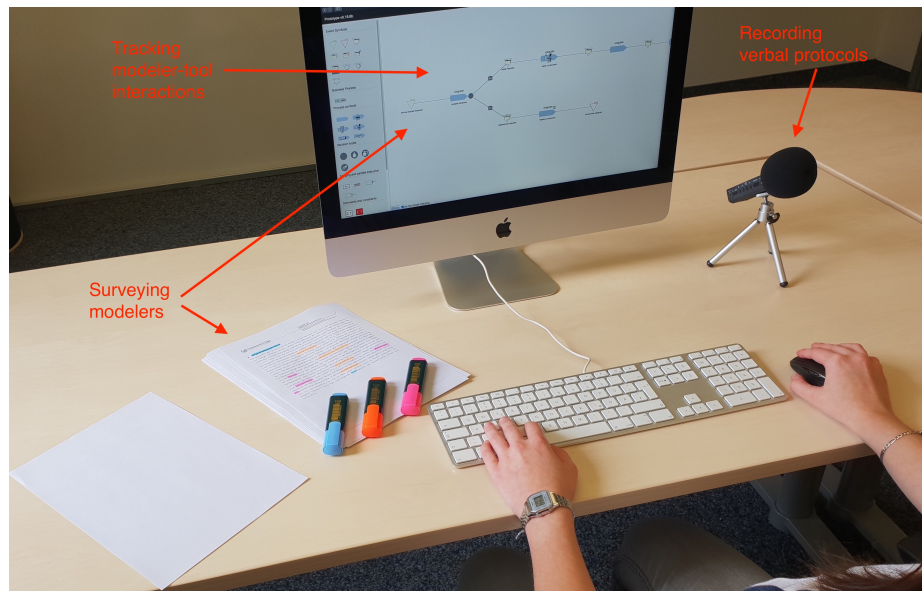


## Supplementary material for “Modeling Difficulties in Data Modeling: Similarities and Differences between Experienced and Non-experienced Modelers”

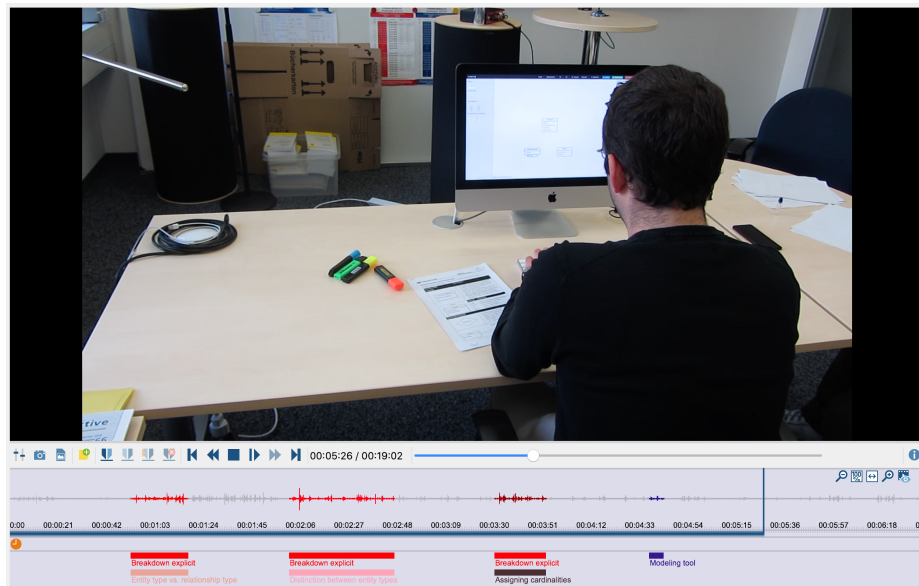
Rosenthal, K., Strecker, S., Pastor, O.: Modeling Difficulties in Data Modeling: Similarities and Differences between Experienced and Non-experienced Modelers. In: ER 2020. Vienna, Austria (2020)

**Table 1.** Observation modes.

Step	Activities
(a) Recording verbal protocols:	This mode targets the reasoning of modelers while modeling via verbalization, and aims at understanding cognitive processes during modeling. Subjects are instructed to verbalize all their thoughts during the modeling, and their comments are recorded.
(b) Videotaping modelers:	Subjects are videotaped from an ‘over-the-shoulder’ perspective to capture the overall interaction with the written material and the software tool during the modeling and the modelers’ behavior as movements and gestures entailing additional information on their modeling process, e.g., cues about modeling challenges and difficulties.
(c) Recording modeler tool-interactions:	This mode of observation aims to observe modelers’ interactions with the graphical editor of the modeling tool. Every interaction during the modeling is recorded as a time-discrete event as basis for visualizing the modeling processes.
(d) Surveying modelers:	Subjects fill in a survey comprising closed-ended and open-ended questions before modeling and one after modeling. The aim is to gather additional data on prior experience, a retrospective self-assessment of the modeling process and demographic information—aimed at achieving an overview of the sample of subjects and to identify peculiarities and outliers.



