WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Automatisierte Analyse von Multiple-Choice Prüfungen

Karl Ledermüller, Michaela Nettekoven, Maria Weiler

Internationale Tagung des Netzwerks für Qualitätsmanagement und Qualitätsentwicklung, 10. Oktober 2014

Inhalt

- Einleitung
- Software
- Statistische Auswertungen
  - Noten und Prozentwerte
  - Klassifizierung von Prüfungsfragen
  - Detailanalyse der Fragen und Antworten
  - Auswertungen gemäß der Item Response Theory
- Zusammenfassung

Inhalt

Einleitung

Evaluierung der Lehrqualität an Universitäten:

- häufig eingesetzt: Lehrveranstaltungsevaluierung
- oft vernachlässigt: Evaluierung der Prüfungen

Die Art der Wissensüberprüfung hat einen entscheidenden Einfluss auf das Lernverhalten, ein unterstützendes Feedbacksystem für Lehrende zur Erstellung von Prüfungen — insbesondere bei Multiple-Choice Prüfungen — wäre daher wünschenswert.

Multiple-Choice Prüfungen

Vorteile:

- maschinelle und zeitsparende Auswertung und Benotung
- vergleichsweise einfache statistische Analysemöglichkeit

Die Erstellung von „guten“ MC-Prüfungen ist aufwändig, insbesondere sollten folgende grundlegenden Punkte beachtet werden:

- Festlegung des oder der zu messenden Konstrukte („Kompetenzen“)
- Abstimmung der MC-Fragen untereinander bzw. auf das zu messende Konstrukt
- Abstimmung der Antwortoptionen untereinander

Praktische Probleme

Schwierigkeiten bei der Prüfungserstellung:

- mangelnder Überblick über frühere Prüfungen bzw. Prüfungsfragen
- fehlende Informationen über einfache und schwierige Fragen bzw. Antwortoptionen (ex ante)
- fehlende Informationen über den Zusammenhang zwischen Schwierigkeit (Fragen) und Fähigkeit (Studierende)
- fehlende Sensibilität über die Abhängigkeit zwischen Antwortoptionen (richtig/falsch), „Ankreuzverhalten“ und Punktevergabe

Dadurch ist es schwierig, basierend auf Erfahrungswerten „bessere“ Prüfungen zu erstellen.

Zielsetzung

Eine umfassende Analyse von MC-Prüfungen soll sowohl die Zusammenstellung zukünftiger Prüfungen erleichtern als auch deren Qualität verbessern, insbesondere durch

- verbesserten Einblick in die eigenen Prüfungen
- Identifikation einfacher und schwieriger Fragen
- bessere Abstimmung der Fragenzusammenstellung
- bessere Einschätzung der Fähigkeiten der Studierenden
- Zuordnung von Prüfungsfragen zu Konstrukten (z.B. Kompetenzen)
- Sichtbarmachen von Problembereichen und Verbesserungsmöglichkeiten
- Lernen aus vergangenen Prüfungsterminen

Analyse-Tool für MC-Prüfungen

Um obige Ziele zu erreichen, wurde ein Tool erstellt, das die Prüfungsergebnisse vom Prüfungsserver automatisiert auswertet und den Prüfungsverantwortlichen einen Bericht mit statistischen Kennzahlen, Interpretationshinweisen sowie Erläuterungen zur Verfügung stellt.

Der Bericht enthält

- deskriptive Statistiken zur Noten- und Punkteverteilung sowie zu den einzelnen Fragen und Antworten
- Sensitivitätsanalysen zu den einzelnen Fragen
- Faktoren- und Clusteranalyse zur Gruppierung der Prüfungsfragen
- Auswertungen gemäß der Item Response Theory

Wichtig: Die statistischen Kennzahlen sind an sich nicht wertend, sondern bedürfen immer der kontextabhängigen Interpretation!

Inhalt

Software

Das Tool basiert gänzlich auf Open Source Software

- Softwarepaket R (inkl. diverse Zusatzpackages wie beispielsweise knitR)
- Textsatzsystem LaTeX
- Prüfungsserver basiert auf OpenACS sowie .lrn

Inhalt

- Einleitung
- Software
- Statistische Auswertungen
  - Noten und Prozentwerte
  - Klassifizierung von Prüfungsfragen
  - Detailanalyse der Fragen und Antworten
  - Auswertungen gemäß der Item Response Theory
- Zusammenfassung

Noten und Prozentwerte

Grundlegende Fragen:

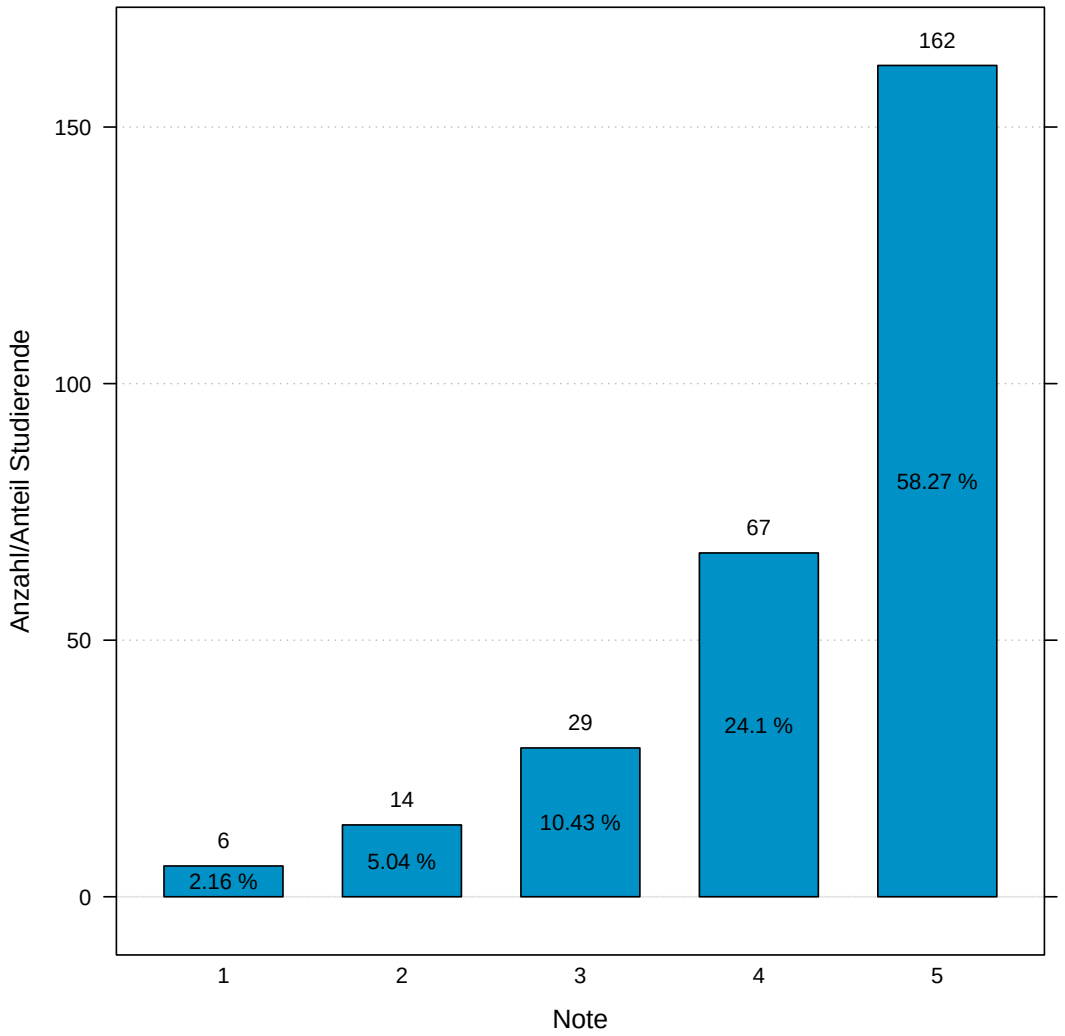
- Wie verteilen sich die Noten, wie die erreichten Prozentwerte auf Fragen- sowie auf Antwortebene?
- Unterscheiden sich die Fragenergebnisse, wenn man diese nach Noten gruppiert? Welche Fragen sind besonders gut geeignet, Studierende mit guten und schlechten Noten zu „trennen“?
- Wie würden sich die Noten und erreichten Prozentwerte ändern, wenn eine Frage aus der Prüfung gestrichen würde?

Deskriptive Statistiken:

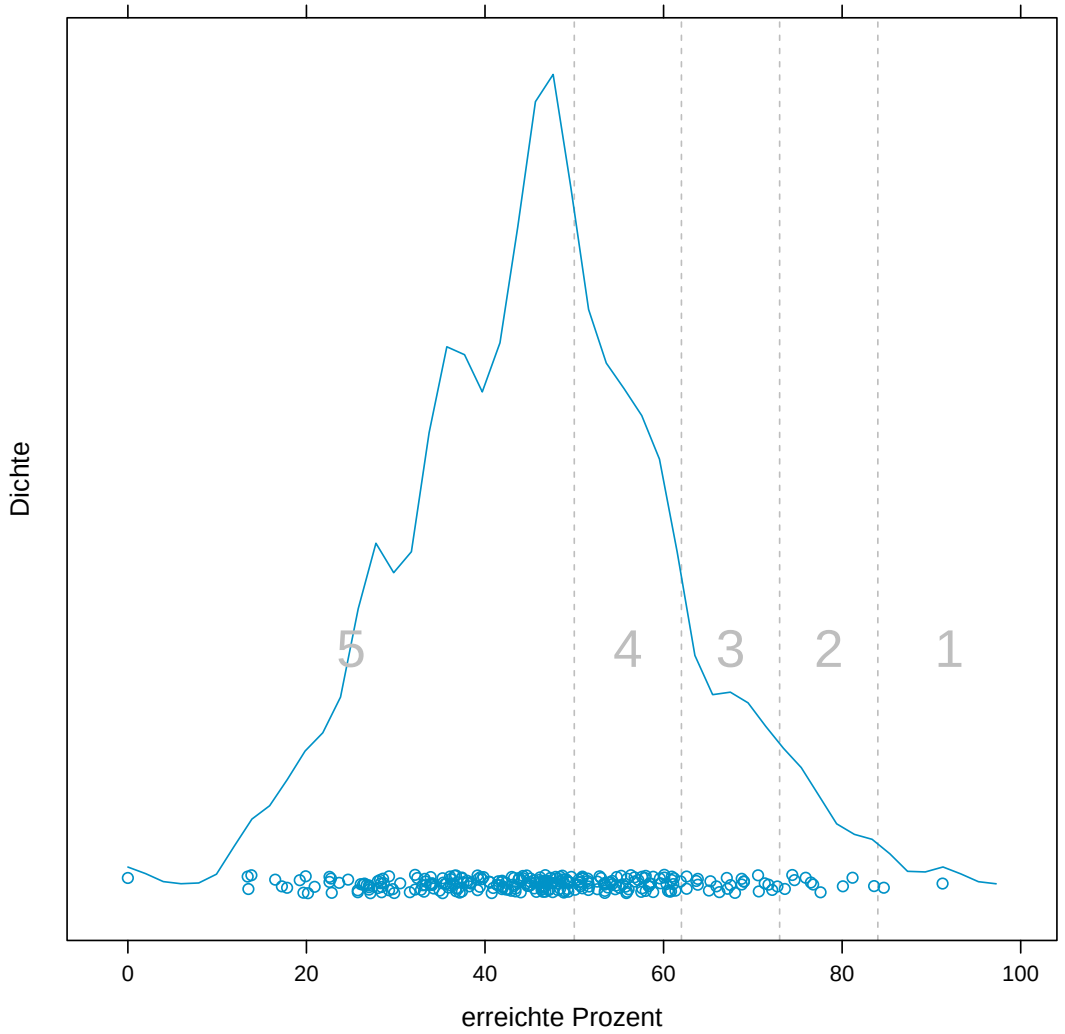
- Notenverteilung
- Dichtefunktion der erreichten Gesamtprozent
- für jede Prüfungsfrage: durchschnittlich erreichte Prozent
- für jede Antwortoption: Anteil der Studierenden, die diese angekreuzt haben
- durchschnittlich erreichte Prozent je Frage, gruppiert nach Noten

