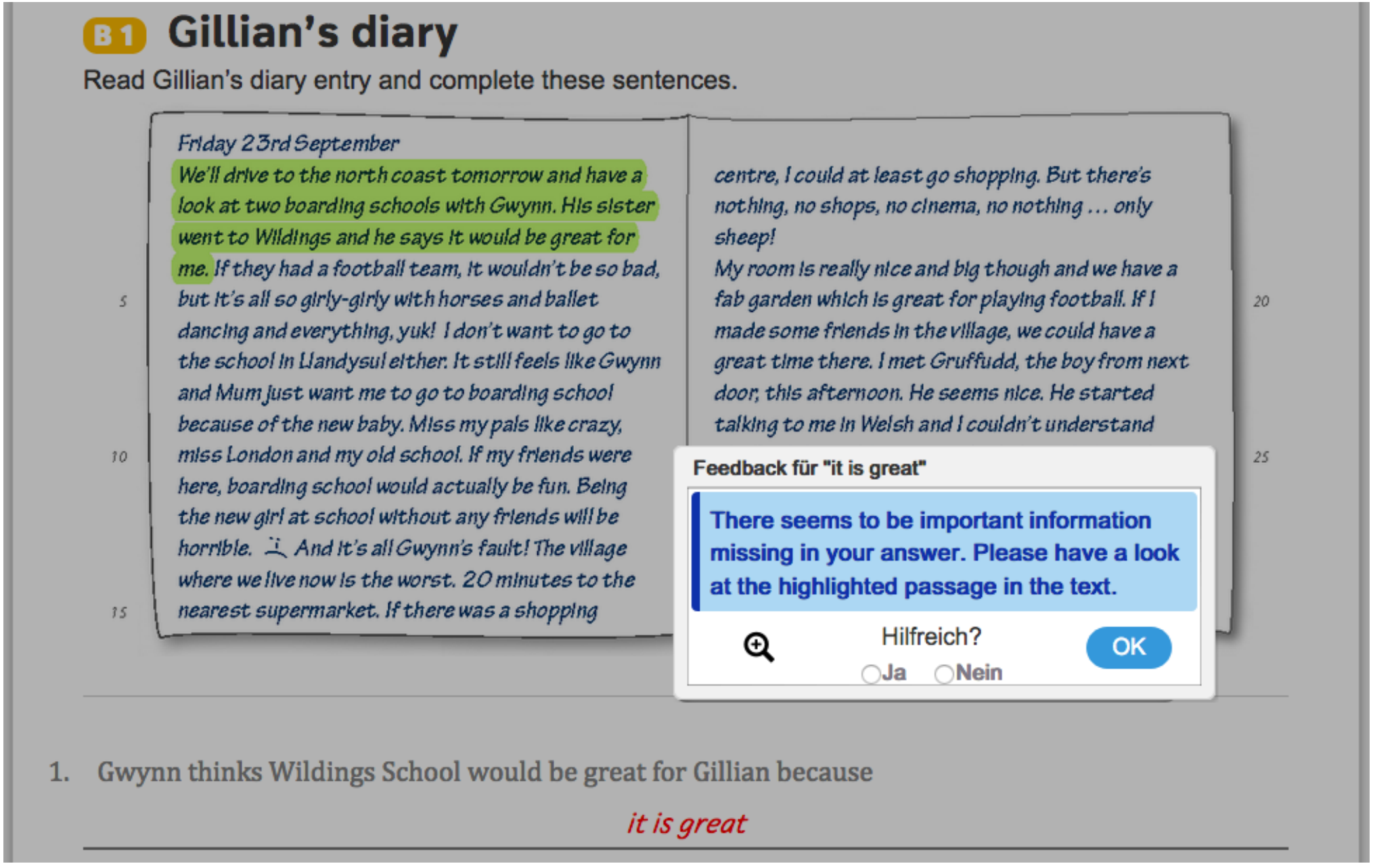
My room is really nice and big though and we have a fab garden which is great for playing football. If I made some friends in the village, we could have a great time there. I met Gruffudd, the boy from next door, this afternoon. He seems nice. He started talking to me in Welsh and I couldn't understand ANYTHING. He then spoke English and told me he plays rugby. Well, it's not football but I might have to learn to like it. They all love rugby here.

Aargh, if I had some credit on my phone, I could call Caroline. Hope I get some pocket money tomorrow.

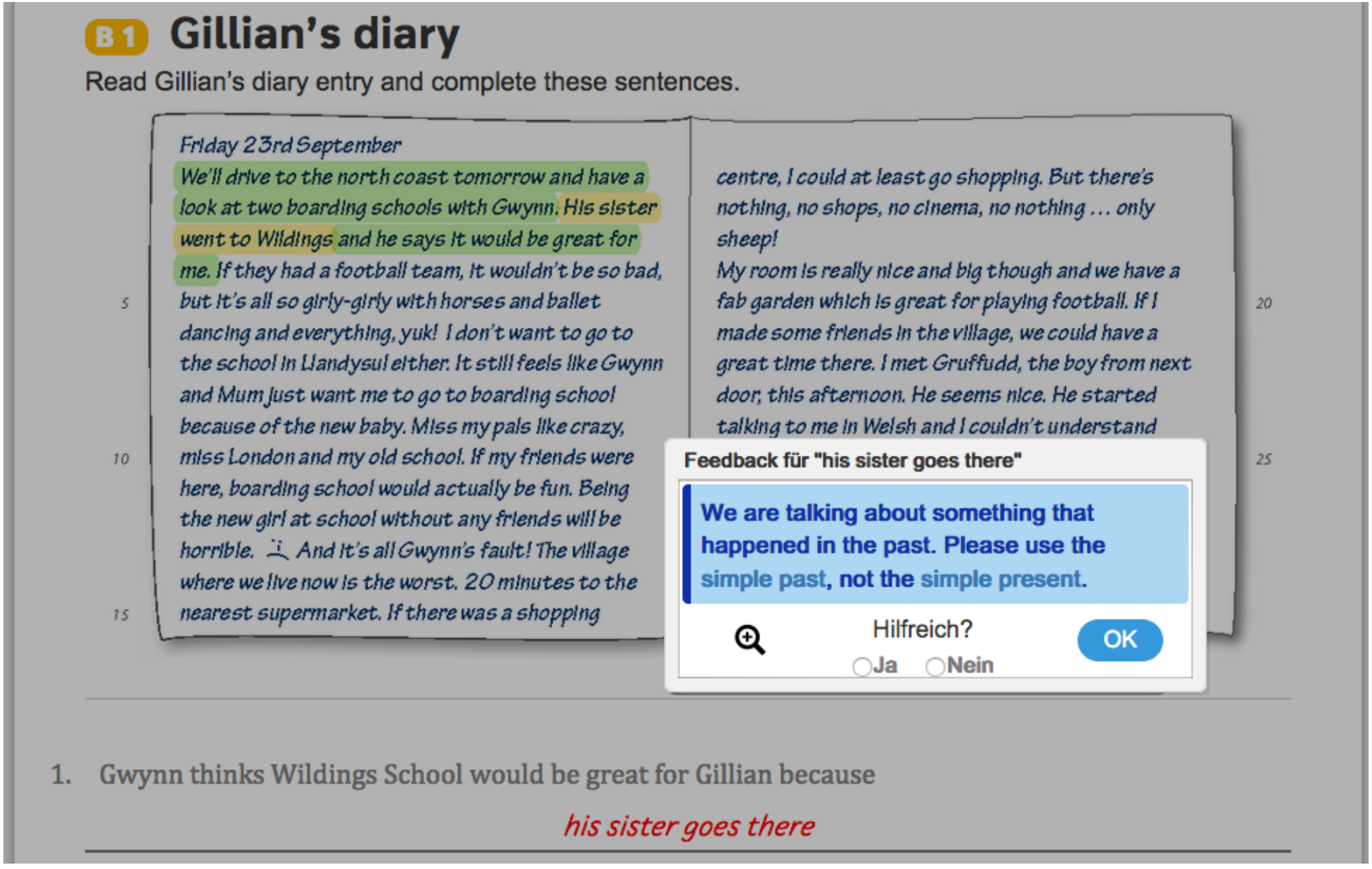
1. Gwynn thinks Wildings School would be great for Gillian because

