Simon Schäfer[1]
Dr. Rüdiger Tiemann[1]
[1]Humboldt-Universität zu Berlin

# Bayes'sche Item-Response-Modellierung mit R

## Motivation

Ausgangslage für die Beschäftigung mit Bayes'scher Statistik stellte die Entwicklung eines Testinstruments zur Messung fachdidaktisch-deklarativen Wissens bei Chemie-Lehramtsstudierenden dar. Die Aufgaben wurden im Multiple-Choice-Format konzipiert, sodass meist mehr als ein Attraktor und mehrere Distraktoren gegeben sind. Die eingesetzten Messmodelle basieren auf der Item Response Theorie (IRT), da diese im Vergleich zur klassischen Testtheorie einige Vorteile bezüglich der Messung von Kompetenzen und der Untersuchung von Kompetenzmodellen (vgl. Hartig & Frey, 2013) bietet. Beispielsweise lassen sich konkurrierende Item-Response-Modelle (IR-Modelle) auf Grund ihres prädiktiven Charakters anhand von Informationskriterien miteinander vergleichen.

Dabei soll die Auswertung der Daten mit möglichst wenig Informationsverlust realisiert werden. Daher wurde sich sowohl gegen eine dichotome Auswertung der Aufgaben als Ganzes als auch gegen ein Partial-Credit-Modell entschieden. Stattdessen werden die einzelnen Antwortmöglichkeiten jeweils als eigenes dichotomes Item behandelt, deren lokale Abhängigkeit (bedingt durch den gemeinsamen Itemstamm) mit Hilfe des Testlet-Modells (Wainer, Bradlow & Wang, 2007) berücksichtigt wird.

Als mehrdimensionales (und ggf. auch mehrparametrisches) Modell ist das Testlet-Modell als recht komplex zu bezeichnen. Die benötigte Stichprobengröße von IR-Modellen steigt aber sowohl mit der Komplexität des Modells als auch mit der avisierten Bestimmtheit der zu treffenden Aussagen. Es finden sich diverse Studien mit Aussagen über die Mindeststichprobengröße, deren Ergebnislage aber recht heterogen ist (vgl. Sahin & Anil, 2017). Jedoch werden für mehrdimensionale, mehrparametrische Modelle zumeist mehrere tausend Personen als Voraussetzung für eine belastbare Auswertung genannt.

Da eine solch große Stichprobe nicht akquiriert werden konnte, wurde nach einer Alternative gesucht. Fujimoto und Neugebauer (2020) zeigten, dass für Testlet-Modelle Parameterschätzungen mit geringem Bias schon für Stichproben ab 100 Personen möglich sind, wenn die Modellierung mit Methoden der Bayes'schen Statistik unter Verwendung informativer Prior abläuft.

## Zugang zur Bayes'schen Statistik

Vor der Entwicklung leistungsfähiger Computer und der Einführung der Markow-Chain-Monte-Carlo-Algorithmen [1990] (kurz MCMC) konnten Bayes'sche Methoden nur auf relativ einfache Modelle angewendet werden (Kruschke & Liddell, 2018a). Seitdem wurden mit WinBUGS [1997], JAGS [2007] und Stan [2012] drei bedeutsame Software-Pakete für Bayes'sche Analysen veröffentlicht. Dennoch ist die Nutzung Bayes'scher Methoden trotz seiner theoretischen Vorteile (Kruschke & Liddell, 2018b) noch nicht zur Norm geworden.

Ein Grund hierfür ist sicherlich das Fehlen benutzerfreundlicher Software, was in vergleichbarer Weise von Rost (1999) als einer der Gründe für das „Schattendasein" des Rasch-Modells in Bezug auf die klassische Testtheorie angeführt wurde. Das Erlernen einer eigenständigen Programmiersprache für das Durchführen Bayes'scher Analysen ist eine hohe Hürde. Hier bietet JASP (JASP Team, 2021) mittlerweile eine Oberfläche, die von der

Bedienung vergleichbar mit SPSS ist und sowohl Bayes'sche als auch frequentistische Analysen ermöglicht. JASP greift dabei auf die Funktionen aus R-Paketen zurück. Die Berechnung von Item-Response-Modellen ist mit JASP aber nicht möglich.
Ein weiterer Grund ist, dass statistische Lehrveranstaltungen noch immer vorrangig frequentistische Methodik lehren. Schon Chernoff schrieb in einem Kommentar zu Efrons (1986) Artikel mit dem Titel „Why Isn't Everyone a Bayesian" von der Verzögerung, mit der eine Veränderung in den Lehrplänen zu erwarten sei. Zu empfehlen sind für das Selbststudium die Bücher von McElreath (2020) und Kruschke (2015) und deren Adaptionen für das R-Paket brms (Bürkner, 2017) durch Kurz (2021). Die Lehrbücher beinhalten zwar keine Kapitel zum Thema Item Response Theorie bereiten aber auf die Anwendung generalisierter linearer Modelle (GLM) vor.