**Fig. 1.** Multi-modal observations (Video recording of a modeler from an ‘over-the-shoulder’ perspective).



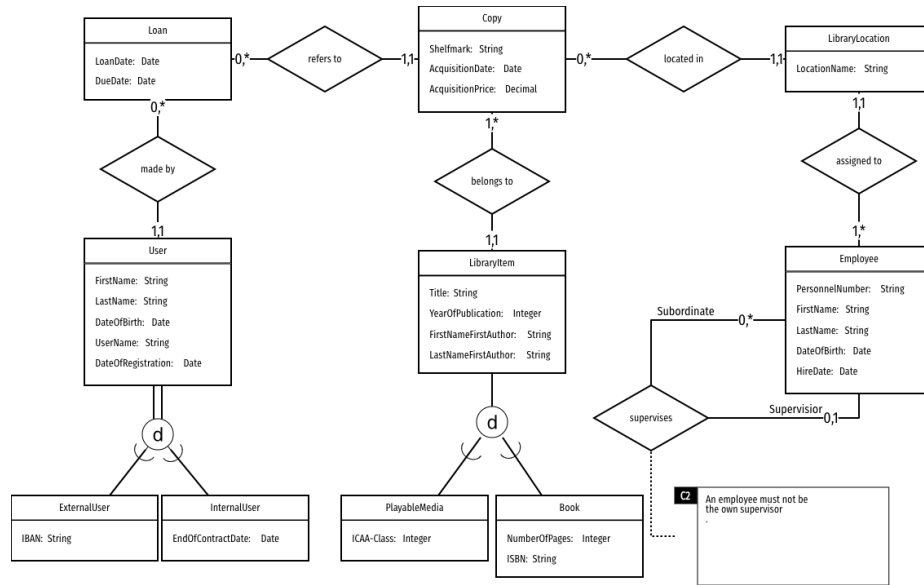
**Fig. 2.** Coding audio-visual protocols in MAXQDA (screenshot from a pre-test).

**Table 2.** Data collection procedure.

Step	Activities
(1)	After being informed about the purpose of the study and completing a consent form, subjects were instructed to study a short description of the semantics of the modeling concepts and the graphical notation of the ER model by themselves.
(2)	This was followed by watching a short video introduction into the used modeling tool of ca. 3 minutes.
(3)	As next step, the subject was provided think aloud instructions. The observer instructed the subject to verbalize all thoughts while modeling “as if alone in the room” and informed the subject that reminders would be given after a predetermined period of silence of 30 seconds—except the subject is reading the modeling task—with the precise wording “Please keep talking”.
(4)	In a warm-up modeling task, each subject was asked to construct a simple data model, to become familiar with the modeling tool and to practice verbalizing thoughts while modeling. The observer subsequently answered the subject’s questions about the procedure.
(5)	Next, the subject was required to fill in a pre-modeling survey asking closed-ended and open-ended questions on prior conceptual modeling experience and perceived familiarity with the domain of the modeling task, and a test with six yes/no-type questions on theoretical knowledge of conceptual data modeling with the ER model.
(6)	Next, the subject was given the main modeling task from the library domain described in natural language and presented on paper. Each subject was instructed to use the modeling tool to construct a conceptual data model based on the natural language description. During modeling, the verbalizations of the participant and a video of the modeler’s behavior were collected using a camcorder. The participant was requested to let the observer know when he/she had finished the task. The observer terminated the modeling at a convenient moment after about 45 minutes.
(7)	After completing the main modeling task, each participant was required to fill in a post-modeling survey comprising closed-ended and open-ended questions on encountered modeling difficulties, difficulties with think aloud and domain knowledge, as well as demographic information.

