Automated feedback on the structure of hypothesis tests

Sietske Tacoma[1][0000-0002-9662-8489], Bastiaan Heeren[1,2], Johan Jeuring [1,2][0000-0001-5645-7681] and Paul Drijvers[1][0000-0002-2724-4967]

[1] Utrecht University, Utrecht, The Netherlands
[2] Open University of the Netherlands, Heerlen, The Netherlands
s.g.tacoma@uu.nl

Abstract. Hypothesis testing is a challenging topic for many students in introductory university statistics courses. In this paper we explore how automated feedback in an Intelligent Tutoring System can foster students' ability to carry out hypothesis tests. Students in an experimental group (*N* = 163) received elaborate feedback on the structure of the hypothesis testing procedure, while students in a control group (*N* = 151) only received verification feedback. Immediate feedback effects were measured by comparing numbers of attempted tasks, complete solutions, and errors between the groups, while transfer of feedback effects was measured by student performance on follow-up tasks. Results show that students receiving elaborate feedback solved more tasks and made fewer errors than students receiving only verification feedback, which suggests that students benefited from the elaborate feedback.

Keywords: Domain reasoner, Hypothesis testing, Intelligent tutoring systems, Statistics education.

1 Introduction

Hypothesis testing is widely used in scientific research, and is therefore covered in most introductory statistics courses in higher education [2]. The topic is challenging for many students, because it requires an ability to follow a complex line of reasoning involving several abstract concepts and uncertainty [4; 6]. Students struggle to understand the role and interdependence of the concepts, or, in other words, the structure of hypothesis tests [14]. Appropriate feedback might support students in comprehending this structure. It should not only address the content of a current step, but also its relation to earlier steps. An Intelligent Tutoring System (ITS) can provide such sophisticated feedback on the level of steps and can provide diagnostics of student errors [11]. Feedback on the step level is generally more effective than feedback on the level of complete solutions [16].

Although ITSs vary considerably in design, they generally contain an expert knowledge module, a student model module, a tutoring module, and a user interface module [11]. Of these four components, the expert knowledge module, also referred to as domain reasoner [7], is the most domain-dependent. Two important paradigms for constructing domain reasoners are model-tracing, in which the ITS checks that a student follows the rules of a model solution [1], and constraint-based modeling, in which the

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-23207-8_52

ITS checks whether a student violates constraints [10]. There exist ITSs that support hypothesis testing based on either of these approaches [9]. We combined the two in a single ITS supporting hypothesis tests. The contribution of this paper is a thorough evaluation of the impact of the combined ITS's feedback, which especially addresses the structure of hypothesis tests, on students' problem-solving behavior. It is guided by the question: does automated intelligent feedback on the structure of hypothesis tests contribute to student proficiency in carrying out hypothesis tests?

2 Methods

The domain reasoner for hypothesis testing is based on the Ideas framework [8], with a model-tracing approach as starting point, adding constraint-based modeling to identify inconsistencies in solution structure. For a description of its design, see [13].

The study consisted of a randomized controlled experiment in the context of a compulsory statistics course for first-year psychology students at a Dutch university. Students enrolled in the course were divided randomly into an experimental group (310 students) and a control group (309 students). Consent for the study was given by 163 students in the experimental group and 151 students in the control group. Participants were between 17 and 31 years old (*M* = 19.3, *SD* = 1.7) and 77% were female.

In five weeks of the ten-week course students received online homework sets in the Freudenthal Institute's Digital Mathematics Environment (DME; see [3]). The three homework sets that concerned hypothesis testing each contained two tasks in which students were asked to construct hypothesis tests by selecting steps from a drop-down menu and to completing these steps. For an example, see [13].

Two versions of the homework sets were designed: an experimental version with feedback on steps in the hypothesis testing procedure by the domain reasoner, and a control version with verification feedback on the contents of single steps only. Consequently, in the experimental version correct solutions needed to include four essential steps, since otherwise constraints would be violated. In the control version correct solutions only needed to include a correct conclusion about the null hypothesis.

Data for this study consisted of logs of the students' actions on the online homework sets, including all attempts students made to find correct answers, and all feedback requests. After exporting the logs from the DME, logs from students who did not give consent were deleted and all other logs were anonymized.