it is great

FeedBook: Feedback on language meaning and forms



FeedBook: Feedback on language meaning and forms



FeedBook: Feedback on language meaning and forms

B1 Gillian's diary

Read Gillian's diary entry and complete these sentences.

Friday 23rd September

We'll drive to the north coast tomorrow and have a look at two boarding schools with Gwynn. His sister went to Wildings and he says it would be great for me. If they had a football team, it wouldn't be so bad, but it's all so girly-girly with horses and ballet dancing and everything, yuk! I don't want to go to the school in Llandysul either. It still feels like Gwynn and Mum just want me to go to boarding school because of the new baby. Miss my pals like crazy, miss London and my old school. If my friends were here, boarding school would actually be fun. Being the new girl at school without any friends will be horrible. ☹ And it's all Gwynn's fault! The village where we live now is the worst. 20 minutes to the nearest supermarket. If there was a shopping centre, I could at least go shopping. But there's nothing, no shops, no cinema, no nothing … only sheep!

My room is really nice and big though and we have a fab garden which is great for playing football. If I made some friends in the village, we could have a great time there. I met Gruffudd, the boy from next door, this afternoon. He seems nice. He started talking to me in Welsh and I couldn't understand ANYTHING. He then spoke English and told me he plays rugby. Well, it's not football but I might have to learn to like it. They all love rugby here.

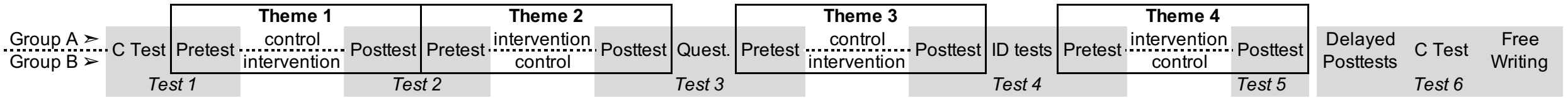
Aargh, if I had some credit on my phone, I could call Caroline. Hope I get some pocket money tomorrow.

1. Gwynn thinks Wildings School would be great for Gillian because

his sister went there

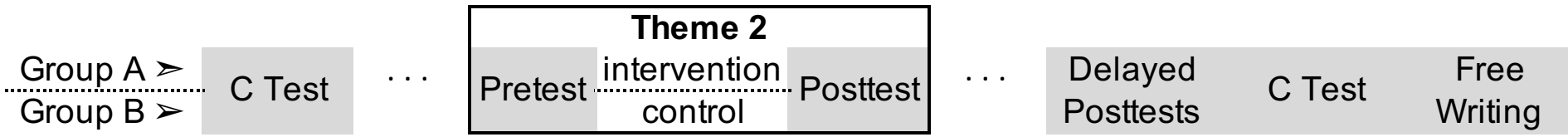
Conducting research integrated in real-life school teaching and learning

- How can a system be fully embedded so that it supports regular teaching and a randomized controlled field study?
- Cover all curricular language targets for the full year.
- Formulate research questions also addressing the needs of education stakeholders:
  - Does immediate scaffolded feedback on form during homework improve learning of the targeted language means? (Meurers et al. 2019)
- Everyone uses the system (avoids temporary multimedia effect) and benefits.
  - everyone receives feedback on meaning, orthography, and default feedback
  - within-class randomization on who sees specific feedback on which grammar topics, alternating groups



