USING THEOREMA AS A PROOF TUTOR IN THE CLASSROOM

Wolfgang Windsteiger

Research Institute for Symbolic Computation (RISC)
Johannes Kepler University Linz (JKU)

RISC Forum — January 17, 2022

JOHANNES KEPLER
UNIVERSITY LINZ

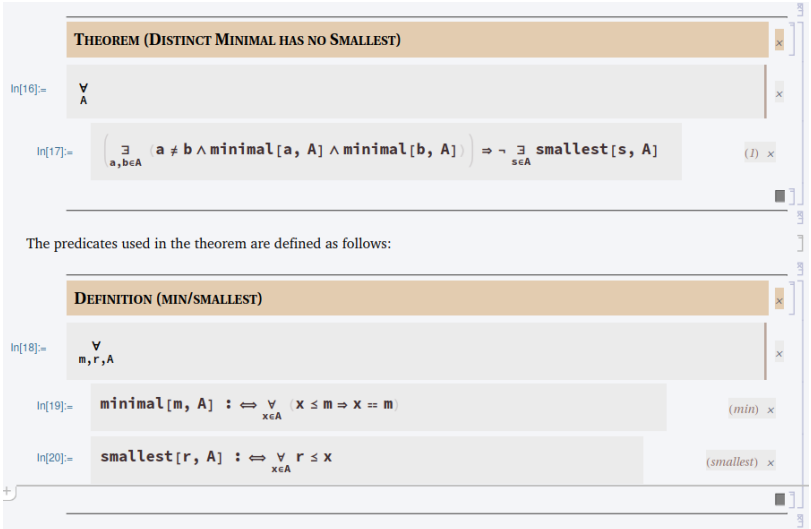
INTRODUCTION: A NEW MODERN LOGIC COURSE

- Modern topics in addition to traditional ones . . .
  - Module Propositional Logic + SAT
  - Module Predicate Logic
    + Pragmatics: How to specify problems? How to do real mathematical proofs?How to do real mathematical proofs?
  - Module Satisfiability Modulo Theories (SMT)
- Modern presentation by showing “logic in action” with logic software.
  - Limboole (SAT solver)
  - RISC-AL (by W. Schreiner)
  - TheoremaTheorema
  - Z3, Yices, CVC4, Boolector (SMT Solvers)
- Modern grading
  - Minitests, bonus exercises, lab exercises.
  - No final exam.

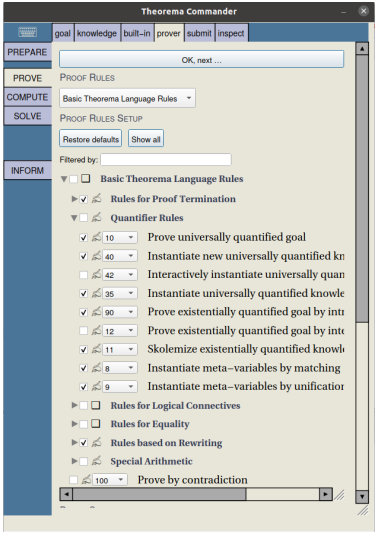
WHY AUTOMATED THEOREM PROVING IN THE COURSE?

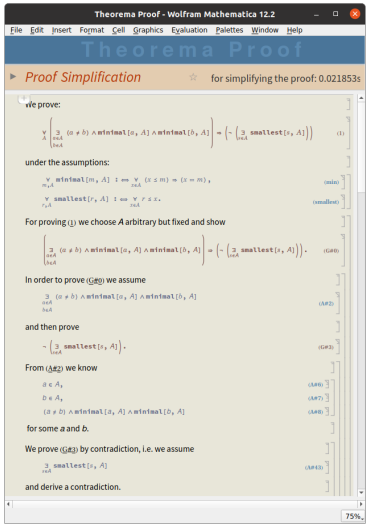
- One of the teaching goals of the course (Module Predicate Logic):
  Students should be able to do (simple) mathematical proofs
  by hand correctly and completely.
- Method:
  Use software (Theorema) as tutoring system for students
  on a voluntary basis in the frame of bonus exercises.

THEOREMA DEMO



THEOREMA DEMO





HOW THEOREMA IS USED IN THE COURSE

- Structure of Module Predicate Logic B:

- Theorema only in voluntary parts (bonus and lab exercises).
- Bonus exercises: students submit automated proofs for problems of previous exercise, which they already did by hand.
- Lab exercise: students generate automated proof and submit a proof done by hand for the same problem.

