Individualisierte, kompetenzorientierte E-Klausuren

Lara Weiß, Aron Fink & Jule Kunz

Partnership AKGU, Goethe-Universität Frankfurt

Projektbeschreibung

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Learning Design: kompetenzorientierte, adaptive E-Klausuren
- Klausuren nicht nur als formaler Abschluss eines Moduls, sondern
  - Im Sinne des Constructive Alignments als zusätzliche Lerngelegenheit implementiert und
  - Konsequent auf die intendierten Lernziele des Moduls bezogen
- E-Klausuren auf Basis des im BMBF-Projekts KAT-HS entwickelten Klausurkonzepts

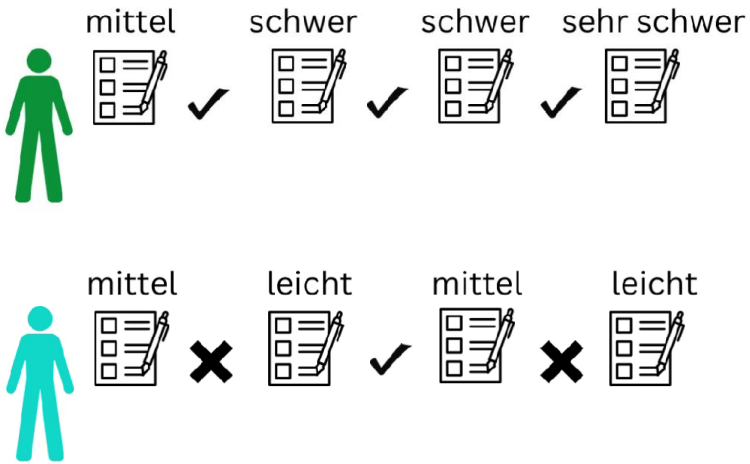
KAT-HS-Konzept

- Orientiert sich an etablierten Standards der **Kompetenzdiagnostik** und ermöglicht so eine systematische und präzise **Erfassung individueller Kompetenzstände** unter Einsatz **psychometrischer Modelle**
- **Interdisziplinär:** Für alle Studiengänge, in denen Lernziele mit standardisierten Klausuren geprüft werden können
- **Direkt umsetzbares Handlungswissen**

Spoden, C. & Frey, A. (Hrsg.). (2021). *Psychometrisch fundierte E-Klausuren für die Hochschule*. Lengerich: Pabst Publishers.

Computerisiertes adaptives Testen

- Auswahl der zur Bearbeitung vorgelegten Items **passt sich dem vorherigen Antwortverhalten der Testperson an** (z. B. Frey, 2023).
- Auswahl von Items mit möglichst **hoher diagnostischer Information** über die individuelle Merkmalsausprägung.
- Dafür Nutzung von Modellen der **Item Response Theory** (z. B. van der Linden, 2016)