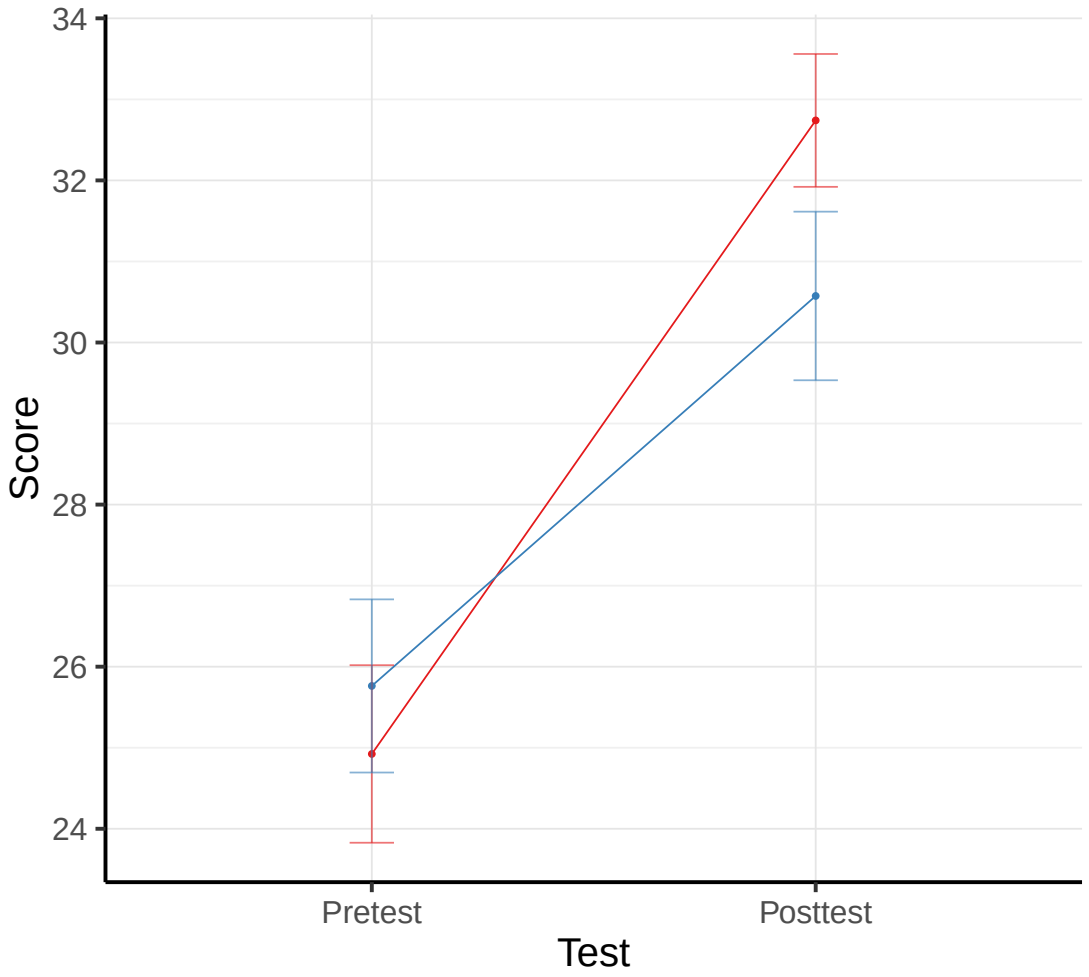
Zooming in on specific results

- Analysis of Theme 2 data (Meurers et al. 2019):



- Phenomena covered in Theme 2: comparatives, conditional clauses, relative clauses
- 205 students completed pre- and posttest for Theme 2.

Results on effectiveness of scaffolded feedback: pre-/posttest scores

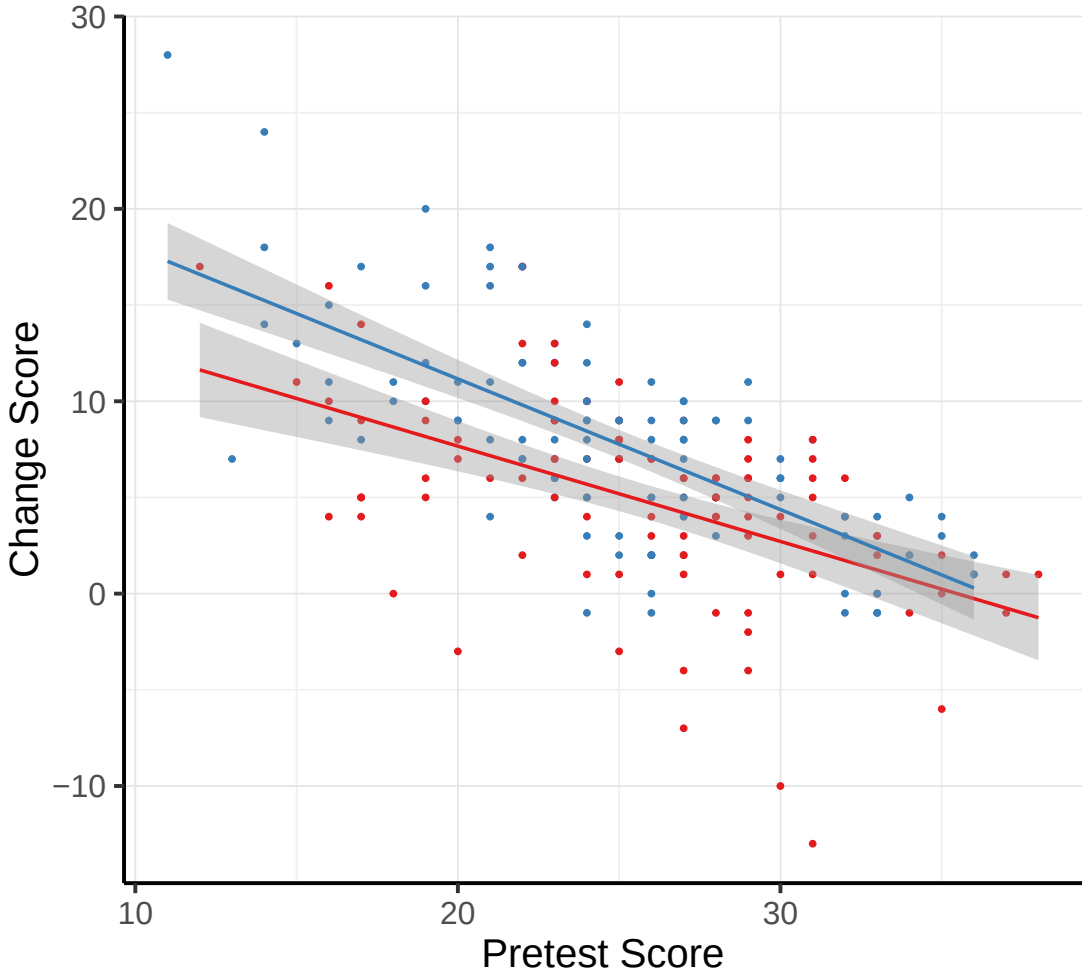


- Intervention group improved significantly more than control ($p < 0.0001$)
  - 7.82 for intervention group
  - 4.81 for control group

→ intervention group learned 62% more than control group

→ Effect size: Cohen's d = 0.56

Results: pretest & group predicting change score



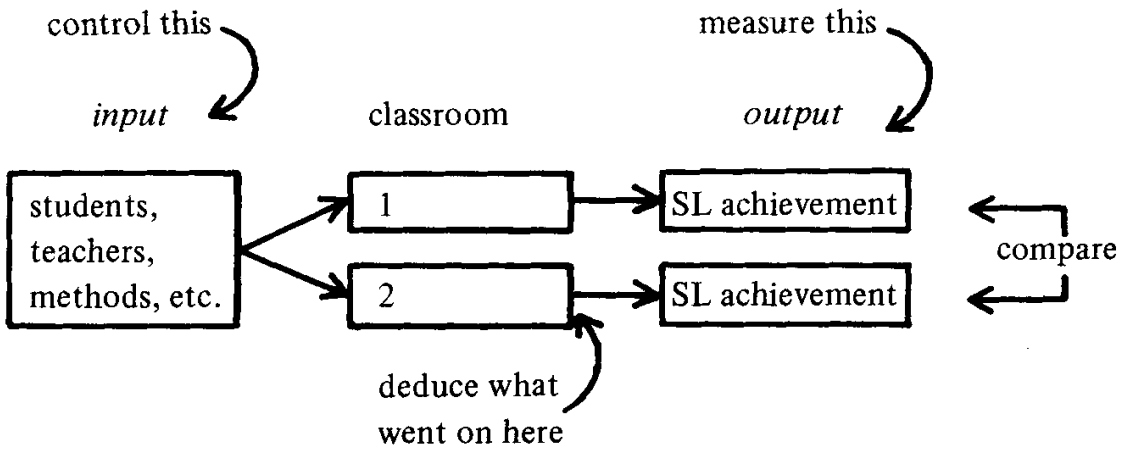
- "Worse" students improved more, but interaction possibly a ceiling effect of measuring instrument (max. score: 40)