THEOREMA AS A PROOF TUTOR

Our didactical hypothesis:

Students can impove their performance in proving when . . .

- . . . they watch, which steps Theorema uses in order to do a proof and
- . . . they watch, how Theorema presents a proof in “(almost) natural language”.

We try to avoid difficulties in handling the Theorema system by

- providing notebooks containing all formulas and by
- providing hints for the prover configuration (if necessary).

PERFORMANCE IN MINITESTS

We show p-values of a one-sided Student T-Test testing for equal mean values, i.e. $p \leq 0.05$ says that mean values differ statistically significantly.

- Minitest 2: Group “Bonus 1” is better than all others whereas Group “no Bonus” is worse even than average.

	⌀	All	Bonus 1
All (307)	3.28	—	—
Bonus 1 (139)	3.62	0.002	—
no Bonus (168)	3.00	0.006	1.21×10^{-6}

- Population of groups (in parentheses) high $\rightsquigarrow$ no random numbers!

PERFORMANCE IN MINITESTS

- Minitest 3: Group “Bonus 1+2” is significantly better than Group “no Bonus”.

	∅	All	Bonus 1	Bonus 1+2
All (286)	3.34	—	—	—
Bonus 1 (135)	3.42	0.20	—	—
Bonus 1+2 (104)	3.47	0.10	0.33	—
no Bonus (141)	3.26	0.22	0.08	0.04

- Group “Bonus 1” is almost significantly better than Group “no Bonus”.
- Group “Bonus 1+2” is almost significantly better than average.

IMPACT ON MATHEMATICS SKILLS IN GENERAL

- Exam Discrete Structures: Group “all Bonus exercises” is significantly better than Group “no Bonus” and better than average.

		all	Bonus=3	Bonus=0	with Lab	Lab+B=3
	ø	13.56	14.73	13.19	13.70	15.00
all	13.56		0.0240	0.1778	0.4472	0.1028
Bonus=3	14.73			0.0078	0.1882	0.4082
Bonus=0	13.19				0.3195	0.0636
with Lab	13.70					0.1866
Lab+B=3	15.00					↑
		low number, all score 16 and one scores 9				

- Group “Lab+Bonus” is spoiled by one weak participant, otherwise . . .

SELF-ASSESSMENT QUESTIONS: SUCCESSFUL PROOF

1. I did not try or was not able to do the examples by hand, but now I think would be able to do them.
2. I did not try or was not able to do the examples by hand. I think I would still not be able to do such proofs.
3. I had no problems doing the proofs by hand. However, they are different from the Theorema proofs and I'm confused now whether my proofs are wrong.
4. I had no problems doing the proofs by hand. However, they are slightly different from the Theorema proofs because Theorema uses certain rules that I did not know. Still, I think my proofs are fine.
5. I had no problems doing the proofs by hand. However, they are slightly different from the Theorema proofs and in the future I would do my proofs differently.
6. I had no problems doing the proofs by hand. After doing the proofs with Theorema I realized that at least one of my original proofs was wrong.
7. I had a hard time doing the proofs by hand. However, I think when doing the next proof by hand, it will be equally difficult, doing the proof with Theorema did not help me for improving my own skills.
8. I had a hard time doing the proofs by hand. After doing the proof with Theorema I understand much better how all of this works. I feel that my own skills improved by using Theorema.
9. I don't see any connection between the examples from the exercises and the Bonus Exercise with Theorema

SELF-ASSESSMENT QUESTIONS: PROOF FAILURE

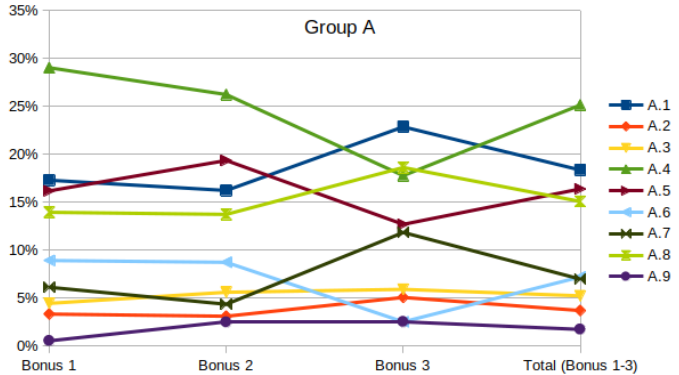
10. I did not try or was not able to do these examples by hand. I wanted to see how Theorema does the proofs, but I failed to produce a compete proof.
11. I did not try or was not able to do these examples by hand. Theorema is much too complicated for me to use it for such exercises.
12. I had no problems doing the proofs by hand. Unfortunately, I failed to produce a complete proof with Theorema. It would have been interesting to compare.
13. I had no problems doing the proofs by hand. I'm not interested how an automated proof looks, I have done them by hand anyway.
14. I had a hard time doing the proofs by hand. Unfortunately, I failed to produce a complete proof with Theorema. It would have been interesting to compare.
15. I had a hard time doing the proofs by hand. I'm not interested how an automated proof looks, I have done them by hand anyway.
16. I don't see any connection between the examples from the exercises and the Bonus Exercise with Theorema.