Three measures were used to assess immediate effects of feedback condition on the students' ability to solve hypothesis testing tasks: the number of tasks in which students attempted to construct steps, the number of tasks that students solved, and the number of errors students made in hypothesis test structure. Since samples were large, independent samples *t*-tests were used for all comparisons between groups [5]. Besides *t*-tests to compare groups over all tasks simultaneously, graphical representations were used to assess the differences between groups over time.

As promising effects of feedback on student performance do not automatically guarantee transfer to new tasks [12], student performance on follow-up tasks was also evaluated. From the three homework sets follow-up tasks on hypothesis testing were selected. For each student who received feedback on constructed steps at least once the ratio between number of selected tasks immediately answered correct and number of selected tasks attempted was calculated and ratios were compared between groups.

3 Results

Table 1. Mean number of tasks students worked on, constructed steps for and solved

	Experimental group (*N* = 163)	Control group (*N* = 151)	*t* (df = 312)	*p*
Tasks worked on	4.8 (1.5)	4.9 (1.5)	0.86	.391
Tasks tried constructing steps	3.8 (1.7)	3.9 (1.6)	0.62	.537
Tasks with complete solution	1.7 (1.8)	2.0 (1.7)	1.33	.184

In the hypothesis testing tasks students could choose to only fill in final answers, without constructing steps. Table 1 contains the mean number of tasks students worked on, the mean number of tasks in which they attempted to construct steps, and the mean number of complete solutions. In both groups, students attempted to construct steps for almost 80% of the tasks they worked on. The *t*-tests yielded no significant differences between groups. For the number of complete solutions, however, examining individual tasks did reveal different patterns. Figure 1 (left) displays the percentage of students who found complete solutions per task, as percentage of students who attempted to construct steps. For the first three tasks the control group outperformed the experimental group, while for the latter three tasks this was reversed.

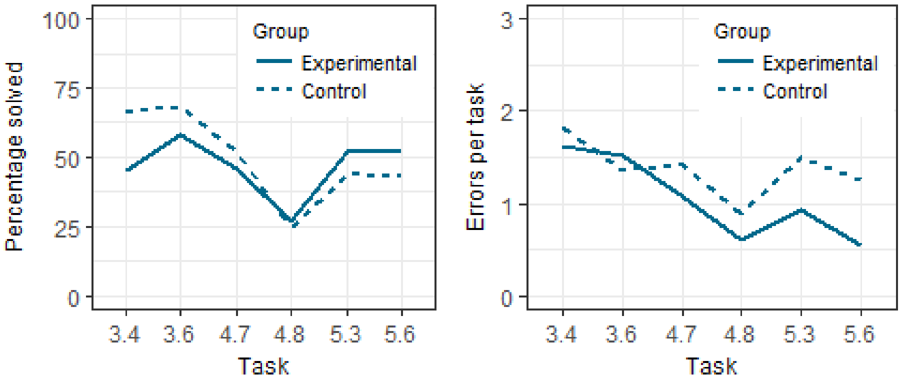

Fig. 1. Percentage of students who correctly solved tasks according to group's assessment criteria (left) and mean number of errors in solution structure (right)

The final measure of immediate feedback effects was the number of errors students made in the structure of their hypothesis tests. The domain reasoner could diagnose 15 different errors in hypothesis test structure, such as a missing alternative hypothesis. On average, students in the experimental group made 1.12 (*SD* = 0.79) different structure errors per solution, while students in the control group made 1.42 (*SD* = 0.86) errors, which was significantly more, $t(312) = 3.22$, $p = .001$, Cohen's $d = .36$. The graph in Figure 1 (right) shows that in both groups the number of structure errors decreased over tasks, but this trend was stronger in the experimental group.

Regarding transfer to follow-up tasks, students in the experimental group ($N = 158$) and the control group ($N = 147$) were found to perform similarly: the mean ratio of correct answers was 0.72 ($SD = 0.07$) in the experimental group and 0.71 ($SD = 0.08$) in the control group. This implies that the domain reasoner feedback did not lead to better performance on follow-up tasks than verification feedback alone.