Results relevant to stakeholders in practice and research

- Students
  - + receive individualized support while practicing → 63% learning improvement
  - + at their level → automatic internal differentiation
  - + regardless of family characteristics → educational equity
- Teachers
  - + have a reduced burden to provide written feedback and internally differentiate in class
  - + can work in class with better-prepared students
  - + are better informed about abilities of individuals and class
- Textbook authors and curriculum designers
  - + learner & activity analytics support systematic improvement of materials & curriculum
- Researchers in authentic settings can investigate the effects of
  - + different types of feedback, activities, learning targets, . . . given individual learner differences

FeedBook as a platform for randomized field trials II

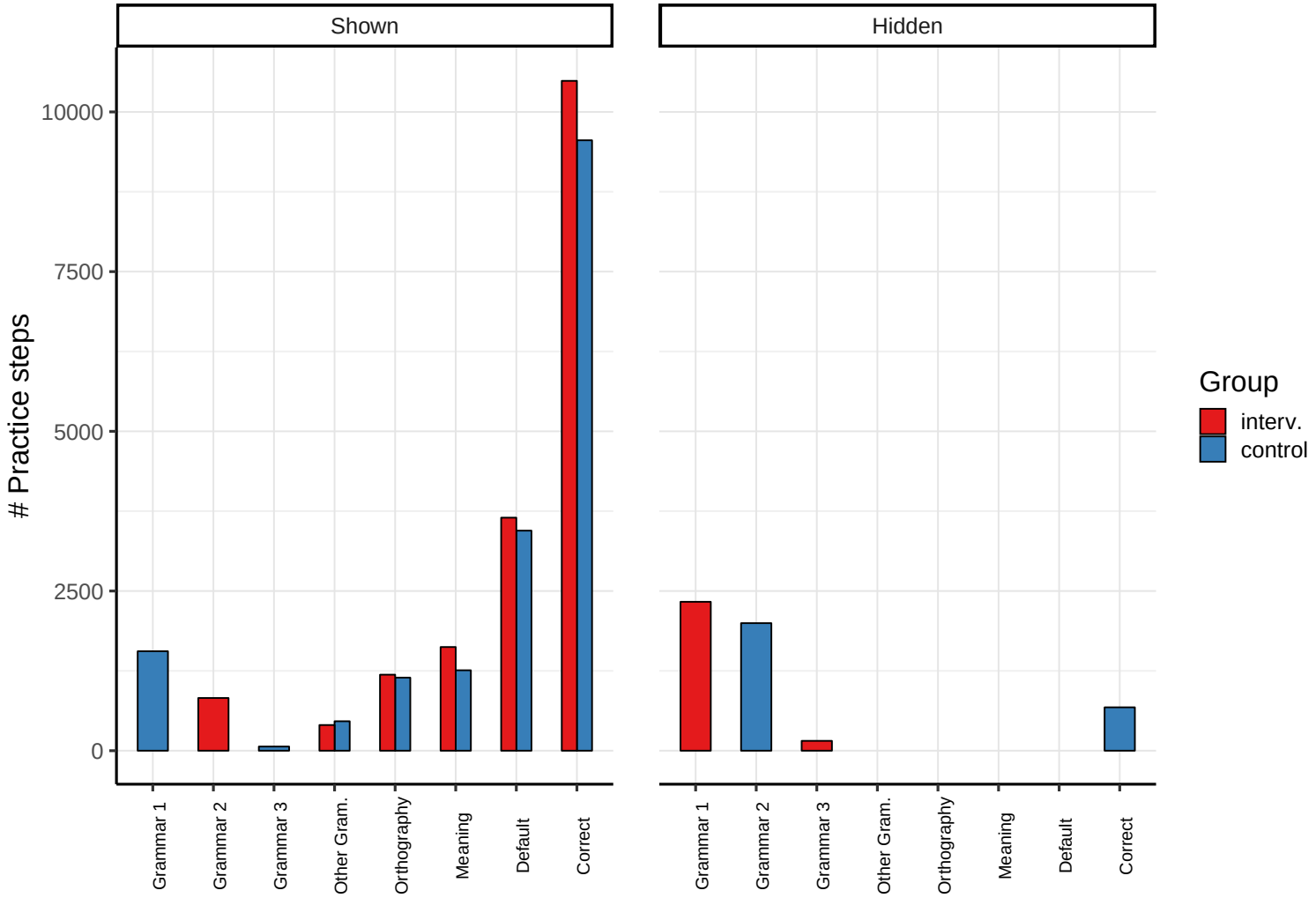
- Black box problem pointed out already by Long (1980):



Problem: Loss of control when scaling research up to authentic school context.

⇒ But learning analytics using system log data shows us what the students actually did!

Who saw which type of feedback?



Learning analytics: For whom is micro-adaptivity effective?

- How are learning gains and interaction patterns linked? (Hui, Rudzewitz & Meurers 2023)

⇒ Clustering identifies four groups of students, based on:

+/- accuracy focus of student

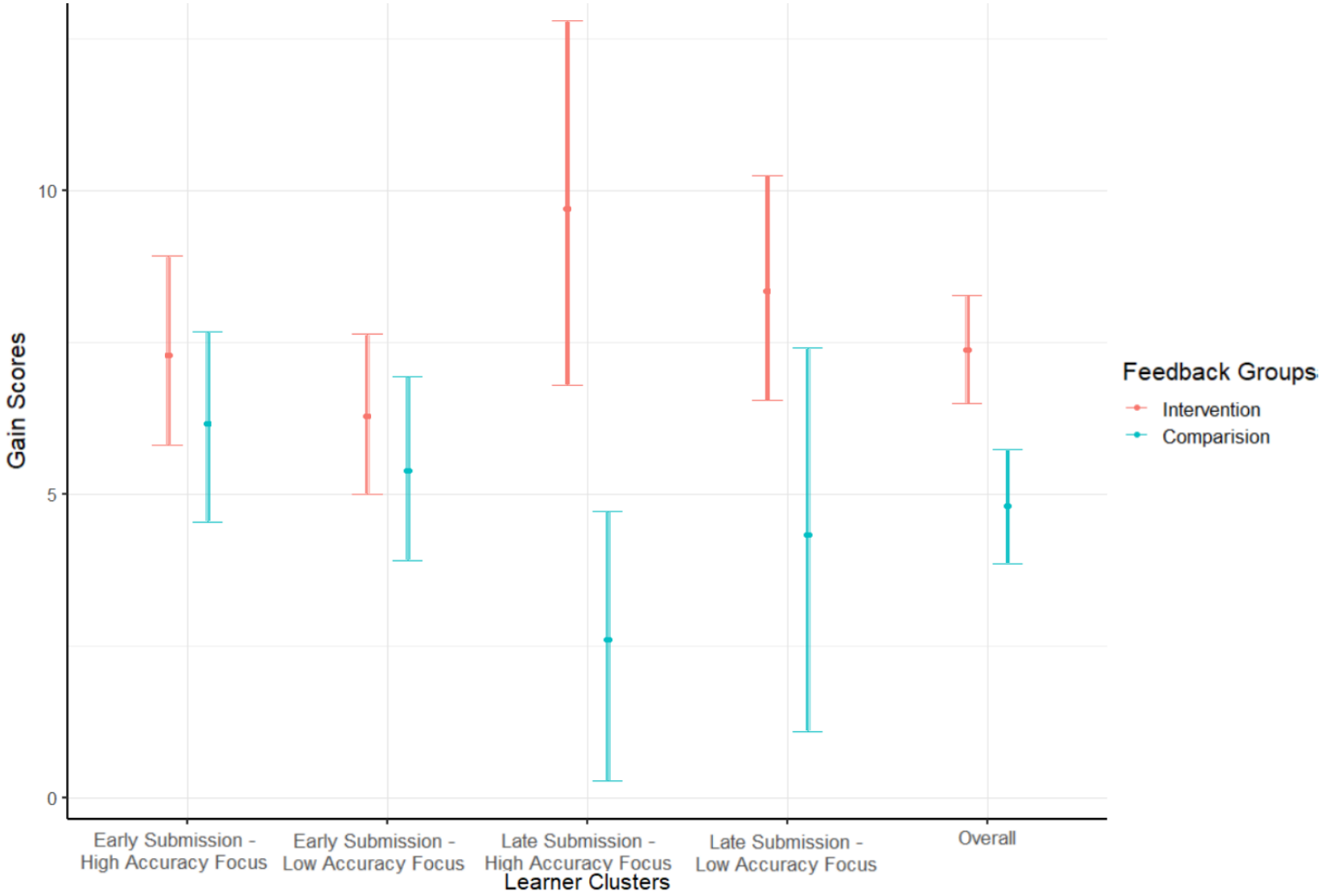
How many of the items were filled out correctly, not left empty, answered correctly at first try?

+/- submission time of student relative to peers

⇒ Scaffolded feedback has the biggest effect for students who

- systematically attempt to solve the exercises correctly
- submit later than their peers

Who benefits most from the specific feedback?



Learning analytics complements but does not replace learning product data

- Does the effect of practice of specific language forms transfer to open, functional language production? (De Kuthy & Meurers 2022)
- Writing topic chosen so that practiced language means are relevant

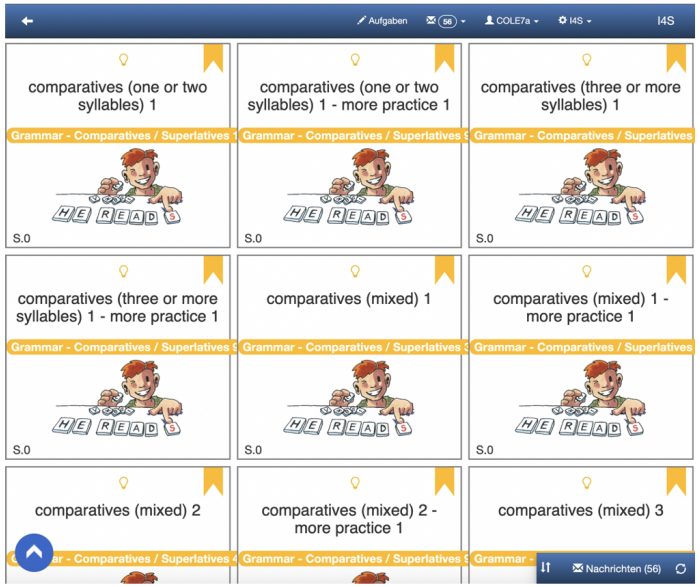
  Write a text about your holidays. Please include the following aspects:
  - *Compare two of your holiday trips (weather, duration, ...),*
  - *describe your next holiday trip, and*
  - *outline what you would do if you could spend 1000€ during your next holiday.*

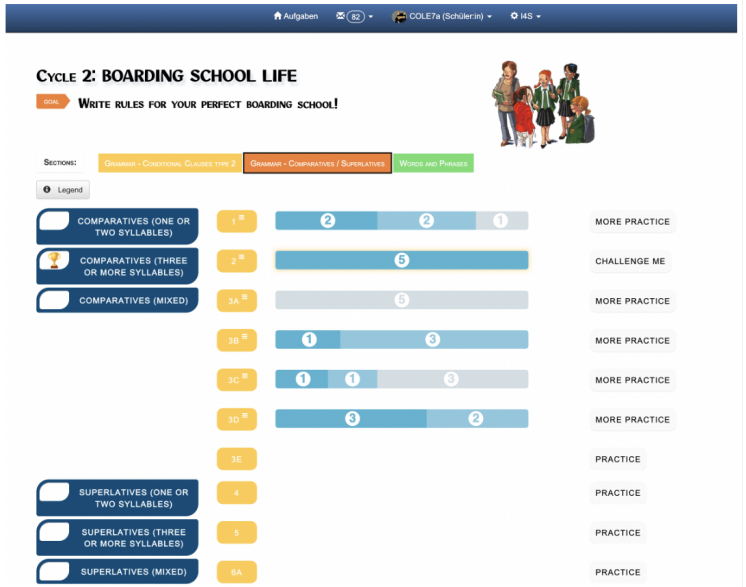
- Automatic identification of practiced language forms in the learner writing

$\Rightarrow$ The score on the form-specific delayed posttest is a significant predictor of the number of uses of those forms in the free writing!