Computerisiertes adaptives Testen

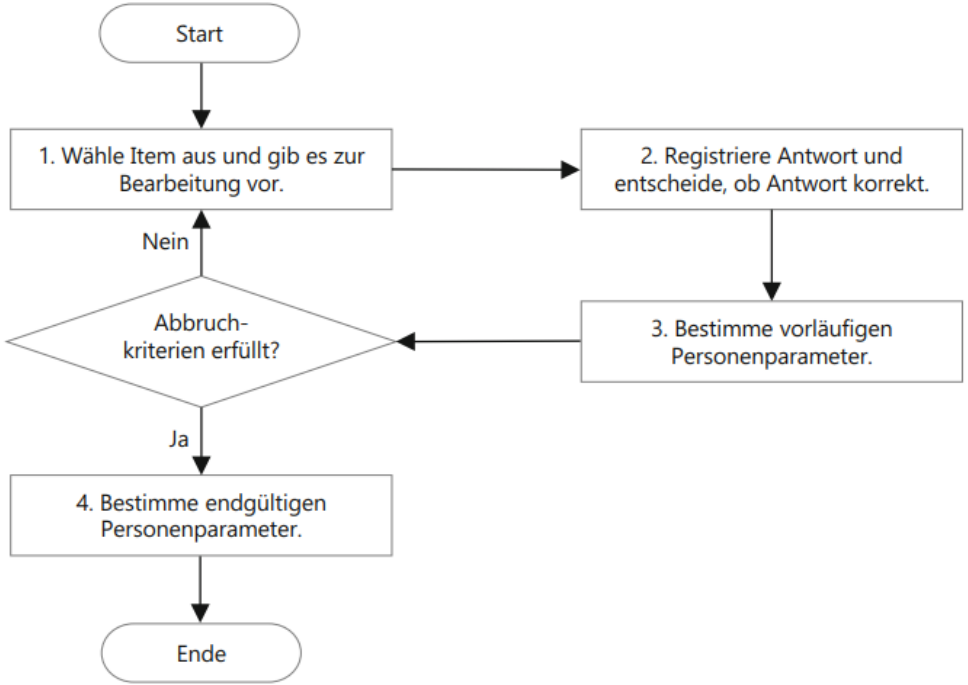
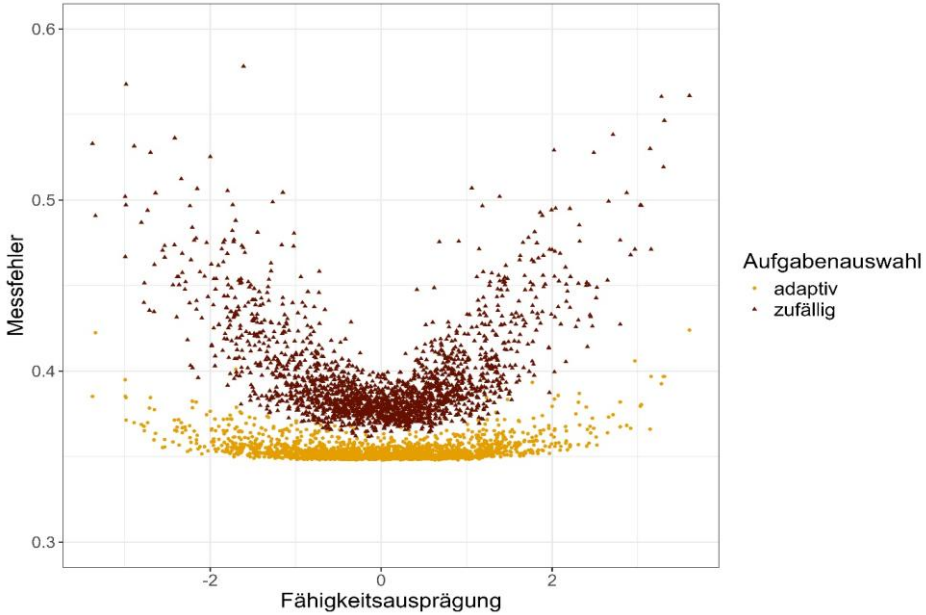


Abb. 20.1 Flussdiagramm zum typischen Ablauf computerisierter adaptiver Tests

Quelle: Frey (2020, S. 505)

Computerisiertes adaptives Testen

- Vorgehen führt zu **erhöhter Messeffizienz** im Vergleich zu linearen Tests (Segall, 2005).
- **Präzision der Klausurergebnisse** und der daraus abgeleiteten **Noten** kann so über das **gesamte Fähigkeitsspektrum** angeglichen werden.