**Bayes'sche Item-Response-Modellierung mit R**

Bürkners Artikel zur Spezifizierung von IR-Modellen als GLM im R-Paket brms bieten einen Leitfaden für die Auswertung sowohl dichotomer als auch ordinaler IR-Modelle mittels Bayes'scher Methoden (Bürkner & Vuorre, 2019; Bürkner, 2019, 2020). Mit brms lassen sich aber auch nichtlineare Modelle (z.B. mehrparametrische IR-Modelle) formulieren. Der Fokus der referenzierten Artikel liegt dabei auf der Auswertung der Genauigkeit der geschätzten Parameter, der Modellierung von Kovariaten und dem Modellvergleich mittels leave-one-out cross-validation (LOO CV). Des Weiteren demonstriert er die Methode der Posterior Predictive Model Checks (PPMC), welche auf alle klassischen Itemfitindizes (z.B. Log-Likelihood, Infit, Outfit, G2, etc.) anwendbar ist und sowohl für die Selektion unpassender Items als auch Modellvergleiche genutzt werden kann.
Mit Vorerfahrungen in der Verwendung von R-Paketen wie TAM (Robitzsch, Kiefer & Wu, 2021) oder mirt (Chalmers, 2012) und der Methodik der Varianzanalyse wird man klassische Auswertungsmethoden für IR-Modelle wie Item Characteristic Curves oder Wrightmaps und auch das in der Bayes'schen Methodik weniger übliche $R^2$-Maß vermissen. Das R-Paket birtms (Schäfer, 2021) baut auf die Funktionalität von brms auf und stellt diese und weitere Methoden zur Auswertung Bayes'scher IR-Modelle zur Verfügung. Derzeit sind diese zusätzlichen Auswertungsmethoden aber noch nicht für alle spezifizierbaren Modelle implementiert.
Eine der wichtigsten Funktionen von birtms ist die Möglichkeit, die marginale Log-Likelihood zu berechnen. Als Basis diverser Informationskriterien wird die Log-Likelihood für den Vergleich der Modellpassung und die Wahl des am ehesten unterstützten psychometrischen Modells genutzt. Von diesem Vergleich ausgehend werden dann häufig Schlüsse über das zugrundeliegende theoretische Konstrukt abgeleitet.
Merkle, Furr und Rabe-Hesketh (2019) zeigten, dass die Informationskriterien für IR-Modelle auf Basis der marginalen Log-Likelihood berechnet werden müssen, da ansonsten komplexere Modelle im Modellvergleich bevorzugt werden. Dies konnte an den eigenen Datensätzen reproduziert werden. Da das bei Bürkner (2019) vorgestellte Vorgehen für den Modell-vergleich auf einer Funktion beruht, die die konditionale Log-Likelihood verwendet, ist eine Alternative nötig. Daher ist in birtms eine Funktion zur Berechnung der marginalen Log-Likelihood implementiert. Mit dieser kann die Methode PSIS-LOO CV (Vehtari, Gelman & Gabry, 2017) für den Modellvergleich genutzt werden. Auch hier ist die Methode noch nicht für alle spezifizierbaren Modelle verfügbar. Alternativ kann immer die grouped K-fold CV (Vehtari, 2020) genutzt werden, welche aber deutlich längere Laufzeiten hat.

**Entscheidungshilfe: Bayes'sche Item-Response-Modelle**

Oben wurde dargelegt, was den Autor zur Auseinandersetzung mit der Bayes'schen Statistik bewegte. Es wurde beschrieben, dass sich die Voraussetzungen für den Einsatz Bayes'scher Methoden allgemein verbessert haben und Möglichkeiten zum Selbststudium genannt. Nachfolgend werden einige Vorteile der Bayes'schen und frequentistischen Methodik aufgelistet, von denen die meisten bei Kruschke und Liddell (2018b) ausgeführt werden.

*Tab. 1: Gegenüberstellung Bayes'sche und frequentistische Methoden*

| **Vorteile Bayes'scher Methodik** | **Vorteile frequentistischer Methodik** |
|---|---|
| Flexible Modellspezifikation | Bekanntheit und Akzeptanz |
| Valide für kleinere Stichproben | Softwareverfügbarkeit* |
| Vorwissen nutzbar* | Hardware- und Zeitanforderungen* |
| Parameterverteilungen informieren über Genauigkeit der Ergebnisse | Bekannte Fehlerraten |
| Intuitiv interpretierbar | |
| Unabhängigkeit von geplanter Testanzahl | |
| Überwindung von Meehls Paradoxon (Meehl, 2013) | |
| Unsicherheit in Folgeauswertungen berücksichtigen | |

Dass auf Seite der Bayes'schen Methodik mehr Vorteile aufgelistet sind, soll nicht darüber hinwegtäuschen, dass die Nachteile im Bereich der IR-Modellierung aus Anwendungssicht äußerst relevant sind. Im Vergleich zu frequentistischen Methoden dauert die Modellierung um den Faktor $10^4$ mal länger, bedarf das Speichern der Modelle 100 mal mehr Platz und für manche der selbst implementierten Auswertungsmethoden reichen selbst 32 GB RAM nicht aus. Die Berechnung von 3PL-Testlet-Modellen auf Basis von etwa 28000 Beobachtungen (resultierend aus knapp 280 Items und gut 100 Personen) benötigte um die 24 Stunden. An anderer Stelle ist es erforderlich Funktionen selbst anzupassen oder neu zu implementieren, um die für speziellere Modelle nutzen zu können. Und je nach Denkschule wird die Möglichkeit, Vorwissen in Form von Priors zu nutzen nicht als Vorteil gesehen, sondern als subjektiv und manipulationsanfällig abgelehnt.