SELF-ASSESSMENT OF STUDENTS: GROUP SIZES

- by hand: 1–2 not able, 3–6 no problems, 7–8 hard time, 9 no connection

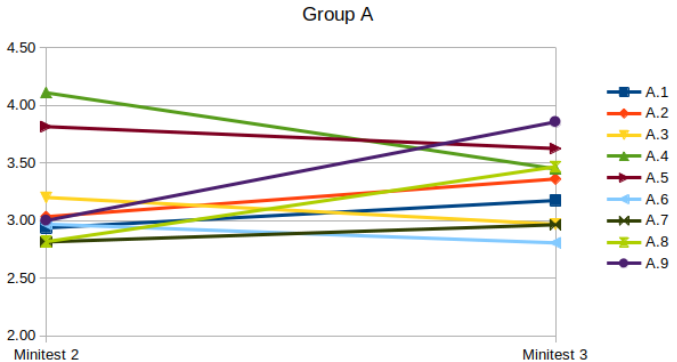
	Submissions	Surveys	Reasons for Success (Group A)								
Bonus 1	157	274	179	65%							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
			31	6	8	52	29	16	11	25	1
		100%	17%	3%	4%	29%	16%	9%	6%	14%	1%
Bonus 2	147	251	160	64%							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
			26	5	9	42	31	14	7	22	4
		100%	16%	3%	6%	26%	19%	9%	4%	14%	3%
Bonus 3	100	180	118	66%							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
			27	6	7	21	15	3	14	22	3
		100%	23%	5%	6%	18%	13%	3%	12%	19%	3%
Total (Bonus 1-3)		705	457	65%							
			84	17	24	115	75	33	32	69	8
		100%	18%	4%	5%	25%	16%	7%	7%	15%	2%
Overall (Groups A and B)			**12%**	**2%**	**3%**	**16%**	**11%**	**5%**	**5%**	**10%**	**1%**

SELF-ASSESSMENT: GROUP SIZE DEVELOPMENT



- Top 4 vs. rest always 3:1
- In bonus 3 the top answer is 1: “not able by hand but now I would be”.
- Big gains in bonus 3: 1,7,8 (not able/hard time), drop: 4–6 (no problems).

SELF-ASSESSMENT VS. PERFORMANCE



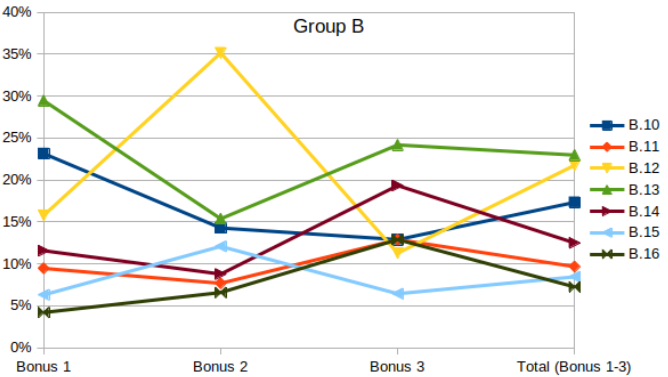
- Interesting: A.8 (hard time by hand but after Theorema yes): rank 14 $\rightarrow$ rank 4.
- Interesting: A.9 (no connection): rank 8 $\rightarrow$ rank 1.
- A.1 (not able by hand but after Theorema yes): rank 10 (but second-biggest group!).

