Electronic Supplementary Material 1

Facilitating Justification, Disconfirmation, and Transparency in Diagnostic Argumentation: Effects of Automatic Adaptive Feedback in Teacher Education

Elisabeth Bauer[1*], Michael Sailer[1], Jan Kiesewetter[2], Martin R. Fischer[2], Iryna Gurevych[3], Frank Fischer[1]

[1]LMU Munich, Munich, Germany
[2]University Hospital, LMU Munich, Munich, Germany
[3]Technical University Darmstadt, Darmstadt, Germany

***Correspondence:** Elisabeth Bauer, Education and Educational Psychology, LMU Munich, Leopoldstr. 13, 80802 Munich

Zeitschrift für Pädagogische Psychologie

Overview

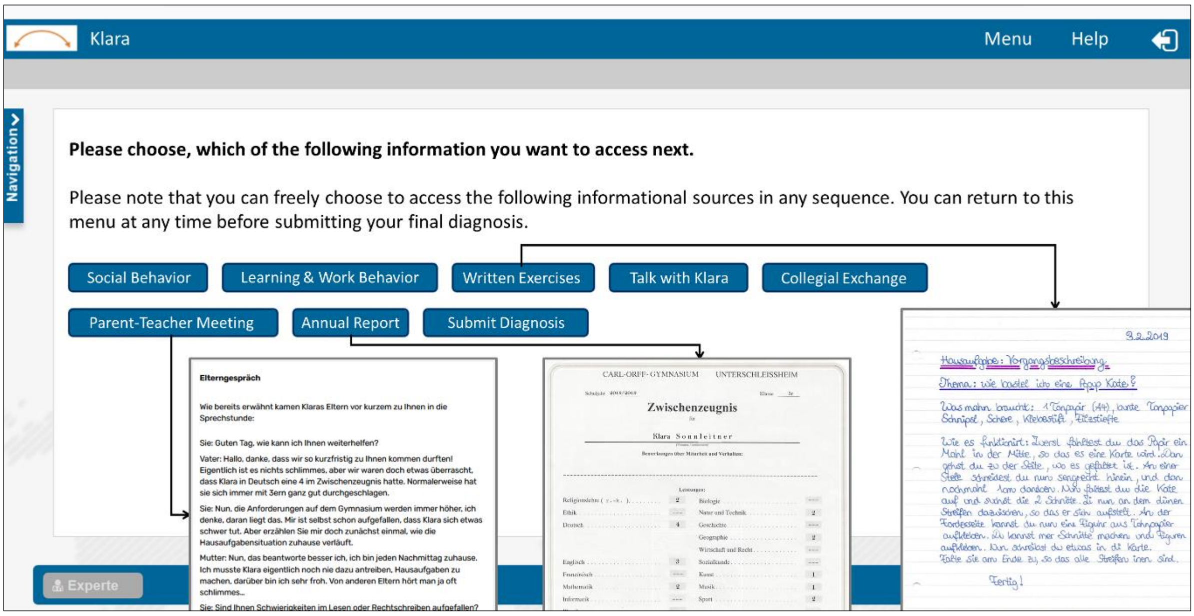
https://doi.org/10.1024/1010-0652/a000363

Supplement A. Simulated Post-test Cases

The six learning cases and the two post-test cases started with a brief description, in which a pupil was introduced as having some particular learning difficulties or behavioral problems. The participating pre-service teachers were asked to take on the role of the pupil’s teacher and further proceed with the case. On the next page, the participants saw a menu, which provided access to different sources of evidence (see Supplementary Figure 1 and Supplementary Figure 2). All of the simulated cases in the learning phase and in the post-test contained the same sources of evidence: An observational report of the pupil’s behavior toward their peers (e.g., during recess); an observational report of the pupil’s behavior during learning activities in the classroom; samples of the pupil’s written assignments (e.g., exercises and tests); the latest school certificate; a transcript of a conversation with colleagues that teach the pupil in other subjects; a transcript of a conversation with the pupil; and a transcript of a conversation with the pupil’s parents during a parent-teacher meeting (see Supplementary Figure 1 and Supplementary Figure 2). Learners were free to choose, which sources of evidence they would like to examine and in which order they would like to do so. To complete a case, learners could click on “Submit Diagnosis” and write an explanation concerning their diagnostic reasoning.