Software

- KAT-HS-App zur **Konstruktion, Administration** und **Auswertung** kriteriumsorientierter, adaptiver und nicht-adaptiver Tests (Fink et al., 2021)
- **KAT-HS-Konzept** bildet psychometrische Grundlage
- Programmiert in **R** (R Core Team, 2024)
- Zielt darauf ab **Zugänglichkeit** IRT-basierter Methoden zu verbessern und gleichzeitig die **methodisch angemessene Nutzung** sicherzustellen
- **Frei Verfügbar, Open Source**
- GUI, für viele Anwendungsgebiete geeignete Voreinstellungen, Vorlagen, Benutzerhandbuch, …

Weiterentwicklung der App – Erste Förderphase

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Entwicklung neuer Aufgabenformate (Mehrfachauswahl, KPrim, R-Code)
- Funktion zur Klausureinsicht
- Skalierbare Serverinfrastruktur

Aufgabe 2 / 4

Bitte auf Weiter klicken, um zur nächsten Aufgabe zu gelangen.

Weiter

Zurück

Aufgabe R-Editor

Schuhgröße

Aufgabe

In Ihrem Studiengang wurde von einer kleinen Stichprobe die Schuhgröße erhoben. Sie bekommen die Daten in einer Datenmatrix (siehe Tabelle 1):

Tabelle 1

Schuhgrößen

ProbandIn	Schuhgröße
1	36
2	38
3	40
4	38
5	42
6	46

Berechnen Sie Mittelwert, Standardabweichung sowie die Varianz der Schuhgröße in der Stichprobe. Nutzen Sie hierfür den R-Editor. Welche Werte ergeben sich?

Bitte kreuzen Sie die korrekte Antwort an.

- (●) M = 40, SD = 4.2, Var = 17.64
- ( ) M = 40, SD = 3.58, Var = 12.8
- ( ) M = 41, SD = 3.58, Var = 12.8
- ( ) M = 41, SD = 4.2, Var = 17.64

Evaluation

- Durchführung Klausur zur Vorlesung „Einführung in Unterricht sowie Diagnostik und Beratung“ für Lehramtsstudierende (*N* = 694)
- In einer Studierendenbefragung in Anschluss an die Klausur (*N* = 350) wurde die Software als übersichtlich und leicht zu bedienen eingeschätzt, ebenso positiv die Vorgabe am PC statt auf Papier

Weiterentwicklung zum Plugin

- **Schnittstelle von Statistiksoftware R** (R Core Team, 2024) **zu Moodle** eröffnet Möglichkeit zur Nutzung von State-of-the-Art Methoden des adaptiven Testens
- Zudem direkte Integration in die E-Learning-Infrastruktur der Goethe-Universität
- Entwicklung eines vollständigen **Moodle-Plugins** zur Konfiguration und Administration adaptiver Tests

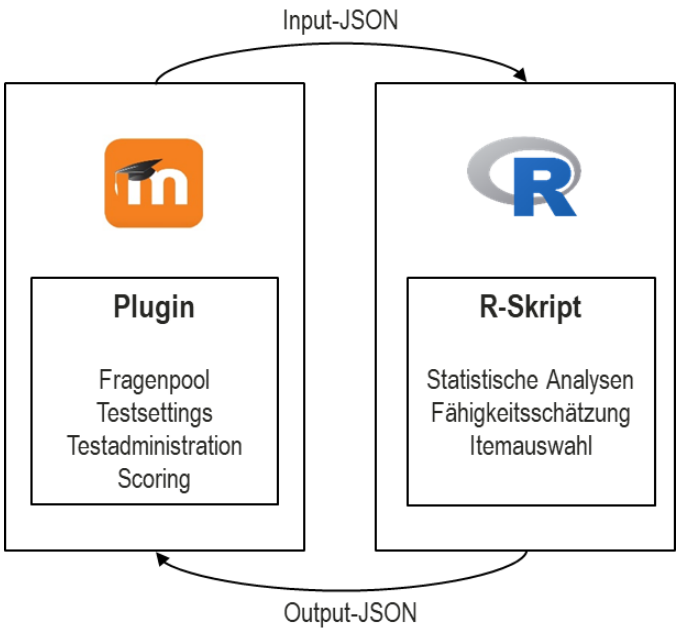
Technische Umsetzung Plugin

Architektur

- Separater R-Server
- **Statistische Berechnungen** laufen serverseitig mit **R**
- **Testvorgabe** und **Scoring** erfolgen in **Moodle**
- **Flexible Anpassung** des Algorithmus durch serverseitiges R-Script möglich, ohne Moodle-Plugin zu verändern
- **Kommunikation** zwischen Moodle und R-Server über **REST-API**

