Appendix A. Overview of the experts' quantitative ratings in EVAL2

Abbrev.	Relev. RQ	Undst.	Real-word fid.	Compl.	Relev. DO1	Relev. DO2	Relev. DO3	Relev. DO4	Relev. DO5
IE1	1	0	2	1	2	3	2	-1	2
IE2	2	2	2	2	3	0	3	3	2
IE3	3	3	1	2	3	1	3	3	3
IE4	3	2	2	3	3	1	2	2	3
IE5	3	3	2	3	3	2	1	2	1
IE6	3	2	2	1	3	2	3	2	1
IE7	2	2	2	2	3	3	3	2	1
RE1	3	3	1	3	2	3	2	3	3
RE2	3	3	3	2	3	3	1	2	3
RE3	3	3	1	2	3	2	1	0	2

Table A.1: Ratings for the Relevance of our Research Question, Understandabilty, Real-World Fidelity, Completeness and Relevance of our DOs (left to right) on a Likert Scale (-3 to 3) by our Interview Participants.

Abbrev.	Impl. DO1	Impl. DO2	Impl. DO3	Impl. DO4	Impl. DO5
IE1	2	3	2	3	2
IE2	3	3	3	2	3
IE3	1	2	2	2	3
IE4	2	3	3	2	3
IE5	3	3	3	2	3
IE6	3	2	2	2	3
IE7	2	2	1	0	2
RE1	2	3	2	3	3
RE2	3	3	1	1	3
RE3	2	3	3	3	3

Table A.2: Ratings for the Implemetation of our DOs in our RA on a Likert Scale (-3 to 3) by our Interview Participants.

Appendix B. Overview of the quantitative ratings in EVAL4

	Helpfulness of the prototype in labeling	Efficiency of the prototype in labeling	Usability of the prototype for labeling	Expectation fulfillment by the prototype
Expert's rating	3	2	2	3
	Meaningfulness of the process instances	Ability to identify new process steps	Confidence in clustering	
Expert's rating	2	2	2	

Table A.1: Quantitative ratings from the dataset author on a Likert Scale (-3 to 3) on the results produced by our prototype.