**Main modeling task:** As part of a project for the introduction of a new information system in the university library, you are asked to create a conceptual data model that reconstructs the following facts representing a simplified description of a library:

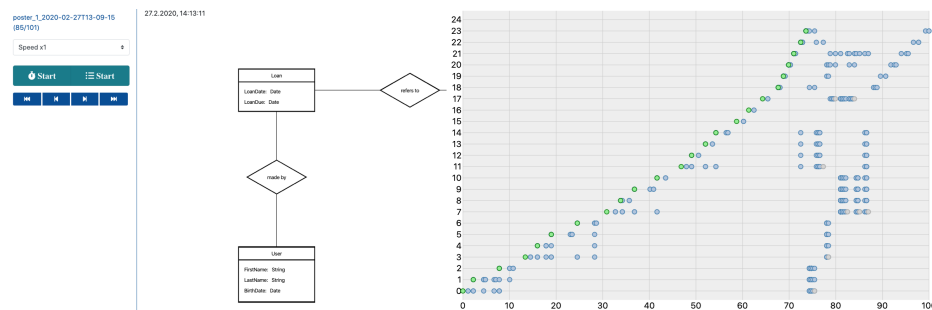
- The current stock of the university library consists of library items, which comprise books and playable media such as DVD. Beyond books and playable media, there are further library items as, for example, journals and microfiches. Library items are described by title and year of publication. In addition, the first name and last name of the first author of the library item are recorded. For books, the number of pages is recorded as well as the International Standard Book Number (ISBN) for unique identification. For playable media, the classification of the ICAA (Instituto de Cinematografía y Artes Audiovisuales - Cinematography and Audiovisual Arts Institute) is recorded which comprises 6 different classes of media defined by the ICAA.
- There may be one or more copies of a library item. Each copy of a library item has a unique shelfmark. To determine the age of the library's holdings, the acquisition date is recorded for each copy. In order to be able to identify particularly valuable items, the acquisition price is also recorded for each copy.
- The university library has several different physical locations (different buildings). Copies of library items can be found at exactly one of the university library's locations. A name is assigned to each location. Library employees are assigned to exactly one library location. At least one employee works at each location.
- All employees of the library are assigned an alphanumeric personnel number and their first name and last name, date of birth and hire date are recorded. Employees can have a supervisor. Thereby, employees can be supervisors for any number of other employees.
- The university library is used by internal users (members of the university) and external users (non-members of the university). Users of the library must be natural persons. When registering as user, the first name, last name, date of birth as well as the date of registration are entered and a unique alphanumeric user name is assigned. For external users, the International Bank Account Number (IBAN) for the collection of fines for overdue items is recorded, which is not considered necessary for internal users. For internal users, the date is recorded at which the contract with the university terminates. This is initially the case for one year.
- Users of the university library can borrow any number of copies of library items. A loan always refers to exactly one copy. For a loan, the date of the loan and the due date are recorded in order to be able to determine if the loan is overdue.



**Fig. 3.** Reference solution for main modeling task.

**Table 3.** Coding scheme for coding the audio-visual protocols. Codes marked with an asterisk (\*) are codes which emerged during coding.

Category	Cognitive breakdowns	General codes
(Sub) Codes	<b>Cognitive breakdown</b> <ul style="list-style-type: none"> <li>– Decide between attribute/generalization*</li> <li>– Choose data type of attribute*</li> <li>– Specify generalization hierarchies*</li> <li>– Decide between entity/relationship type*</li> <li>– Establish relationship types*</li> <li>– Model recursive relationship types*</li> <li>– Develop identifiers for relationship types*</li> <li>– Determine cardinalities*</li> <li>– Counteract compromising of model integrity*</li> </ul>	<b>Actions outside of the modeling tool</b> <ul style="list-style-type: none"> <li>– Reading/marking the modeling task*</li> <li>– Paper-based modeling*</li> <li>– Taking notes*</li> </ul> <b>Non-task-related issues</b> <ul style="list-style-type: none"> <li>– Modeling tool/notebook*</li> <li>– Think aloud*</li> <li>– Variant of the ER model*</li> <li>– Completion time*</li> </ul> <b>Evaluation of the task at meta-level</b> <ul style="list-style-type: none"> <li>– Problem of comprehension*</li> </ul> <b>Silent periods</b> <ul style="list-style-type: none"> <li>– Arrange visual layout of model*</li> </ul>



**Fig. 4.** Replay of a modeling process (left) and dot diagram visualizing a modeling process (right).

PLEASE ANSWER THE QUESTIONS IN THE ORDER GIVEN AND MARK THE CORRECT ANSWER WITH A CROSS OR FILL IN THE TEXT FIELD READABLE. IF YOU MADE A MISTAKE, PLEASE BLACKEN THE WHOLE BOX.

