# Structural Equation Modeling in HCI Research using SEMinR

André Calero Valdez
University of Lübeck
Lübeck, Germany

Lilian Kojan
University of Lübeck
Lübeck, Germany

Nicholas Danks
Trinity College Dublin
Dublin, Ireland

Soumya Ray
National Tsing Hua University
Hsinchu, Taiwan, R.O.C.

## ABSTRACT

Structural equation models (SEMs) are statistical techniques that help to identify models of latent variables in survey data. This allows researchers to test both the quality of the measurement instrument—the survey—as well as the hypothesized relationships using a single model. Partial least squares structural equation modeling (PLS-SEM) is a subset of SEM that works well with small sample sizes and non-parametric data, which frequently occur in HCI research. In this course, we will provide a short introduction to SEMinR, an open-source library for the R language. SEMinR is an easy-to-use domain-specific language for defining, estimating, visualizing, and validating SEMs using the PLS method. SEMinR provides means for scientific reporting and can be used by academics and practitioners alike.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; User studies; • **General and reference** → *Estimation*; *Measurement*.

## KEYWORDS

structural equation modeling, statistical methods, data analysis, psychometric methods, survey methods, causal analysis, SmartPLS, SEMinR

**ACM Reference Format:**
André Calero Valdez, Lilian Kojan, Nicholas Danks, and Soumya Ray. 2023. Structural Equation Modeling in HCI Research using SEMinR. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3544549.3574171

## 1 INTRODUCTION

Researchers of human-computer interaction (HCI) inevitably have to deal with a finicky and hard-to-grasp specimen: The human (or user) whose attributes are notoriously hard to measure. To examine user attitudes and perceptions, HCI researchers often utilize methods from psychological and behavioral science. That is, they use instruments like surveys to measure indicators of variables like attitudes. Afterward, they can use the survey results to construct and test models that explain or even predict user behavior.

One powerful method to build and test models is partial least squares structural equation modeling (PLS-SEM). This method allows the researcher to simultaneously test the validity of their measurement instrument (e.g., the survey) and the relationship between model variables, e.g., perceptions, attitudes, and behavior. A further advantage is that PLS-SEM is comparatively robust to small samples and non-parametric data, both issues that often occur in HCI research.

The course that this proposes is an introduction to PLS-SEM for HCI researchers and practitioners that are interested in a streamlined and robust way to work with human data. PLS-SEM will be taught using the open-source SEMinR R package. We will explain why PLS-SEM might be interesting for your research and how to construct and test a model in SEMinR. Lastly, we provide additional materials on reproducibility and community building.

## 2 WHY SEM-PLS

Human-computer interaction research studies the interactions between computer systems and the humans that use these systems. Computers can be viewed as technological artifacts [7]. Thus, the science of HCI can be seen as a design science, with the goal of researching the creation, improvement, and management of specific computer technologies and systems [4]. Often, research focuses on the design attributes of technologies and systems, such as performance, efficacy, and accuracy. Some of these attributes are not directly measurable and must therefore be operationalized through a series of observable variables. *Operationalization* describes the process of creating measurements for something that cannot be measured directly.

Variables that cannot be measured directly are especially common in HCI research conducted on humans. This kind of research is heavily informed by behavioral research and applied psychology. It often focuses on the traits, beliefs, and perceptions of the examined individuals. These are so-called *latent* variables, which means that they cannot be directly observed or measured. Rather, they must be inferred or measured indirectly using measurable or observable variables.

For example, self-efficacy is a concept used a lot throughout psychological and management sciences. It refers to an individual's belief in their own capabilities to perform a certain action and reach certain goals [1]. This belief is a latent variable or latent construct. It cannot be measured directly but must be inferred from (or operationalized through) survey items that measure the manifest symptoms of self-efficacy [1]. That is, having the trait self-efficacy


*CHI EA '23, April 23–28, 2023, Hamburg, Germany*

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3574171

```
measurements <- constructs(
  reflective("Image",      multi_items("IMAG", 1:5)),
  composite("Expectation", multi_items("CUEX", 1:3)),
  composite("Loyalty",     multi_items("CUSL", 1:3),
       weights = mode_B),
  composite("Complaints",  single_item("CUSCO")))

structure <- relationships(
  paths(from = c("Image", "Expectation"), to = c("
       Complaints", "Loyalty"))
```

**Listing 1: A code example creating a measurement model for four variables and then creating a structural model using these variables.**

(latent construct) is viewed as causing the respondent to answer the survey questions in a determinate way (resulting in indicators).

Empirical research in HCI thus requires rigorous and versatile methods that can incorporate not only observable variables but also latent variables, and design variables. PLS-SEM is uniquely placed to meet this requirement in that it can simultaneously estimate how these latent variables are constructed from their indicators, and how the variables relate to each other, therefore producing consistent and unbiased results [2]. The dual explanatory and predictive nature of PLS-SEM makes it relevant not only to scientific researchers but also to practitioners interested in predictive modeling.