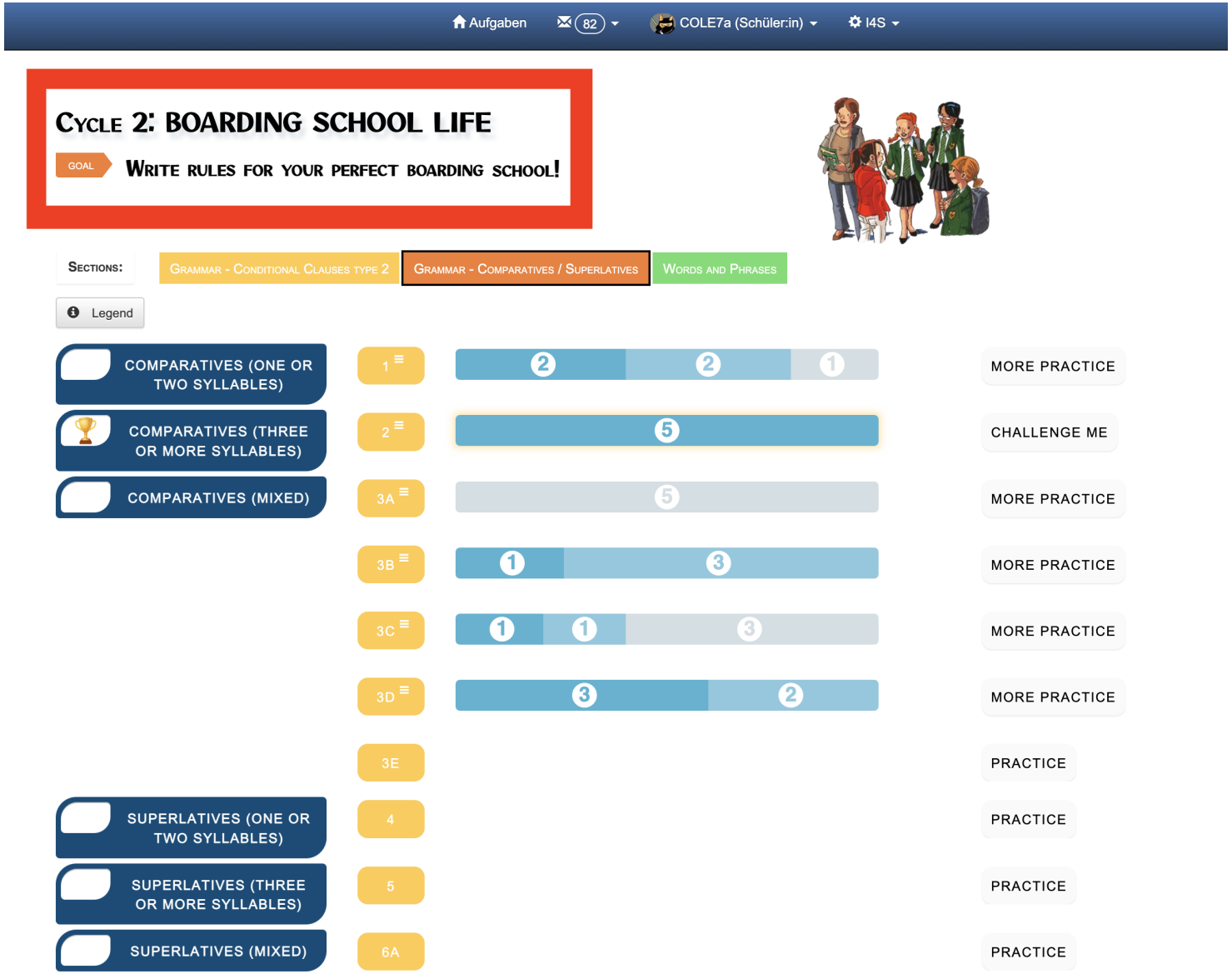
Integration of individual practice & teacher-orchestrated task-based class?

- Organize exercises so they empower students to successfully participate in functional tasks.
  - improves acceptance of practice as pre-task activities in a task-based curriculum
  - fosters intrinsic motivation and self-awareness of ability for students