Installation

- ZIP-Upload in die jeweilige Moodle-Installation
- Moodle-Version > 3.8

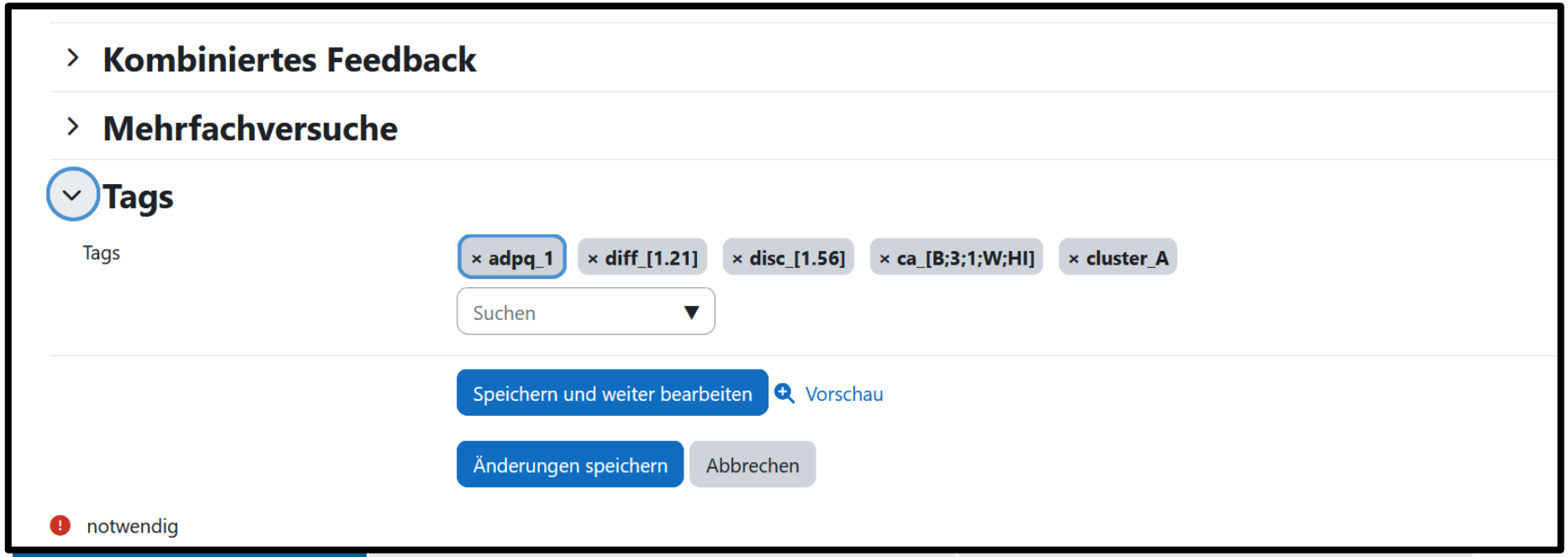


Plugin Itempool-Tags

Tag	Beispiel	Beschreibung
adpq_1	adpq_1	Marker, dass Frage für adaptiven Itempool markiert ist
diff_[array with floats]	diff_[1.2] für dichotome Items diff_[1.2;2.34;5.54] für polytome Items	Schwierigkeitsparameter
disc_[array with floats]	disc_[1.4] für eindimensionale Tests disc_[1.9;0;1.4] für dreidimensionale Tests	Diskriminationsparameter
enemy_[array with ItemID]	enemy_[ItemID1;ItemID456]	ItemID von Enemy-Items
ca_[array with strings]	ca_[Leseverstehen;Verständnis;Märchen;MC]	Inhaltsbereiche/Itemeigenschaften
cluster_CLUSTERNAME	cluster_A, cluster_C1, cluster_L2	Cluster i.S.d. kontinuierlichen Kalibrierungsstrategie