(Daten aus: Prüfung Finanzierung vom 28.06.2013)

Notenverteilung



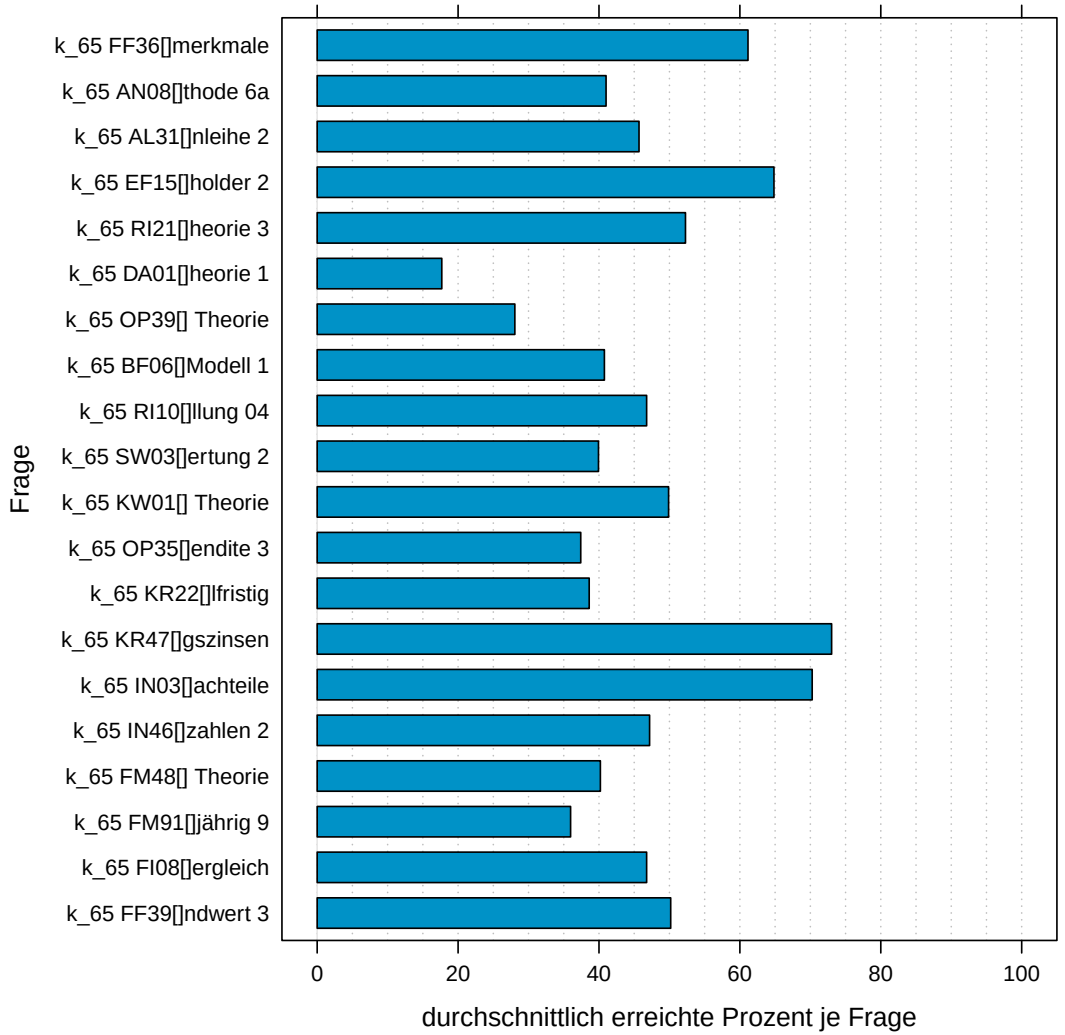
Dichtefunktion der erreichten Gesamtprozent