Supplementary Figure 1

Examples of Materials from the First Post-test Case



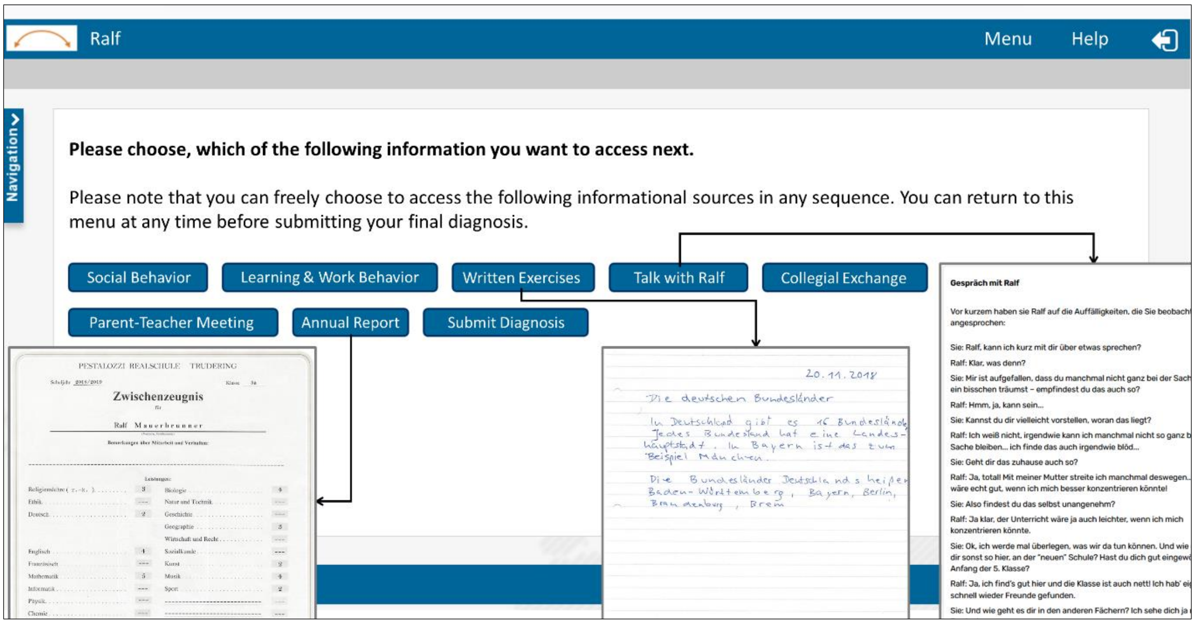
The first post-test case (see Supplementary Figure 1) was concerned with a fifth-grader named Klara. The learners were asked to take on the role of the Klara’s teacher of German and geography. Klara is described as socially well integrated, sharing several friendships with other girls in the class. She is rather calm during the lessons, but she gives

https://doi.org/10.1024/1010-0652/a000363

good answers on question asked by the teacher. Her essays usually contain creative ideas and she seems to spend an adequate amount of effort on completing her assignments. However, her orthography skills are very poor. As also observable in the accessible samples of her written assignments (see Supplementary Figure 1), she tends to confuse orthographic rules and make basic spelling errors, such as omitting characters within words. She is also not consistently making the same mistakes but sometimes misspells the same words in different ways. By contrast, her reading speed and reading comprehension meet the average performance level and she rather seems to enjoy discussing reading assignments. Her latest school certificate as well as a conversation with Klara's teacher of science and technology indicate that Klara generally achieves average to good grades. However, particularly in subjects that require a lot of writing, her grades seem to suffer from her writing. A conversation with Klara as well as a conversation with her parents confirm that she is aware and ashamed of her writing difficulties. However, the conversation with the parents also indicates that during elementary school, Klara used to achieve average grades for her writing. Overall, the case information was designed such that it suggests an isolated spelling disorder as the most likely explanation for Klara's performance problems.

Supplementary Figure 2

Examples of Materials from the Second Post-test Case



The second post-test case (see Supplementary Figure 2) was concerned with a fifth-grader named Ralf. The learners were asked to take on the role of the German teacher, who realized that Ralf tends to be rather inattentive during class. He is known as outgoing and talkative toward his peers. During school recess, other kids with whom he jokes around usually surround him. In contrast, during class, he is usually very quiet and sometimes seems