4 Conclusion and discussion

We have evaluated the influence of ITS feedback addressing hypothesis test structure on student proficiency in carrying out hypothesis tests. The ITS feedback seemed to affect students' success in solving tasks completely; while students receiving ITS feedback performed worse than students receiving only verification feedback on the first three tasks, they outperformed the control group in the final three tasks, even with stricter assessment criteria. Additionally, students receiving ITS feedback made significantly fewer errors in hypothesis test structure than students receiving verification feedback only. This suggests that after familiarization, the ITS feedback effectively supported students in resolving their misunderstandings. This is in line with earlier findings that elaborate feedback is more effective than verification feedback [15]. Performance on follow-up tasks did not differ between groups, which implies that there was no automatic transfer from the positive results of the ITS feedback.

Such a lack of transfer has been found more often [12]. Here it could be caused by the design of the follow-up tasks, none of which specifically addressed the structure of hypothesis tests. From a research perspective, availability of tasks addressing the structure could have provided more insight in transfer of ITS feedback effects. From an educational perspective, availability of such tasks would have been valuable too, to avoid that students rely too much on the ITS feedback [12].

A second limitation of the study was that in this first large-scale implementation of the domain reasoner inevitably some unclarities became apparent. Nonetheless, even though sometimes receiving confusing feedback, students in general kept attempting the tasks and, as the results above show, did still benefit from the feedback.

Overall, this study has demonstrated that combining the model-tracing and constraint-based modeling paradigms can result in effective feedback on the structure of hypothesis tests. A challenging aspect of hypothesis testing that is not yet addressed by the ITS feedback is the role of uncertainty in the interpretation of the results from hypothesis tests [4]. Future research could focus on broadening the scope of the domain reasoner for hypothesis testing to include this reasoning with uncertainty.

Acknowledgments. We thank teachers Jeltje Wassenberg-Severijnen and Corine Geurts for their collaboration in designing teaching tasks and delivering the course. Furthermore, we thank Noeri Huisman, Martijn Fleuren, Peter Boon and Wim van Velthoven who helped with developing the domain reasoner.

References

1. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*(2), 167–207.
2. Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., . . . Wood, B. (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016. *American Statistical Association*. http://www.amstat.org/education/gaise
3. Drijvers, P., Boon, P., Doorman, M., Bokhove, C., & Tacoma, S. (2013). Digital design: RME principles for designing online tasks. In C. Margolinas (Ed.), *Proceedings of ICMI Study 22 Task Design in Mathematics Education* (pp. 55–62). Clermont-Ferrand, France: ICMI.
4. Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*(1), 75–98.
5. Field, A. (2009). *Discovering statistics using SPSS* (3th ed.). London: Sage Publications.
6. Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). Learning to reason about statistical inference. *Developing students' statistical reasoning* (pp. 261–288) Springer, Dordrecht.
7. Goguadze, G. (2011). *ActiveMath - generation and reuse of interactive exercises using domain reasoners and automated tutorial strategies* (Doctoral dissertation, Saarland University, Saarbrücken, Germany). Retrieved from https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/26153/1/goguadzeDiss2011.pdf
8. Heeren, B., & Jeuring, J. (2014). Feedback services for stepwise exercises. *Science of Computer Programming, 88*, 110–129. https://doi.org/10.1016/j.scico.2014.02.021
9. Kodaganallur, V., Weitz, R. R., & Rosenthal, D. (2005). A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal of Artificial Intelligence in Education, 15*(2), 117–144.
10. Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: The constraint-based approach. *IEEE Intelligent Systems,* (4), 38–45
11. Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review, 4*(4), 251–277.
12. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
13. Tacoma, S., Heeren, B., Jeuring, J., & Drijvers, P. (2019). Automated feedback on the structure of hypothesis tests. In: U.T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education,* (pp. xx–yy). Utrecht, the Netherlands: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME.
14. Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute: Proceedings of the Fifty-Second Session of the International Statistical Institute, 58*, 201–204.
15. Van der Kleij, F., Feskens, R., & Eggen, T. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes. A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881
16. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369