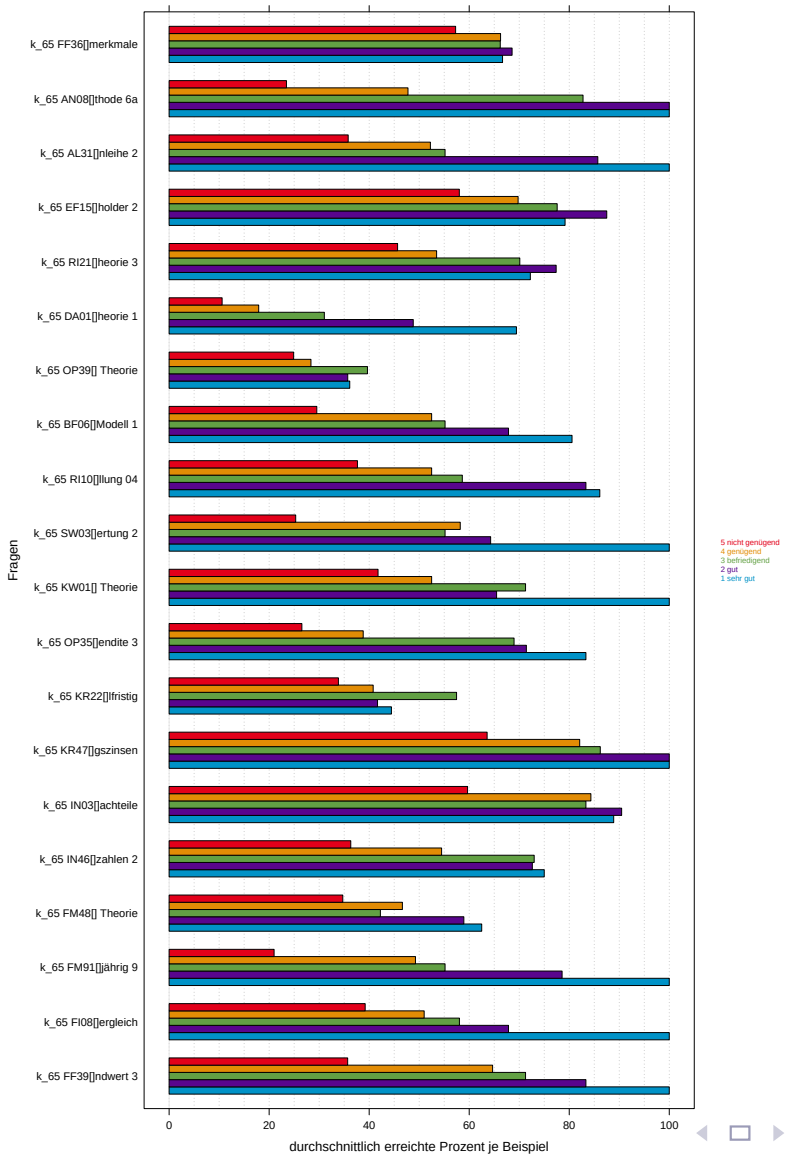
Statistiken pro Prüfungsfrage und Antwortoption

Prozentwerte	RF	Aufgabe 6 k_65 DA01: Dynamische Amortisationsrechnung Theorie 1 Punkte: 5 durchschnittlich erreichte Prozent: 17.69
		Bei der dynamischen Amortisationsrechnung
25.90	0	werden im Unterschied zur statischen Amortisationsrechnung alle mit dem Investitionsprojekt verbundenen Zahlungen berücksichtigt
41.01	1	werden im Gegensatz zur statischen Amortisationsrechnung die zu berücksichtigenden Zahlungen auf den Zeitpunkt t=0 abgezinst
69.78	0	wird der Zeitpunkt ermittelt, zu dem die Summe der bisherigen Einzahlungen erstmals größer als die Summe der bisherigen Auszahlungen ist
86.69	1	wird beim Projektvergleich das Projekt vorgezogen, welches eine kürzere Amortisationszeit aufweist
22.66	1	kann bei Normalinvestitionen der kumulierte Barwert zum Amortisationszeitpunkt nicht größer als der Kapitalwert des Projekts sein

Durchschnittlich erreichte Prozent pro Frage



Durchschnittlich erreichte Prozent pro Frage, gruppiert nach Note



Sensitivitätsanalyse

Was wäre, wenn eine Frage aus der Prüfung gestrichen würde?

	43 Sehr gut	41 Gut	41 Befriedigend	54 Genügend	99 Nicht genügend
k_65 FF36[]merkmale	1	1	-1	-1	0
k_65 AN08[]thode 6a	6	0	5	-5	-6
k_65 AL31[]nleihe 2	9	-3	0	-8	2
k_65 EF15[]holder 2	1	0	1	-1	-1
k_65 RI21[]heorie 3	5	0	-1	-3	-1
k_65 DA01[]heorie 1	10	-3	8	-8	-7
k_65 OP39[] Theorie	10	-3	4	-7	-4
k_65 BF06[]Modell 1	4	-2	4	-2	-4
k_65 RI10[]llung 04	5	-1	2	-3	-3
k_65 SW03[]ertung 2	9	-2	2	-4	-5
k_65 KW01[] Theorie	2	4	1	-5	-2
k_65 OP35[]endite 3	14	-13	6	-5	-2
k_65 KR22[]lfristig	5	-1	4	-4	-4
k_65 KR47[]gszinsen	2	-5	-1	5	-1
k_65 IN03[]achteile	-1	1	1	-1	0
k_65 IN46[]zahlen 2	5	-2	5	-6	-2
k_65 FM48[] Theorie	3	-1	3	-4	-1
k_65 FM91[]jährig 9	13	-2	-1	-6	-4
k_65 FI08[]ergleich	1	1	0	-2	0
k_65 FF39[]ndwert 3	5	0	-3	0	-2

Klassifizierung der Prüfungsfragen

Grundlegende Fragen:

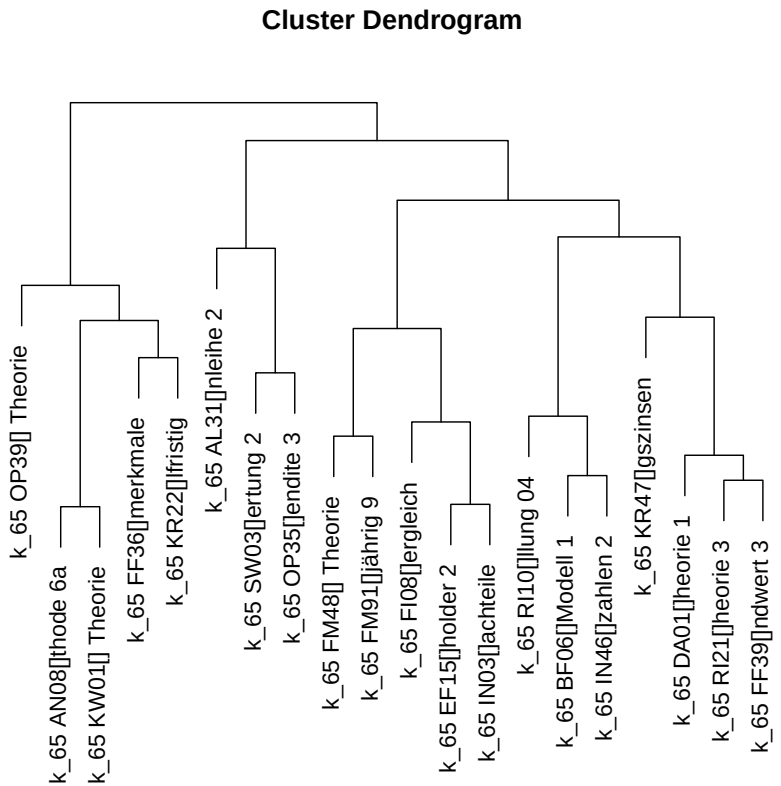
- Messen alle Fragen dasselbe Konstrukt (z.B. eine Kompetenz)?
- Wenn nicht, welche unterschiedlichen Dimensionen lassen sich identifizieren?
- Welche Fragen zielen auf die gleichen Dimensionen ab?

Faktoren- und Clusteranalyse zeigen ähnliche Prüfungsitems auf, bzw. ordnen die Items verschiedenen Gruppen zu.

Wofür ein Faktor bzw. Cluster steht, muss vom Prüfungsverantwortlichen selber interpretiert werden, zum Beispiel:

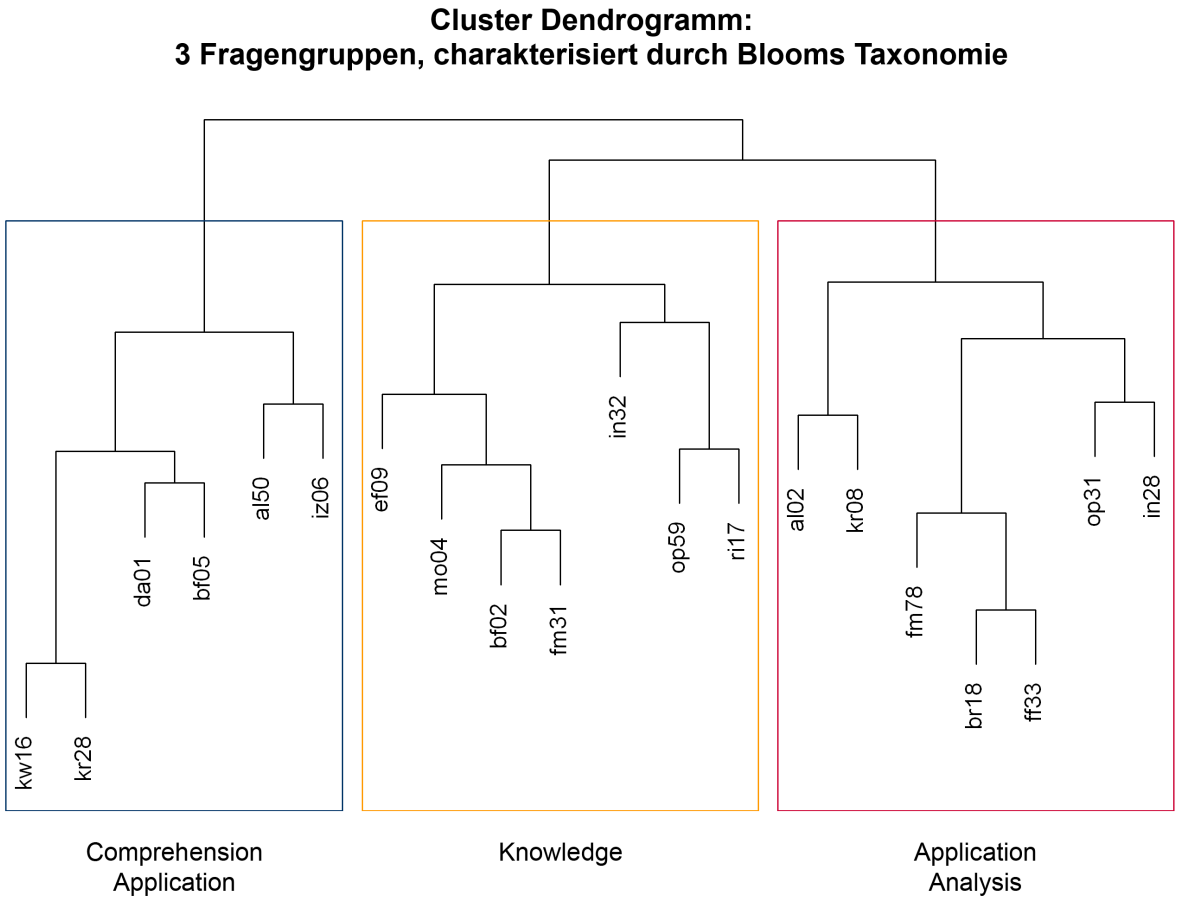
- Kompetenzen
- Zielniveau der Prüfungsfragen (Bloom'sche Taxonomie)
- „Verständnisfragen“ und „Auswendiglern-Fragen“
- Theoriefragen und Rechenbeispiele
- verschiedene Kapitel des Prüfungsstoffes
- beliebige Kombinationen