SELF-ASSESSMENT OF STUDENTS: GROUP SIZES

- by hand: 10–11 not able, 12–13 no problems, 14–15 hard time, 16 no connection

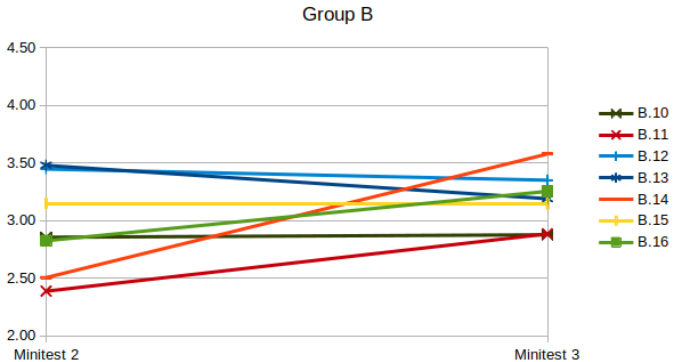
	Submissions	Surveys	Reasons for Failure (Group B)						
Bonus 1	157	274	95	35%					
			10	**11**	**12**	**13**	**14**	**15**	**16**
			22	9	15	28	11	6	4
		100%	23%	9%	16%	29%	12%	6%	4%
Bonus 2	147	251	91	36%					
			10	**11**	**12**	**13**	**14**	**15**	**16**
			13	7	32	14	8	11	6
		100%	14%	8%	35%	15%	9%	12%	7%
Bonus 3	100	180	62	34%					
			10	**11**	**12**	**13**	**14**	**15**	**16**
			8	8	7	15	12	4	8
		100%	13%	13%	11%	24%	19%	6%	13%
Total (Bonus 1-3)		705	248	35%					
			43	24	54	57	31	21	18
		100%	17%	10%	22%	23%	13%	8%	7%
Overall (Groups A and B)			**6%**	**3%**	**8%**	**8%**	**4%**	**3%**	**3%**

SELF-ASSESSMENT: GROUP SIZE DEVELOPMENT



- Less clear picture.
- Big drop bonus 3: B.12 (no problems by hand, wanted to compare).

SELF-ASSESSMENT VS. PERFORMANCE



- Interesting: B.14 (hard time by hand, wanted to compare): rank 15 $\rightarrow$ rank 3.

ALL DATA: MINITEST 2

		overall	Bonus 1	Bonus 1+2	no Bonus	A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	A.9	B.10	B.11	B.12	B.13	B.14	B.15	B.16
	⌀	3.28	3.62		3.00	2.94	3.03	3.20	4.11	3.81	2.97	2.82	2.82	3.00	2.86	2.39	3.45	3.48	2.51	3.15	2.83
overall	**3.28**		0.0015		0.0067	0.040	0.2797	0.4322	3E-07	0.009	0.1151	0.0837	0.0271	only 1	0.0352	0.0275	0.2933	0.1729	0.0193	0.3971	0.0607
Bonus 1	**3.62**				1.21E-06	0.001	0.0994	0.1986	0.002	0.1997	0.0102	0.0143	0.0011	only 1	0.0014	0.0075	0.282	0.2501	0.0035	0.1858	0.0127
no Bonus	**3.00**					0.378	0.4684	0.3413	4E-10	0.0004	0.4532	0.2878	0.2245	only 1	0.2705	0.0819	0.0786	0.0155	0.0808	0.3845	0.2485
A.1	**2.94**						0.4149	0.304	1E-06	0.001	0.4581	0.3683	0.3374	only 1	0.386	0.1135	0.0726	0.0225	0.1258	0.3448	0.3508
A.2	**3.03**							0.3941	0.0203	0.0585	0.4461	0.3354	0.3219	only 1	0.3508	0.1332	0.208	0.1711	0.1588	0.4273	0.3285
A.3	**3.20**								0.0496	0.1269	0.3331	0.25	0.2355	only 1	0.2569	0.1004	0.3289	0.2955	0.1185	0.4706	0.2404
A.4	**4.11**									0.117	0.0002	0.001	8E-06	only 1	1E-05	0.0012	0.0251	0.0051	0.0003	0.0518	0.0014
A.5	**3.81**										0.0058	0.0072	0.0009	only 1	0.0011	0.0037	0.1527	0.1182	0.0015	0.1210	0.0039
A.6	**2.97**											0.3493	0.3229	only 1	0.3641	0.1134	0.1075	0.0553	0.1289	0.3722	0.3335
A.7	**2.82**												0.4983	only 1	0.458	0.2007	0.0738	0.0425	0.2445	0.2848	0.4896
A.8	**2.82**													only 1	0.4507	0.1775	0.0464	0.0145	0.2136	0.2729	0.4895
A.9	**3.00**													only 1	only 1	only 1	only 1	only 1	only 1	only 1	only 1
B.10	**2.86**															0.1565	0.0554	0.0184	0.1852	0.2954	0.4618
B.11	**2.39**																0.0224	0.0145	0.408	0.1212	0.1740
B.12	**3.45**																	0.4637	0.0194	0.3040	0.0533
B.13	**3.48**																		0.0093	0.2725	0.0254
B.14	**2.51**																			0.1439	0.2093
B.15	**3.15**																				0.2780