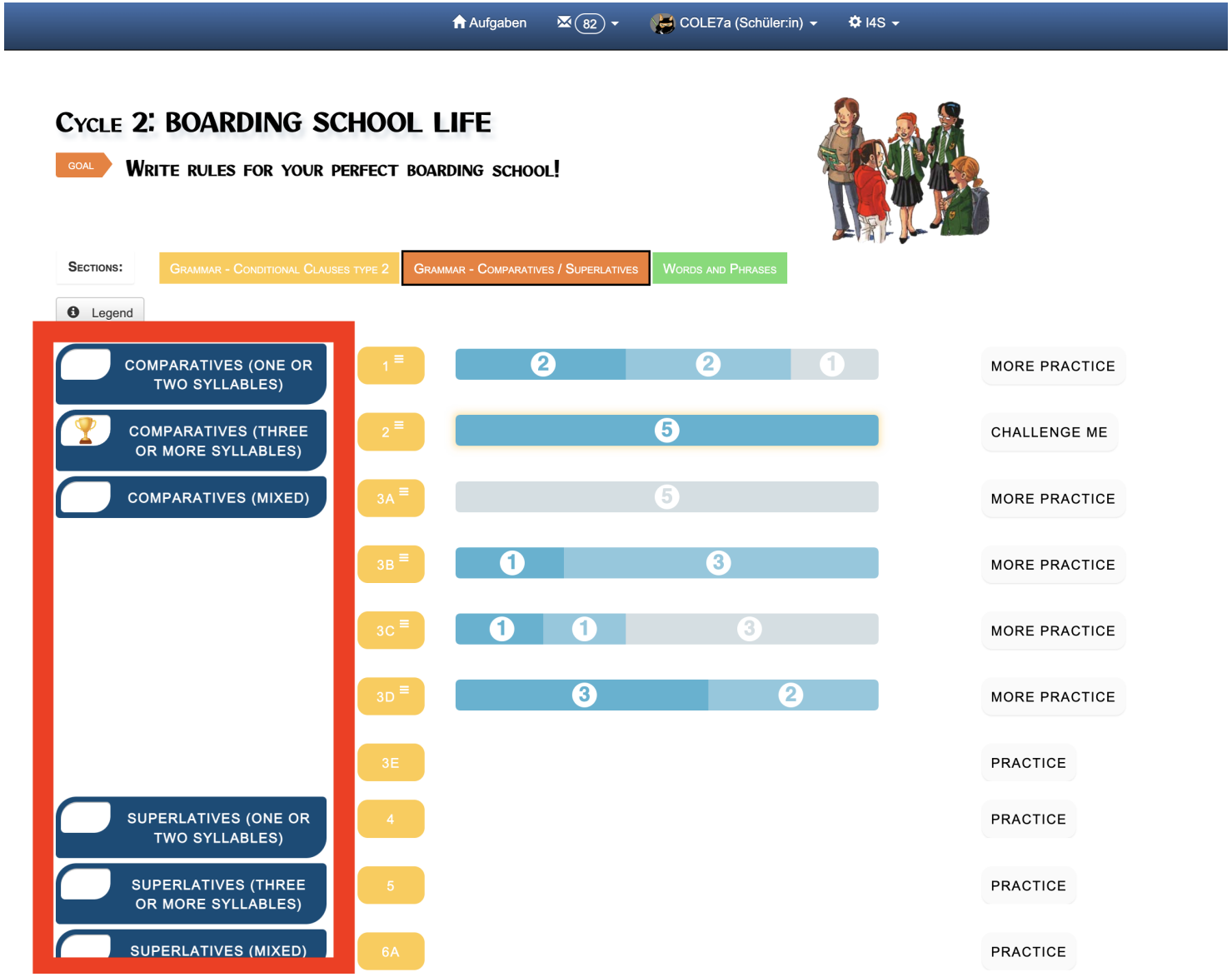




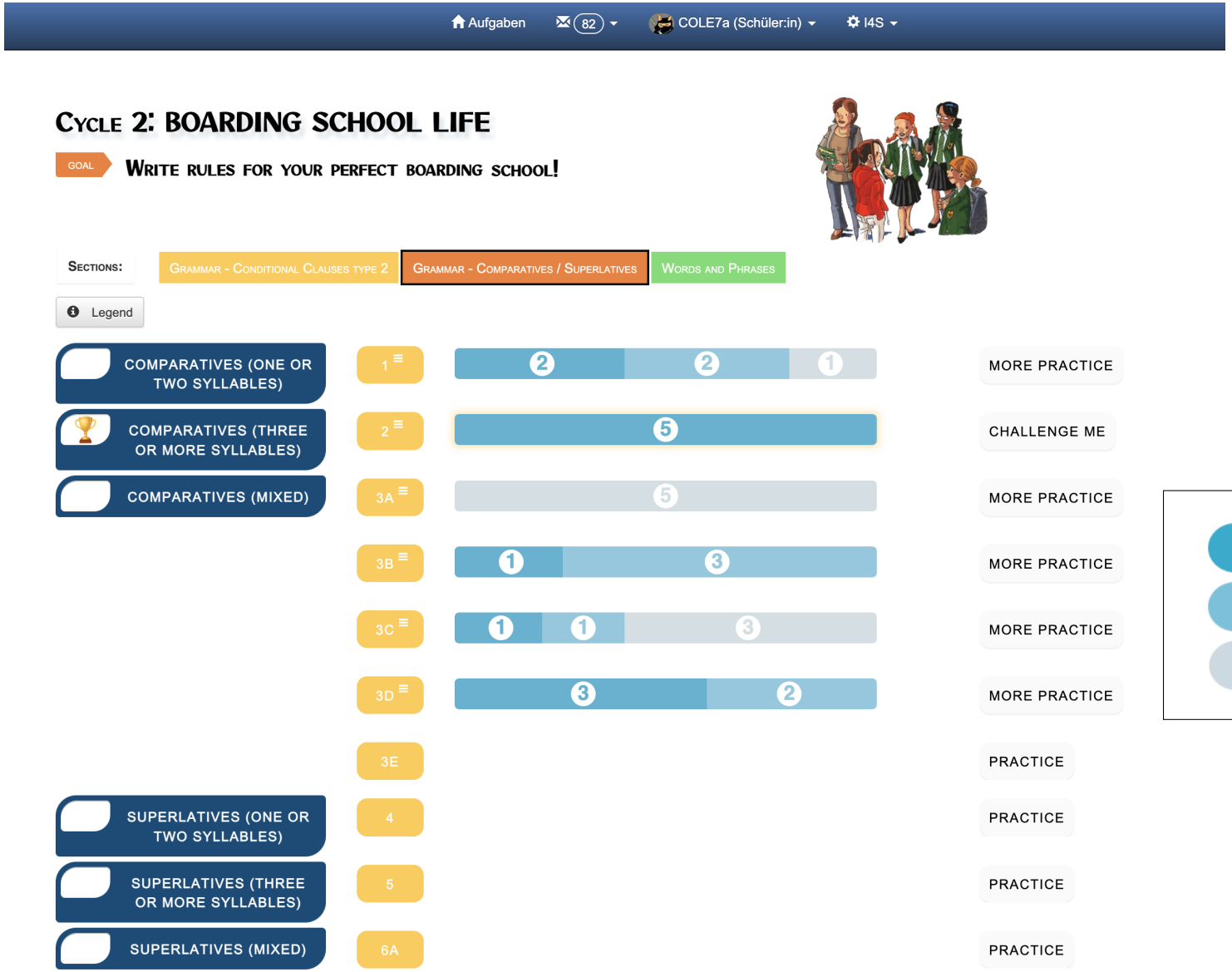
1. Make final task explicit



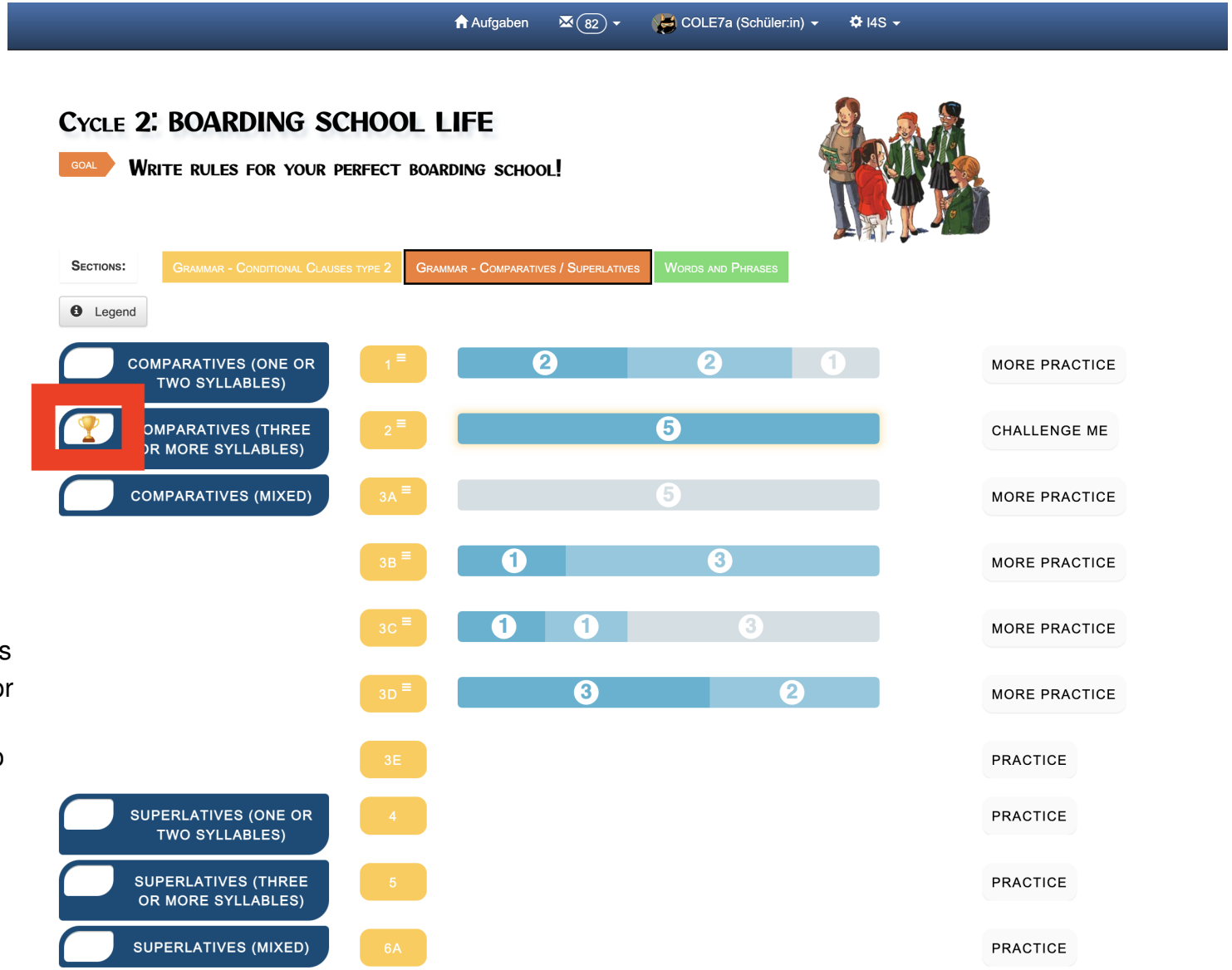
2. Organize practice by task-essential language means