Plugin Itempool-Tags

Tag	Beispiel	Beschreibung
		Kalibrierungsstrategie



Usability Test und Evaluation Plugin

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

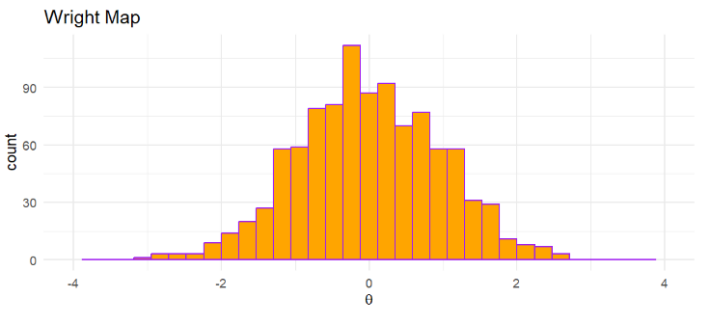
- Performance-Test: ca. 500-600ms R-Script + 70-80ms Kommunikation Schnittstelle
- Erster Usability-Test im Rahmen von 3 Seminaren an der Goethe-Universität verlief ohne Zwischenfälle
- Nachbefragung mit *N* = 51 Personen zeichnet positives Bild

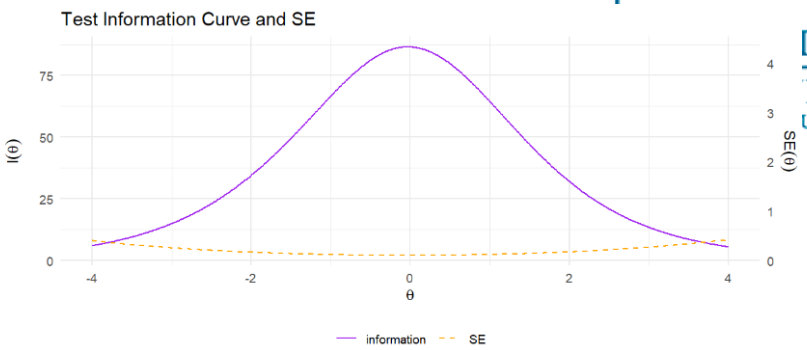
Item	*M*	*SD*
Ich finde, die Benutzeroberfläche der Testsoftware ist übersichtlich gestaltet.	3.47	0.58
Die Bedienung der Testsoftware fiel mir leicht.	3.62	0.60
Die Zeit zwischen dem Klick auf „weiter“ und dem Anzeigen der nächsten Aufgabe, empfand ich als zu lang. (-)	1.54	0.67
Die Ladezeiten in der Testsoftware waren angemessen.	3.48	0.71
Die Navigation in der Testsoftware fiel mir leicht.	3.57	0.57

4-stufige Antwortskala („trifft gar nicht zu“ bis „trifft völlig zu“)