## 3 WHY SEMINR

SEMinR is an R package designed to make structural equation models (SEM) approachable to practitioners across disciplines. It is an *open-source* and *free* tool that practitioners can use immediately with minimal setup. SEMinR's syntax offers a domain-specific language to define structural equation models using R syntax. The syntax is designed with the needs and knowledge of SEM practitioners foremost in mind. While we focus on partial least squares structural equation modeling (PLS-SEM) in this tutorial, SEMinR also includes covariance-based SEMs and covariance-based confirmatory factor analysis. SEMinR has functions named after the major features of an SEM model. For example, consider a PLS-SEM model that has several *composite* or *reflective constructs* that are measured by *multiple items* (e.g., survey questions), with causal *paths from* a certain set of constructs *to* others. The syntax for this model might look like Listing 1.

Comparing our description of the model to the syntax, we can see that the functions we use correspond to the nomenclature of SEM models that practitioners use: the measurement model is defined using functions called `constructs()`, `reflective()`, `composite()`, and `multi_items()`; and the structural model is defined by functions using `relationships()`, and `paths()`. This makes SEMinR's syntax highly readable and interpretable to new and expert SEM practitioners alike.

### 3.1 High-level and Opinionated

The syntax of SEMinR also resolves many data-munging tasks that practitioners would otherwise have to do by themselves. For example, models that include interaction terms would ordinarily require the user to manually create a new interaction variable based on item products. But SEMinR's *high-level syntax* allows users to specify the overall goal of creating an interaction term and SEMinR automatically conducts the above steps: `interaction_term(iv = "Image", moderator = "Expectation")`

The `interaction_term()` function, like many other functions in SEMinR, can take several parameters to specify the method of estimation, but its default parameters are based on the *published and demonstrated best practices* found in the literature.

### 3.2 Friendly Reporting Tools

SEMinR has an expanding set of reporting tools that produce highly readable tables and appealing visualizations that summarize the results of an estimated model. For example, SEMinR can generate a visualization (see Figure 1) of an estimated model as a directed acyclic graph, including the estimated path coefficients and bootstrapped confidence intervals.

### 3.3 Reproducibility and Reuse

While commercial solutions abound for SEM analysis, the programmatic syntax of SEMinR confers many advantages for scientific research. The evolution of a model over time can be *versioned* and tracked by version control tools like git. The syntax can be rerun on different machines to *reproduce* a solution (SEMinR is tested on a variety of platforms). Finally, researchers can *reuse* elements of syntax in other research efforts.

### 3.4 Resources and Community

SEMinR not only has a dedicated authoring team but also a fast-growing and engaged community of users who help develop resources for learning the package and discovering its ins and outs.

Most content provided in the workshop can also be found in the workbook by Hair Jr et al. [3] for future reference. A Facebook group for casual questions is available [1]. A large series of video tutorials are available on Youtube[2]:

Our workshop at CHI will take users through the underpinnings of SEM, the basics of our syntax, and how to best use the resources available.

## 4 INTENDED AUDIENCE(S)

The intended CHI audiences of the course are: Academic researchers including Masters and doctoral students as well as industry researchers that are interested in thoroughly understanding their users including the users' motivations and behavior. The key takeaways for these audiences are: Academic researchers will learn 1) how to transform a research question into a structural equation model, including relationships between variables and the relationships between variables and measurement, 2) how to estimate a model using empirical data, 3) how to validate findings using bootstrapping, 4) how to report a model for a scientific publication, and 5) how to evaluate a reported model as a reviewer. Industry researchers will learn how to utilize SEM to improve products effectively.

[1] Facebook group: https://www.facebook.com/groups/seminr
[2] Youtube Playlist: https://www.youtube.com/playlist?list=PLb7vm6tsQ3Ks0TyMWw3EUlgoMr06_7U8S