3. Feedback on success relative to support received



4. Feedback on acquisition criterion



"Ready-to-Go-ness":

- tackled all difficulty levels
- 60% correct at first try for most complex
- Trophies "collect dust" to model forgetting and require revision or use

First results of Interact4School RCT using task-oriented FeedBook

- 847 students from 36 classes (English 7th grade high school)
- broad range of analyses in preparation, incl. classroom videography by partners at Univ. of Lüneburg: Torben Schmidt, Diana Pili-Moss, Carolyn Blume
- Focus here: 618 students in FeedBook condition (7 schools, 24 classes)

$\Rightarrow$ Analysis of learning success (pretest/posttest) shows significantly higher learning gains, when comparing new dashboard version to original FeedBook (Parrisius et al. submitted)

$\Rightarrow$ Interestingly, motivational feedback through avatar was detrimental after initial phase.

Summary

- Adaptivity is essential to address the **multi-dimensional student heterogeneity**.
- To digitally support individual learning, AI methods allow us to
  - **macro-adaptively** select developmentally proximal learning activities from a **rich space of activities**
  - **micro-adaptively** support students interactively.
- We presented randomized field studies using two ITS we developed for school practice:
  - ALEE: adaptive economics education offering exercises parameterized in cognitive, linguistic and domain-specific complexity
  - FeedBook: individualized English practice with feedback complementing teacher-led task-oriented class
- Adaptive learning can successfully be integrated in authentic school contexts.
  - adaptive support particularly helpful for less intrinsically motivated students, and students evaluate learning experience and self-efficacy higher
  - learning outcomes are better for learners practicing with feedback and when functionally embedded
- Let's conduct more basic research in real-life formal education contexts!

References I

Baumert, J., R. Lehmann, M. Lehrke, B. Schmitz, M. Clausen, I. Hosenfeld, O. Köller & J. Neubrand (1997). *TIMSS—Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: deskriptive Befunde*. Springer-Verlag.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety*. San Francisco: Jossey-Bass. Originally published in 1975.

Daróczy, G. (2021). Linguistic and Arithmetic Factors in Word Problems. Ph.D. thesis, Universität Tübingen. URL http://dx.doi.org/10.15496/publikation-55310.

Daroczy, G., C. Artemenko, D. Meurers, M. Wolska & H.-C. Nuerk (2020a). Influence of Task Characteristics on Eye-Movement Patterns Related to Numerical and Textual Information in Arithmetic Word Problems. *PsyArXiv* URL https://doi.org/10.31234/osf.io/6mpcx.

Daroczy, G., D. Meurers, J. Heller, M. Wolska & H.-C. Nürk (2020b). The interaction of linguistic and arithmetic factors affects adult performance on arithmetic word problems. *Cognitive processing* 21(1), 105–125.

Daróczy, G., M. Wolska, W. D. Meurers & H.-C. Nuerk (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology* 6(348). URL http://www.frontiersin.org/developmental_psychology/10.3389/fpsyg.2015.00348/abstract.

Daroczy, G., M. Wolska, H.-C. Nuerk & D. Meurers (2014). On the influence of linguistic and numerical complexity in word problems. *Cognitive Processing* 15(1), S21–S21.

De Kuthy, K. & D. Meurers (2022). Extending experimental research on the effectiveness of an Intelligent Tutoring System. A corpus study of authentic ESL student essays. In *Learner Corpus Research Conference. Book of Abstracts*. Padua, Italy.

References II

Hui, B., B. Rudzewitz & D. Meurers (2023). Learning Processes in Interactive CALL systems: Linking Automatic Feedback, System Logs, and Learning Outcomes. *Language Learning and Technology* 27(1), 1–23. URL https://hdl.handle.net/10125/73527.

Höhmann, K. (2009). Heterogenität: Eine begriffliche Klärung. In K. Höhmann, R. Kopp, H. Schäfers & M. Demmer (eds.), *Lernen über Grenzen: Auf dem Weg zu einer Lernkultur, die vom Individuum ausgeht*, Opland & Farmington Hills: Verlag Barbara Budrich, pp. 27–35.

Kunze, I. & C. Solzbacher (eds.) (2008). *Individuelle Förderung in Sekundarstufe I und II*. Schneider Verlag Hohengehren.

Letzel, V. (2021). Binnendifferenzierung in der Schulpraxis -– Eine quantitative Studie zur Einsatzhäufigkeit und zu Kontextfaktoren der Binnendifferenzierung an Sekundarschulen. Doktorarbeit, Universität Trier. URL https://doi.org/10.25353/ubtr-xxxx-b711-2fc7.

Long, M. H. (1980). Inside the “black box”: Methodological issues in classroom research on language learning. *Language Learning* 30(1), 1–42.

Meurers, D., K. De Kuthy, V. Möller, F. Nuxoll, B. Rudzewitz & R. Ziai (2018). Digitale Differenzierung benötigt Informationen zu Sprache, Aufgabe und Lerner. Zur Generierung von individuellem Feedback in einem interaktiven Arbeitsheft. *FLuL – Fremdsprachen Lehren und Lernen* 47(2), 64–82. URL http://purl.org/dm/papers/Meurers.DeKuthy.ea-18.pdf.

Meurers, D., K. De Kuthy, F. Nuxoll, B. Rudzewitz & R. Ziai (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39, 161–188.

Parrisius, C., K. Wendebourg et al. (eingereicht). Effective features of feedback in an intelligent tutoring system. *Language Learning and Technology* .

References III

Trautmann, M. & B. Wischer (2011). *Heterogenität in der Schule: eine kritische Einführung*. VS Verlag für Sozialwissenschaften & Springer.

Vygotsky, L. S. (1986). *Thought and Language*. Cambridge, MA: MIT Press.