Bayes'sche Methoden basieren zwar auf einem konsistenteren und intuitiveren Wahrscheinlichkeitskonstrukt, liefern aber ohne die Verwendung informativer Prior vergleichbare Ergebnisse für die Modellparameter wie die etablierten frequentistischen Methoden. Daher bietet der bloße Wechsel zu MCMC-Algorithmen für den Praktiker kaum Vorteile. Damit findet sich eine weitere Parallele zur Situation, wie Rost (1999) sie bezüglich klassischer Testtheorie und Item Response Theorie beschrieb.

Der praktische Nutzen der Bayes'schen Metodik ergibt sich erst, wenn …

- mit kleineren Stichproben komplexe psychometrische Modelle getestet werden,
- Modelle spezifiziert werden, für die es noch keine klassischen Lösungen gibt,
- die Auswertung auf Grundlage der Parameterverteilungen und PPMC-Methoden basiert,
- oder Vorwissen mittels informativer Prior in die Auswertung mit einbezogen wird.

Interessierten Pionieren bietet die Bayes'sche Methodik aber schon heute die Möglichkeit differenziertere Fragestellungen mit dafür maßgeschneiderten Modellen zu überprüfen.

# Literatur


Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2019, 23. Mai). *Bayesian Item Response Modeling in R with brms and Stan*. Verfügbar unter: https://arxiv.org/pdf/1905.09501

Bürkner, P.-C. (2020). Analysing Standard Progressive Matrices (SPM-LS) with Bayesian Item Response Models. *Journal of Intelligence*, *8*(1). https://doi.org/10.3390/jintelligence8010005

Bürkner, P.-C. & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Chalmers, R. P. (2012). mirt : A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Efron, B. (1986). Why Isn't Everyone a Bayesian? *The American Statistician*, *40*(1), 1. https://doi.org/10.2307/2683105

Fujimoto, K. A. & Neugebauer, S. R. (2020). A General Bayesian Multidimensional Item Response Theory Model for Small and Large Samples. *Educational and Psychological Measurement*, *80*(4), 665–694. https://doi.org/10.1177/0013164419891205

Hartig, J. & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl" für die Modellierung von Kompetenzen? *Zeitschrift für Erziehungswissenschaft*, *16*(S1), 47–51. https://doi.org/10.1007/s11618-013-0386-0

JASP Team. (2021). JASP (Version 0.15) [Computer software].

Kruschke, J. K. (2015). *Doing Bayesian data analysis. A tutorial with R, JAGS, and Stan* (2. ed.). Amsterdam: AP Academic Press/Elsevier. Retrieved from http://www.contentreserve.com/TitleInfo.asp?ID={38F45CF6-6B5C-433C-85F8-A3568420927D}&Format=50

Kruschke, J. K. & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K. & Liddell, T. M. (2018b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kurz, A. S. (2021, 23. Oktober). *Books | A. Solomon Kurz*. Zugriff am 23.10.2021. Verfügbar unter: https://solomonkurz.netlify.app/bookdown/

McElreath, R. (2020). *Statistical Rethinking*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Meehl, P. E. (2013). The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What If There Were No Significance Tests?* (Multivariate Applications Series, S. 393–425). Hoboken: Taylor and Francis. Zugriff am 12.09.2021. Verfügbar unter: https://meehl.umn.edu/sites/meehl.umn.edu/files/files/169problemisepistemology.pdf

Merkle, E. C., Furr, D. & Rabe-Hesketh, S. (2019). Bayesian Comparison of Latent Variable Models: Conditional Versus Marginal Likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Robitzsch, A., Kiefer, T. & Wu, M. (2021). TAM: Test Analysis Modules (Version 3.7-16) [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=TAM

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, *50*(3), 140–156. https://doi.org/10.1026//0033-3042.50.3.140

Sahin, A. & Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*. https://doi.org/10.12738/estp.2017.1.0270

Schäfer, S. (2021). *Famondir/birtms: Bayesian Item Response Theory Models using Stan*. Zenodo. https://doi.org/10.5281/zenodo.5507637

Vehtari, A. (2020, 16. Dezember). *Cross-validation for hierarchical models*. Zugriff am 29.03.2021. Verfügbar unter: https://avehtari.github.io/modelselection/rats_kcv.html

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618765