https://doi.org/10.1024/1010-0652/a000363

to be lost in thought. In addition, he is generally very slow in completing his assignments and often does not fully finish them, which is also observable in the accessible samples of his written assignments (see Supplementary Figure 2). He is rather disorganized and tends to forget his learning materials, such as handouts or schoolbooks, at home. The latest school certificate as well as a conversation with the math teacher indicate that Ralf generally achieves average to poor grades. He has particular performance problems in math and is currently having the lowest average math grade of all pupils in his class. Ralf himself confirms that he tends to get lost in thought and explains that he has difficulties to stay concentrated while doing a task. He emphasizes that it is easier for him to concentrate in some classes, such as German class, which he enjoys more than other classes, such as math. At home, his inattentiveness sometimes causes arguments with his mother. During a parent-teacher conference, his mother explains that she repeatedly needs to remind Ralf to start doing his homework as well as finishing it. Overall, the case information was designed such that it indicates an attention-deficit disorder with a potential comorbid dyscalculia as the most likely explanation for Ralf's performance problems.

Supplement B. Static and Adaptive Feedback

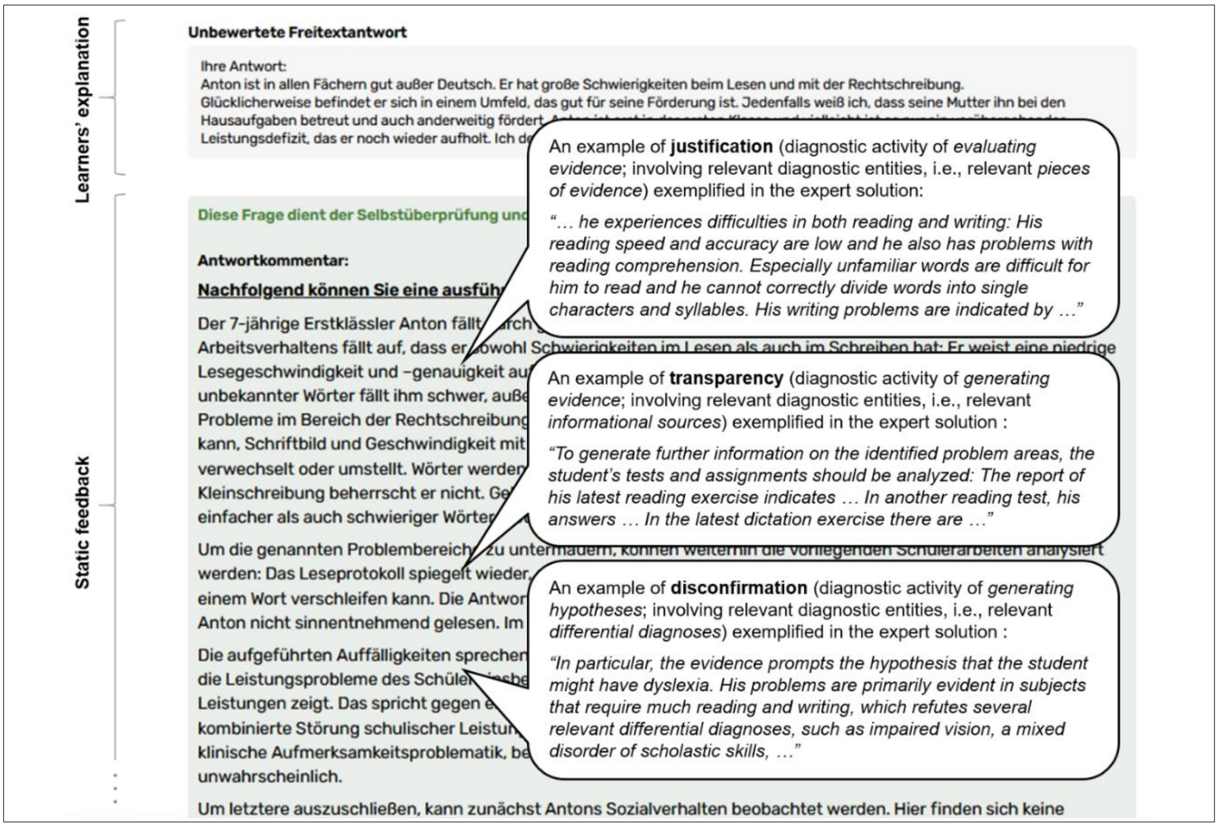
Learners in the *static feedback* condition received case-specific expert solutions, which exemplified the epistemic and the content dimension of how experts would relate the complementary information of justification, disconfirmation, and transparency in their diagnostic argumentation (see Supplementary Figure 3).