ALL DATA: MINITEST 3

		overall	Bonus 1	Bonus 1+2	no Bonus	A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	A.9	B.10	B.11	B.12	B.13	B.14	B.15	B.16
	⌀	3.34	3.42	3.47	3.26	3.17	3.36	2.97	3.45	3.62	2.81	2.96	3.47	3.86	2.88	2.88	3.35	3.19	3.58	3.14	3.25
overall	**3.34**		0.1964	0.0974	0.2225	0.2109	0.4839	0.1835	0.2435	0.0411	0.0174	0.2252	0.2734	0.0382	0.0314	0.0255	0.4715	0.3163	0.2557	0.2981	0.4223
Bonus 1	**3.42**			0.3260	0.0822	0.1226	0.4624	0.1393	0.4298	0.1170	0.0093	0.1826	0.4135	0.0562	0.0175	0.0138	0.3709	0.2320	0.3328	0.2287	0.3491
Bonus 1+2	**3.47**				0.0400	0.0829	0.4281	0.1154	0.4472	0.1947	0.0061	0.1585	0.4919	0.0752	0.0117	0.0092	0.2820	0.1859	0.3874	0.1900	0.3052
no Bonus	**3.26**					0.3377	0.4359	0.2360	0.1391	0.0199	0.0350	0.2721	0.1818	0.0254	0.0594	0.0480	0.3345	0.4087	0.2000	0.3737	0.4929
A.1	**3.17**						0.3847	0.3192	0.1322	0.0354	0.1090	0.3410	0.1520	0.0169	0.1561	0.1446	0.2577	0.4820	0.1639	0.4702	0.4300
A.2	**3.36**							0.2921	0.4438	0.3380	0.2052	0.3001	0.4335	0.2277	0.2333	0.2347	0.4935	0.3993	0.3769	0.3772	0.4408
A.3	**2.97**								0.1336	0.0695	0.3600	0.4971	0.1356	0.0322	0.4190	0.4236	0.1937	0.3257	0.1276	0.3686	0.3061
A.4	**3.45**									0.2070	0.0119	0.1738	0.4711	0.0740	0.0208	0.0156	0.3431	0.2212	0.3698	0.2172	0.3298
A.5	**3.62**										0.0026	0.1070	0.2722	0.1945	0.0050	0.0036	0.1317	0.1015	0.4518	0.1134	0.2055
A.6	**2.81**											0.3819	0.0186	0.0024	0.4113	0.3948	0.0356	0.1539	0.0418	0.2118	0.1754
A.7	**2.96**												0.1721	0.0582	0.4335	0.4381	0.2287	0.3423	0.1531	0.3790	0.3198
A.8	**3.47**													0.1031	0.0297	0.0243	0.3394	0.2238	0.3945	0.2172	0.3223
A.9	**3.86**														0.0039	0.0041	0.0486	0.0409	0.2520	0.0509	0.1077
B.10	**2.88**															0.4890	0.0558	0.1999	0.0555	0.2610	0.2130
B.11	**2.88**																0.0468	0.1952	0.0531	0.2600	0.2121
B.12	**3.35**																	0.3254	0.2876	0.3043	0.4159
B.13	**3.19**																		0.2026	0.4601	0.4494
B.14	**3.58**																			0.1941	0.2748
B.15	**3.14**																				0.4187

CONCLUSION

- Classroom experiment using the automated theorem proving software Theorema in the teaching of logic.
- Software is applied to aid the learning process of students.
- Tutoring-by-software correlates with students' performance.
- Students' experiences being tutored by software.
- Those who had a hard time doing proofs by hand and claimed an improvement of their understanding through being tutored by software showed a significant improvement from one exam to the next.


JOHANNES KEPLER
UNIVERSITY LINZ