Clusteranalyse (Prüfung Juni 2013)



keine eindeutige Interpretation der Gruppen gefunden

Clusteranalyse (Prüfung Mai 2011)



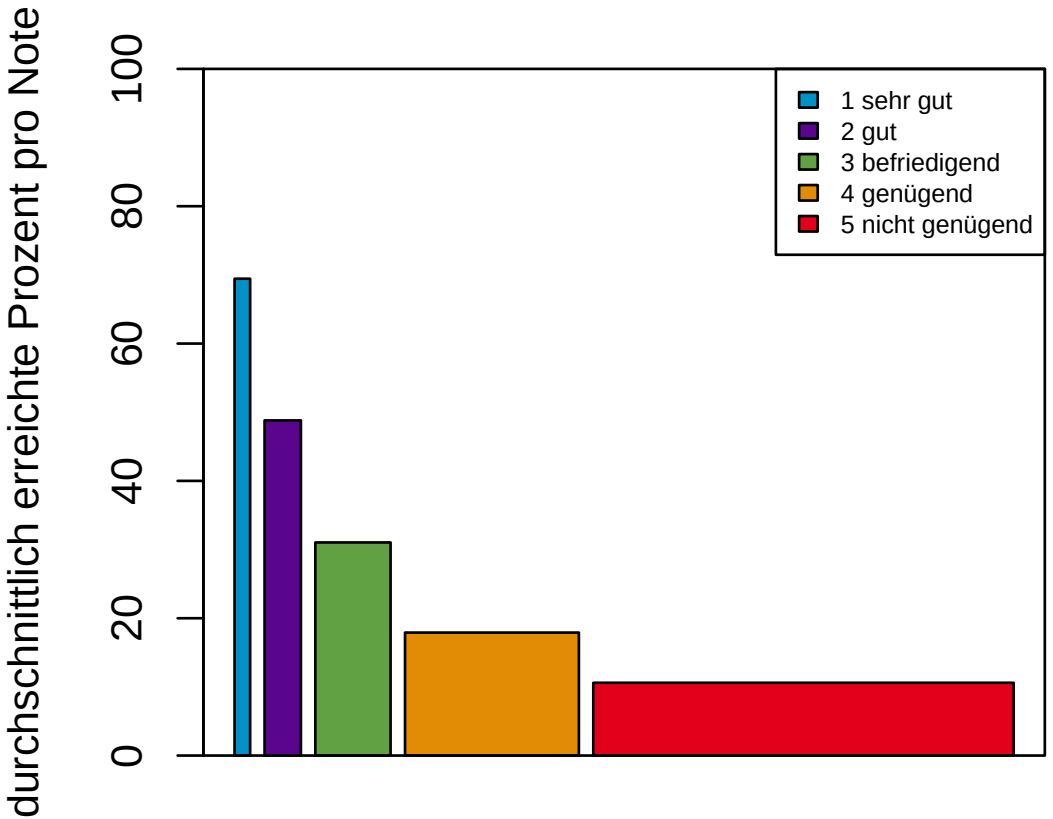
Interpretation der Gruppen nach dem Zielniveau der Prüfungsfragen