In the *adaptive feedback* condition (AFC), learners' explanations were analyzed by an NLP-algorithm, which was trained using the Python-based web service NeuralWeb. The training data (i.e., written explanations on the same simulated cases of 118 preservice teachers) was manually coded regarding diagnostic entities (i.e., content dimension; e.g., hyperactivity) and epistemic activities (i.e., epistemic dimension; e.g., evaluating evidence). Thus, the algorithm could identify diagnostic entities and epistemic activities as correct, incorrect, or missing in new explanations written by learners in the present study. Based on the automatic analysis, a suitable subset of around 40 case-specific feedback paragraphs were adaptively shown to the learner. Parts of the feedback addressed the epistemic activities and their relations (i.e., epistemic dimension) and other parts the diagnostic entities and their relations (i.e., content dimension; see Supplementary Figure 4). The adaptive feedback also offered highlighting diagnostic entities and activities found in a learner's submitted explanation.

https://doi.org/10.1024/1010-0652/a000363

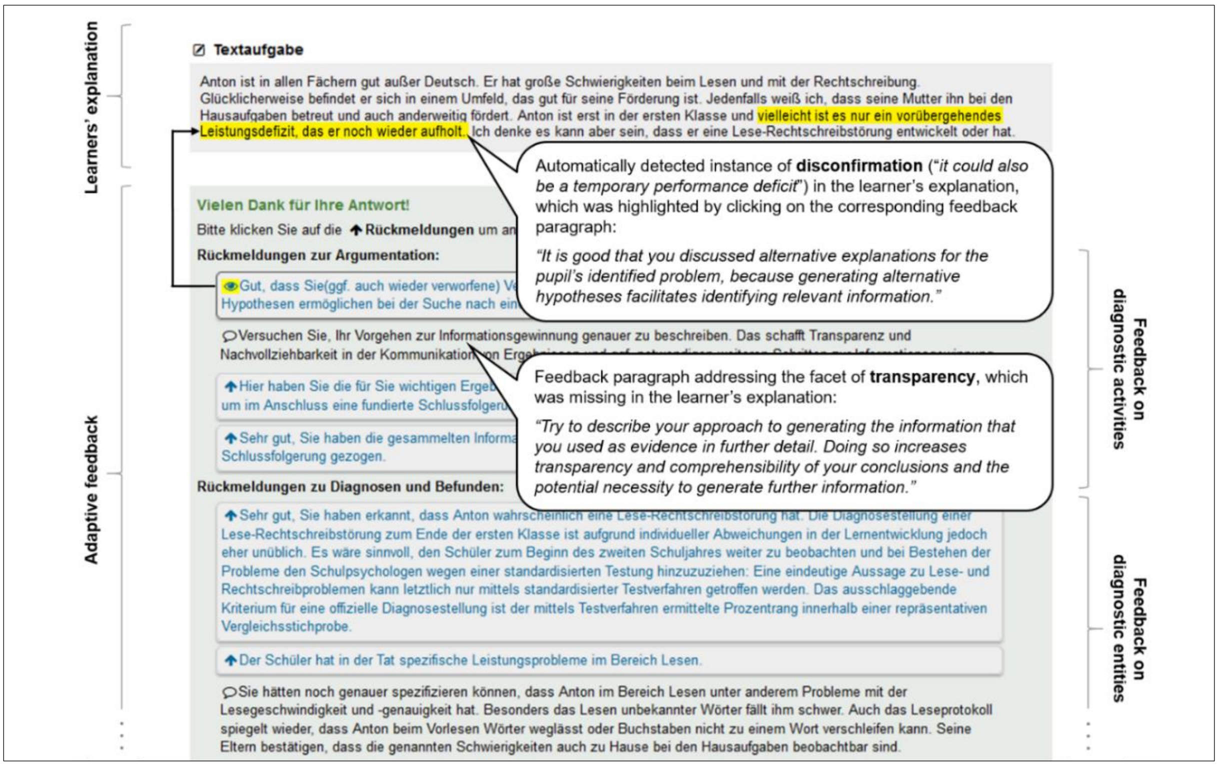
Supplementary Figure 3

Static Feedback



Supplementary Figure 4

Automatic Adaptive Feedback



https://doi.org/10.1024/1010-0652/a000363

Supplement C. Randomization Check

Individual Argumentation Facets before the Feedback Intervention

In the following, we report the descriptive statistics of the individual argumentation facets in the first learning case prior to receiving the first feedback, separated by the experimental conditions. We also calculated a multivariate ANOVA with the independent variable *feedback* and the dependent variables *justification*, *disconfirmation*, and *transparency* to report inferential statistics concerning the a priori differences of the individual argumentation facets between the static feedback condition and the adaptive feedback condition.