FIRST, WE ASK YOU FOR INFORMATION ABOUT YOUR PRIOR EXPERIENCE IN CONCEPTUAL MODELING, FOR EXAMPLE, IN YOUR STUDIES OR IN PRACTICE.

years and  month

(YYYY, e.g., 1996)

models

☐ ☐ ☐ models

models

[illegible]

(Please fill in)

**Fig. 5.** Pre-modeling questionnaire – Part 1

<p><b>THE FOLLOWING QUESTIONS REFER TO YOUR KNOWLEDGE OF THE CHOSEN VARIANT OF THE ER MODEL. PLEASE AGREE TO THE STATEMENTS ("AGREE") OR DISAGREE ("DISAGREE").</b></p>				
<p><b>A7 A relationship type is connected to at least one entity type by at least two edges.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>A8 Conceptual data modeling is aimed at dynamic abstraction.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>A9 Relationship types can be associated with relationship types.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>A10 For an entity type, at least on attribute with a corresponding data type is required.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>A11 Relationship types in the chosen variant of the ER model can be connected with three or more entity types.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>A12 A relationship type can relate entities of one entity type to entities of the same entity type.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				
<p><b>B – Domain knowledge</b></p>				
<p>IN ADDITION, WE ASK YOU FOR INFORMATION ABOUT YOUR KNOWLEDGE OF THE LIBRARY DOMAIN.</p>				
<p><b>B1 Have you ever visited a library?</b></p>				
<input type="checkbox"/> yes				
<input type="checkbox"/> no				
<p><b>B2 How often do you use the services of a library (e.g. a university library or city library)?</b></p>				
Not at all	Rarely	sometimes	Often	very often
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p><b>B3 In your own estimation, how many books have you borrowed from a library so far?</b></p>				
<input type="text"/> <input type="text"/> <input type="text"/> books				
<p><b>B4 I understand what the term "shelfmark" (sp. "la signatura") means in the context of a library.</b></p>				
<input type="checkbox"/> agree				
<input type="checkbox"/> disagree				

**Fig. 6.** Pre-modeling questionnaire – Part 2

**Questionnaire 2 – Own assessment and demographic information (answer after modeling)**

PLEASE ANSWER THE QUESTIONS IN THE ORDER GIVEN AND MARK THE CORRECT ANSWER WITH A CROSS OR FILL IN THE TEXT FIELD READABLE. IF YOU MADE A MISTAKE, PLEASE BLACKEN THE WHOLE BOX.

**C – Own assessment**

FIRST, WE ASK YOU FOR YOUR OWN ASSESSMENT OF WORKING ON THE MODELING TASK.

**C1 I had difficulties to verbalize my thoughts concerning the actual data modeling.**

☐ no

☐ yes, please describe your difficulties in verbalizing your thoughts in detail:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(Please fill in)

**C2 Did you encounter any modeling difficulties while working on the modeling task?**

☐ no

☐ yes, please describe your modeling difficulties in detail:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(Please fill in)

**Fig. 7.** Post-modeling questionnaire – Part 1



**Fig. 8.** Post-modeling questionnaire – Part 5

**Fig. 8.** Post-modeling questionnaire – Part 5

D – Demographic information
FINALLY, WE ASK FOR DEMOGRAPHIC INFORMATION.
<b>D1 What is your gender?</b>
<input type="checkbox"/> male
<input type="checkbox"/> female
<input type="checkbox"/> diverse
<b>D2 Please state your age.</b>
<input type="text"/> <input type="text"/> years
<b>D3 Which language is your first language (mother tongue)?</b>
..... (Please fill in)
<b>D4 What is your highest professional qualification (e.g., Bachelor of Arts, Master of Science, PhD etc.)?</b>
..... (Please fill in)
<b>D5 In which subject(s) did you obtain your highest professional qualification (e.g., Computer Science, Biology etc.)?</b>
..... (Please fill in)
<b>D6 What is your current position (e.g., Master student, PhD candidate, Postdoctoral researcher)?</b>
..... (Please fill in)
<b>D7 In which subjects are you currently studying or working (e.g., Computer Science, Biology etc.)?</b>
..... (Please fill in)
<b>D8 How long is your professional experience?</b>
<input type="text"/> <input type="text"/> years and <input type="text"/> <input type="text"/> month
<b>D9 In which application area or branch of industry did you gain your professional experience?</b>
..... (Please fill in)

**Fig. 9.** Post-modeling questionnaire – Part 3