Detailanalyse der Fragen und Antworten

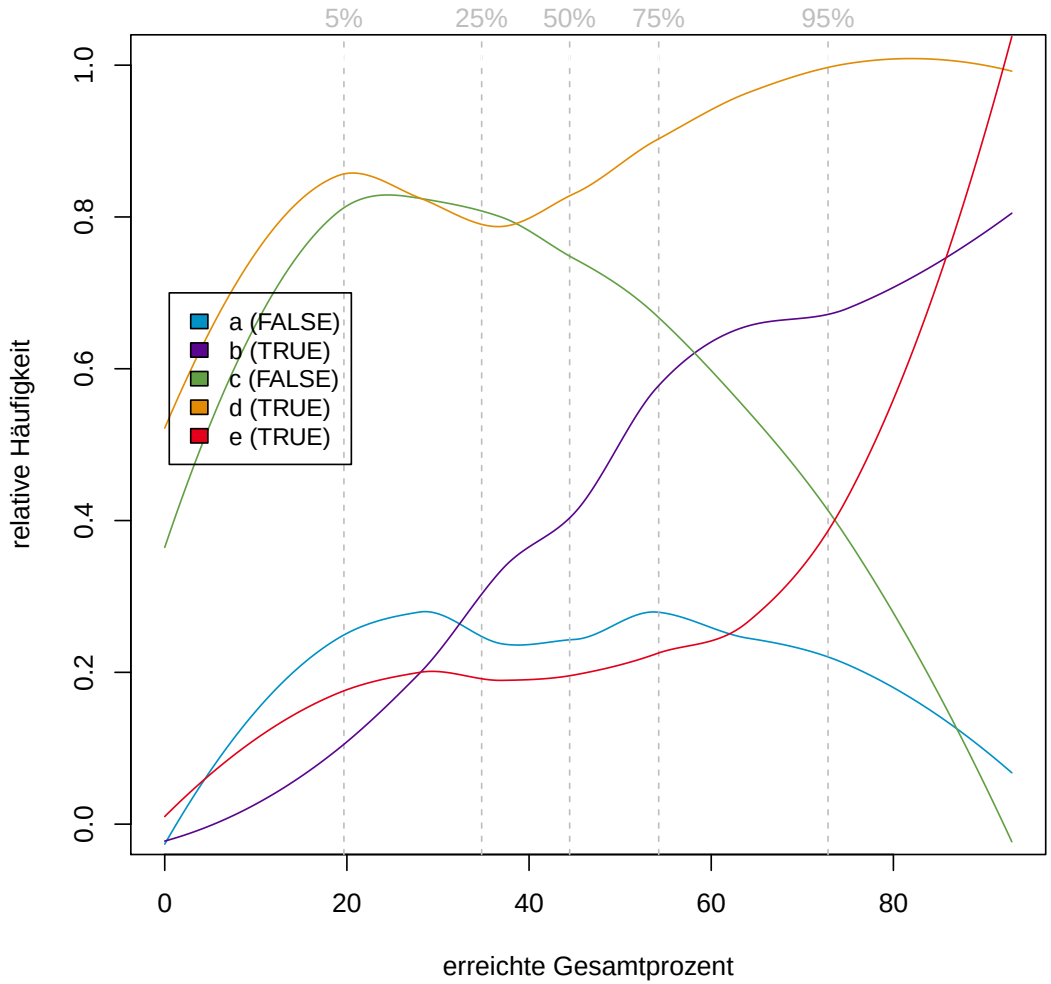
Jeder Report enthält Detailauswertungen zu jeder Frage zu folgenden Themenbereichen:

- Tabelle mit Statistiken pro Antwortoption auf Fragenebene
- durchschnittlich erreichte Prozent pro Note auf Fragenebene
- grafische Darstellung der Sensitivitätsanalyse
- geglättete relative Häufigkeit der Antwortmöglichkeiten nach Gesamtpunkten

durchschnittlich erreichte Prozent pro Note auf Fragenebene



geglättete relative Häufigkeit der Antwortmöglichkeiten nach Gesamtprozenten



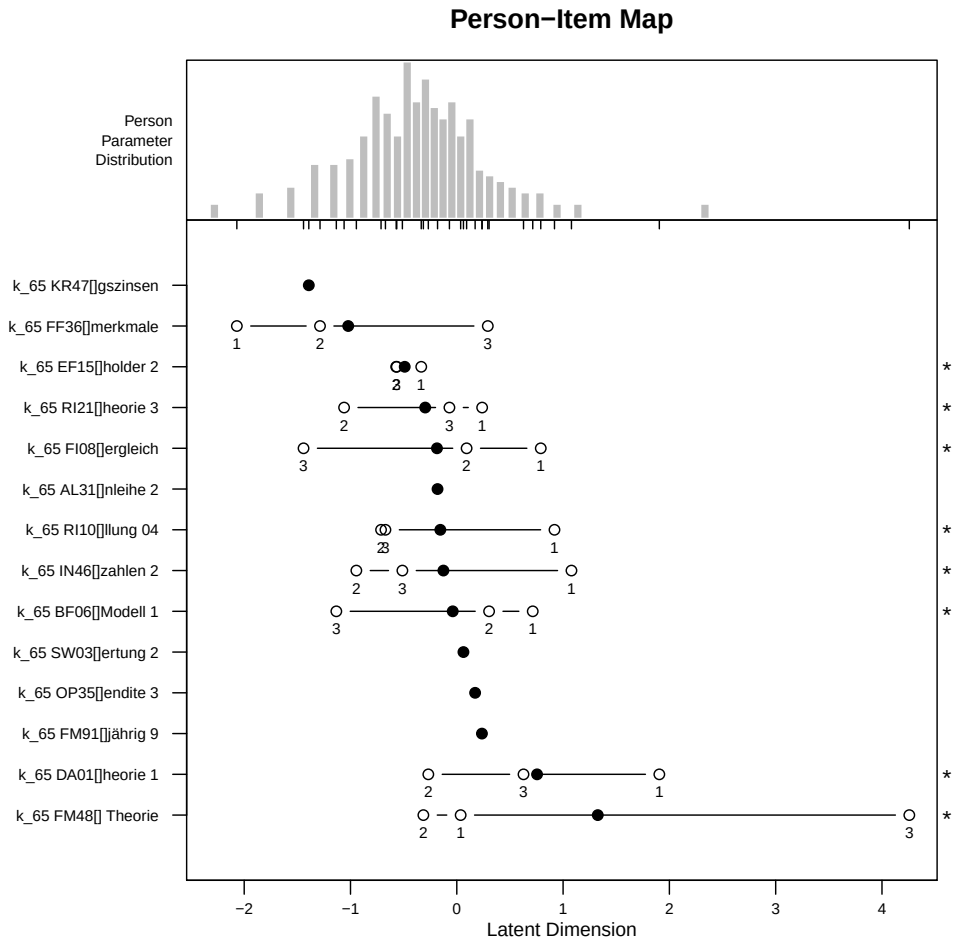
Auswertungen gemäß der Item Response Theory

Item Response Theory Modelle ermöglichen es, Fragenschwierigkeit und Personenfähigkeit getrennt voneinander zu schätzen.
Grundlegende Fragen:

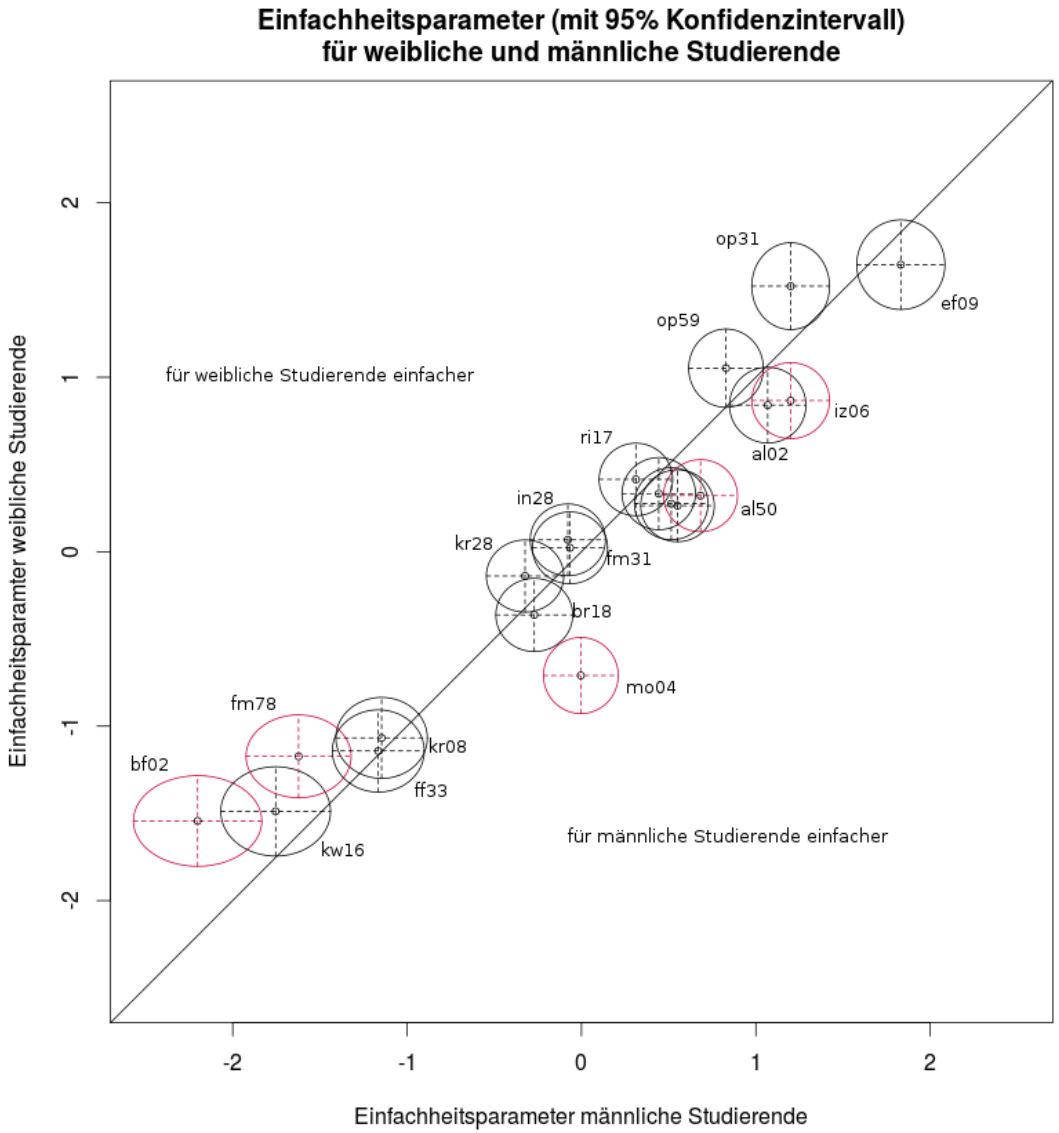
- Wie schwierig sind die Prüfungsfragen, wie fähig sind die Studierenden?
- Passen die Schwierigkeit der Fragen und die Fähigkeit der Studierenden zusammen?
- Benachteiligt die Prüfung Personen mit speziellen Eigenschaften? (Geschlecht, Herkunft, . . .)

Nachteil:
Für die Schätzung eines IRT-Modell müssen mehrere Voraussetzungen erfüllt sein (z.B. lokale stochastische Unabhängigkeit, spezifische Objektivitität, Eindimensionalität), daher kann das Modell korrekterweise nur für eine geeignete Untergruppe der Prüfungsfragen angewendet werden.

Person-Item Map



Grafischer Modellcheck (Prüfung Mai 2011)



Inhalt

Zusammenfassung und weitere Schritte

- Wir stellen Vortragenden ein Tool zur Verfügung, das ihre MC-Prüfungen automatisiert mit verschiedenen statistischen Methoden auswertet.
- Der erstellte Bericht enthält detaillierte Erklärungen,
  - wie die jeweiligen Ergebnisse zu interpretieren sind,
  - welche Schlüsse hinsichtlich der Prüfungsfragen gezogen werden können,
  - und welche Konstellationen Anlass zur genaueren Betrachtung bzw. Überarbeitung einer Frage liefern können.
- Insgesamt soll so den Vortragenden geholfen werden, die Qualität ihrer MC-Prüfungen einzuschätzen und zu verbessern.

Ausblick:

- Integration eines Vergleichs verschiedener Prüfungstermine hinsichtlich Fragenschwierigkeit und Personenfähigkeit
- Workshops für Prüfungsverantwortliche

Vielen Dank für Ihre Aufmerksamkeit!

michaela.nettekoven@wu.ac.at