In terms of justification, participants in the static feedback condition included on average M = 2.63 (SD = 1.52) of the six primary supporting pieces of evidence for the correct diagnosis in their diagnostic argumentation, whereas in the adaptive feedback condition the average was M = 2.50 (SD = 1.11). The difference was not statistically significant, $F(1,58)$ = 0.15, p = .70, η_p^2 = 0.003.

In terms of disconfirmation, participants in the static feedback condition included on average M = 1.27 (SD = 0.89) of the six most relevant differential diagnoses in their diagnostic argumentation, whereas in the adaptive feedback condition the average was M = 1.20 (SD = 1.11). The difference was not statistically significant, $F(1,58)$ = 0.07, p = .79, η_p^2 = 0.001.

In terms of transparency, participants in the static feedback condition included on average M = 1.47 (SD = 1.55) of the six most relevant differential diagnoses in their diagnostic argumentation, whereas in the adaptive feedback condition the average was M = 1.23 (SD = 1.28). The difference was not statistically significant, $F(1,58)$ = 0.41, p = .53, η_p^2 = 0.007.

Overall, we found no significant a priori difference between the static feedback condition and the adaptive feedback condition concerning the individual argumentation facets, which further supports that the randomization was successful.

Relations between the Argumentation Facets before the Feedback Intervention

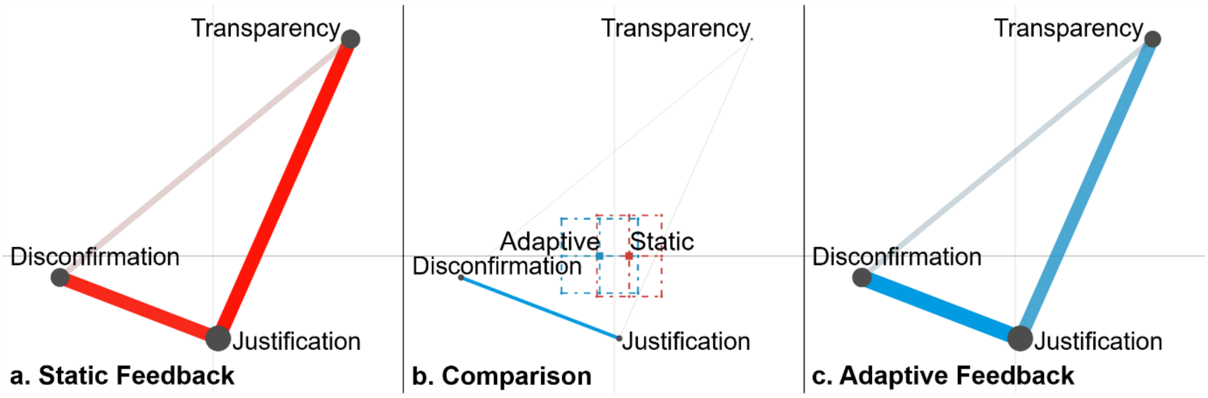
We considered learners' performance in the first learning case as pretest, because learners received the first feedback only after completing the first learning case. Using ENA, we compared diagnostic argumentation networks of the static feedback condition (Supplementary Figure 5a) and of the adaptive feedback condition (Supplementary Figure 5c) in the first learning case (interpretation of the ENA networks is explained in the Results

https://doi.org/10.1024/1010-0652/a000363

section of the manuscript). Using a *t*-test, we found no significant a priori difference between the group mean of the static feedback condition (M = .12, SD = .68) and the group mean of the adaptive feedback condition (M = -.12, SD = .81), $t(56.42) = -1.20$, $p = .23$, Cohen's d = .31. Thus, the randomization is considered successful.

Supplementary Figure 5

Diagnostic Argumentation Networks of the Static Feedback Condition (5a) and the Adaptive Feedback Condition (5c) before the Intervention; The Comparison Plot (5b) Shows Differences Between the Two Networks, Group Means (Colored Squares), and Confidence Intervals (Dashed Boxes)



Supplement D. Time on Task

Time on Task during the Learning Phase with the Feedback Intervention

In the following, we report the descriptive statistics of time on task during the learning phase, separated by the experimental conditions. In doing so, we distinguished between time on task for examining the case materials and time on task writing a diagnostic argumentation. We also calculated a multivariate ANOVA with the independent variable *feedback* and the dependent variables *examination time* and *writing time* to report inferential statistics concerning the differences between the static feedback condition and the adaptive feedback condition.