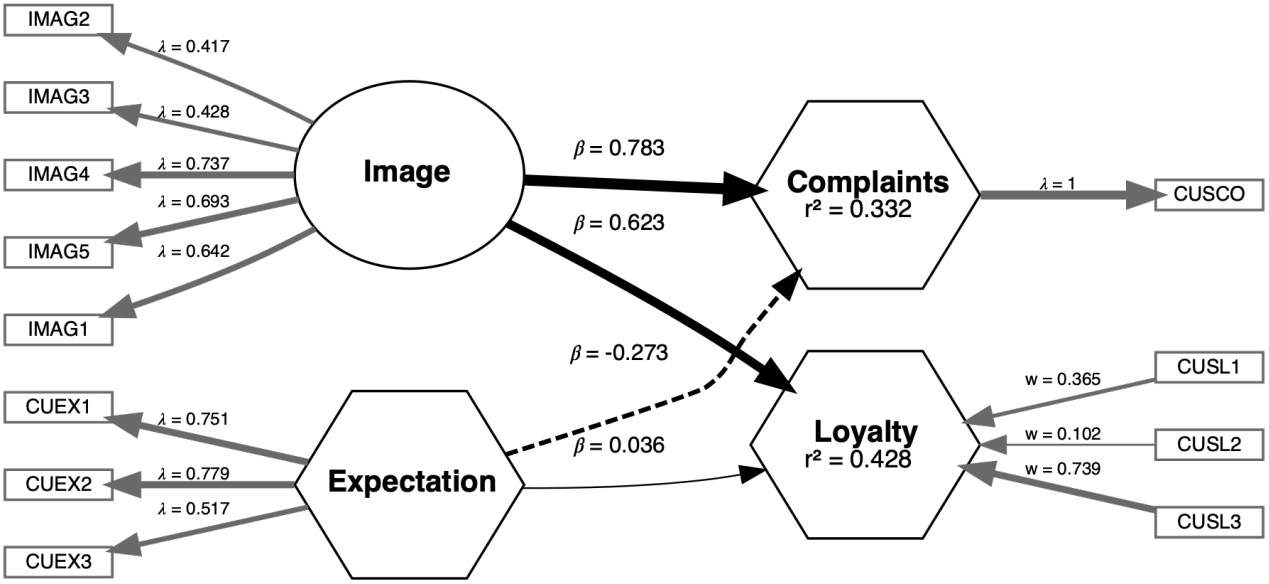


**Figure 1: Estimated SEM model visualized by SEMinR generated by the code in Listing 1**

## 5 CONTENT

The course will consist of three modules of 75 minutes each, with each module including theoretical input and practical exercises.

The first module will provide a basic introduction to R with RStudio (35 minutes), structural equation modeling in general, and using SEMinR (35 minutes).

The second module will provide a deeper understanding of how to construct a model (35 minutes) and introduce estimation techniques using a pre-selected data set (35 minutes).

The last module addresses evaluating and reporting models. It will teach how to draw conclusions from results and inform on how to generate sensible models. This module is a half lecture (35 minutes) and half exercise (20 minutes + 15 minutes presentation). We will finish by providing follow-up material on reproducibility on YouTube.

## 6 ENROLLMENT, PREREQUISITES AND MATERIALS

The course will be advertised through SIGCHI and other relevant mailing lists and social media channels. Participants should have at least some basic programming experience. Basic statistical knowledge (e.g., linear regressions, t-tests, confidence intervals, levels of significance) is an advantage. Knowledge of R is not required.

All exercises will be conducted on RStudio.cloud (now called Posit cloud) instances provided by the instructors, prepared with all necessary files, and free to use for one month. Having a laptop with a chromium-based browser is necessary to access RStudio.cloud.

## 7 ABOUT THE INSTRUCTORS

The instructors include key authors, maintainers, and contributors of the SEMinR package [6]. André Calero Valdez is Professor for human-computer interaction and usable safety engineering in the University of Lübeck. He uses structural equation modeling to identify antecedents of user behavior in safety-critical applications and eHealth applications.

Lilian Kojan is a Ph.D. candidate at the University of Lübeck and extensively uses structural equation modeling to understand the impact of trust and communication on human behavior in topics such as the Covid-19 pandemic [5] or climate change.

Nicholas Danks is a Professor of Business Analytics at Trinity Business School at Trinity College Dublin and one of the maintainers of the SEMinR package.

Soumya Ray is a Distinguished Professor of Service Science at the National Tsing Hua University. He both applies structural equation modeling in empirical research and develops new methodologies around structural equation modeling.

## REFERENCES

[1] Albert Bandura et al. 2006. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents* 5, 1 (2006), 307–337.
[2] Joe F Hair, Christian M Ringle, and Marko Sarstedt. 2011. PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice* 19, 2 (2011), 139–152.
[3] Joseph F Hair Jr, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, Nicholas P Danks, and Soumya Ray. 2021. Partial least squares structural equation modeling (PLS-SEM) using R: A workbook.
[4] Jörg Henseler. 2017. Bridging design and behavioral research with variance-based structural equation modeling. *Journal of advertising* 46, 1 (2017), 178–192.
[5] Lilian Kojan, Laura Burbach, Martina Ziefle, and André Calero Valdez. 2022. Perceptions of behaviour efficacy, not perceptions of threat, are drivers of COVID-19 protective behaviour in Germany. *Nature Humanities and Social Sciences Communications* 9, 1 (2022), 1–15.
[6] Soumya Ray, Nicholas Danks, and André Calero Valdez. 2021. SEMinR: Domain-specific language for building, estimating, and visualizing structural equation models in R. *Estimating, and Visualizing Structural Equation Models in R (August 6, 2021)* (2021).
[7] Herbert A Simon. 1969. The sciences of the artificial MIT Press. *Cambridge, MA* (1969).