Participants in the static feedback condition took on average M = 192.23 (SD = 72.80) seconds per learning case for examining the case materials, whereas in the adaptive feedback condition the average was M = 191.60 (SD = 76.21) seconds per learning case. The difference was not statistically significant, $F(1,58) = 0.01$, $p = .97$, $\eta_p^2 < 0.001$.

Regarding the time on task for writing a diagnostic argumentation, participants in the static feedback condition took on average M = 250.21 (SD = 151.03) seconds per learning case, whereas in the adaptive feedback condition the average was M = 275.31 (SD = 152.78)

https://doi.org/10.1024/1010-0652/a000363

seconds per learning case. The difference was not statistically significant, $F(1,58) = 0.41$, $p = .53$, $\eta_p^2 = 0.007$.

The result indicate that participants in both experimental groups spent similar efforts on processing the simulated cases and learning tasks.

Time on Task during the Post-test Phase without the Feedback Intervention

In the following, we report the descriptive statistics of time on task during the post-test phase, separated by the experimental conditions. In doing so, we distinguished between time on task for examining the case materials and time on task writing a diagnostic argumentation. We also calculated a multivariate ANOVA with the independent variable *feedback* and the dependent variables *examination time* and *writing time* to report inferential statistics concerning the differences between the static feedback condition and the adaptive feedback condition.

Participants in the static feedback condition took on average $M = 95.67$ ($SD = 61.32$) seconds per post-test case for examining the case materials, whereas in the adaptive feedback condition the average was $M = 106.43$ ($SD = 59.00$) seconds per post-test case. The difference was not statistically significant, $F(1,58) = 0.48$, $p = .49$, $\eta_p^2 = 0.008$.

Regarding the time on task for writing a diagnostic argumentation, participants in the static feedback condition took on average $M = 113.90$ ($SD = 50.87$) seconds per post-test case, whereas in the adaptive feedback condition the average was $M = 147.73$ ($SD = 67.02$) seconds per post-test case. The difference was statistically significant with a medium effect, $F(1,58) = 4.85$, $p = .03$, $\eta_p^2 = 0.077$.

The result show that participants in the adaptive feedback condition took on average more time to write a diagnostic argumentation in the post-test cases than participants in the static feedback condition.

Supplement E. Individual Argumentation Facets in the Post-test

In the following, we report the descriptive statistics of the individual argumentation facets in the two post-test cases, separated by the experimental conditions. We also calculated a multivariate ANOVA with the independent variable *feedback* and the dependent variables *justification*, *disconfirmation*, and *transparency* to report inferential statistics concerning the differences of the individual argumentation facets between the static feedback condition and the adaptive feedback condition in the post-test phase.

https://doi.org/10.1024/1010-0652/a000363

In terms of justification, participants in the static feedback condition included on average *M* = 1.93 (*SD* = 0.86) of the six primary supporting pieces of evidence for the correct diagnosis in their diagnostic argumentation in each post-test case, whereas in the adaptive feedback condition the average was *M* = 2.50 (*SD* = 0.64). The difference was statistically significant with a large effect, $F(1,58) = 8.37$, $p = .005$, $\eta_p^2 = 0.126$.

In terms of disconfirmation, participants in the static feedback condition included on average *M* = 1.20 (*SD* = 0.64) of the six most relevant differential diagnoses in their diagnostic argumentation in each post-test case, whereas in the adaptive feedback condition the average was *M* = 1.40 (*SD* = 0.82). The difference was not statistically significant, $F(1,58) = 1.11$, $p = .30$, $\eta_p^2 = 0.019$.

In terms of transparency, participants in the static feedback condition included on average *M* = 0.67 (*SD* = 0.86) of the six most relevant differential diagnoses in their diagnostic argumentation in each post-test case, whereas in the adaptive feedback condition the average was *M* = 1.03 (*SD* = 1.03). The difference was not statistically significant, $F(1,58) = 2.22$, $p = .14$, $\eta_p^2 = 0.037$.

Overall, we found that participants in the adaptive feedback condition achieved descriptively higher post-test scores for the individual argumentation facets. However, the difference was only statistically significant for justification, but not for disconfirmation or transparency.

The ENA analysis reported in the paper suggested that adaptive feedback compared to static feedback facilitated relations between justification, disconfirmation, and transparency in pre-service teachers' diagnostic argumentation (i.e., participants in the adaptive feedback condition rather related the complementary information of all three argumentation facets).

https://doi.org/10.1024/1010-0652/a000363