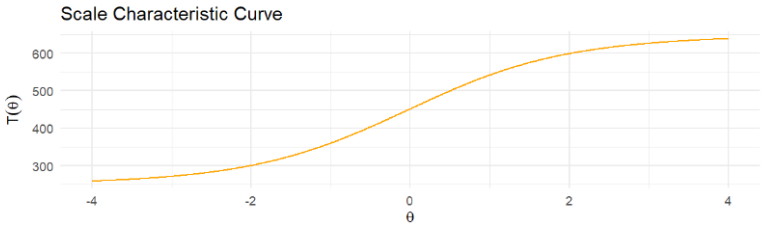
R-Paket

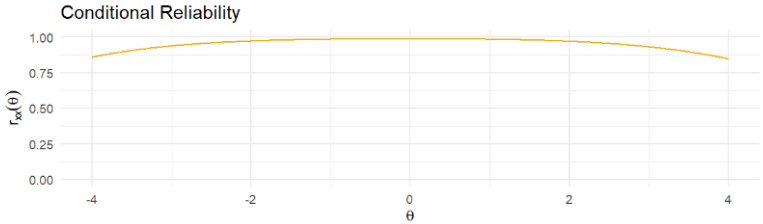
- R-Paket zur einfachen Testauswertung der Antwortdaten aus Moodle
- Klassische Itemstatistiken, IRT-Itemstatistiken, Fähigkeitsschätzungen und weitere Testinformationen
- Analysen auf aktuellem psychometrischen Forschungsstand auch ohne psychometrischen Hintergrund machbar





b category
- b
- b1
- b2
- b3
- b4





Fazit

- Erster Usability-Check des Plugins zeigt, dass **Anbindung** von R an Moodle wie gewünscht **funktioniert**
- Plugin ermöglicht Durchführung **linearer** und **adaptiver Tests auf aktuellem wissenschaftlichen Stand**
- Durch Möglichkeit adaptiven Algorithmus **individuell** anzupassen, lässt sich das Plugin auf **verschiedene Anwendungsbedürfnisse adaptieren**
- Das zugehörige **R-Paket** erleichtert zudem die Analyse der Testdaten
- E-Klausuren können Barrierefreiheit von Klausuren durch Integration passgenauer Maßnahmen erhöhen (z.B. Screenreader)

Ausblick

- Einsatz im Rahmen einer großen Vorlesung im SoSe25 geplant
- Veröffentlichung erste Version Plugin + R-Paket zum Ende der Projektlaufzeit
- Geplant ist zudem eine Softwarenote in einem wissenschaftlichen Fachjournal

Vielen Dank für die Aufmerksamkeit

Anhang

Evaluation

Item	M	SD
Das Antwortformat dieser Klausur hat mir Probleme bereitet. (-)	2.00	0.84
Ich finde, die Klausuraufgaben in Sb1 waren präzise und eindeutig formuliert.	2.87	0.72
Ich finde, die Arbeitsanweisungen zur Beantwortung der Klausuraufgaben waren eindeutig.	3.17	0.73
Ich finde, die Benutzeroberfläche der Klausursoftware ist übersichtlich gestaltet.	2.99	0.91
Die Bedienung der Klausursoftware fiel mir leicht.	3.43	0.68
Ich finde gut, dass die Klausur am Computer vorgegeben wurde und nicht auf Papier.	3.23	0.88

4-stufige Antwortskala („trifft gar nicht zu“ bis „trifft völlig zu“)

R-Server

- Der R-Server wird über **Docker** eingerichtet
- Als Image für den R-Server wird rstudio/plumber verwendet
- Start des R-Servers über kurzes R-Script

```
FROM rstudio/plumber
RUN apt-get update -qq && apt-get install -y \
        libssl-dev \
        libcurl4-gnutls-dev \
        r-base \
        r-base-core \
        r-base-dev \
        r-base-html \
        r-cran-boot \
        r-cran-class \
        r-cran-cluster \
        r-cran-codetools \
        r-cran-foreign \
        r-cran-kernsmooth \
        r-cran-lattice \
        r-cran-mass \
        r-cran-matrix \
        r-cran-mgcv \
        r-cran-nlme \
        r-cran-nnet \
        r-cran-rpart \
        r-cran-spatial \
        r-cran-survival \
        r-doc-html \
        r-recommended \
        r-cran-dplyr \
        r-cran-rjson \
        r-cran-jsonlite

        # Install R-packages
RUN R -e "install.packages('plyr')" > /install_plyr.log
RUN R -e "install.packages('mirtCAT')" > /install_mirtCAT.log
RUN R -e "install.packages('dplyr')" > /install_dplyr.log
RUN R -e "install.packages('rjson')" > /install_rjson.log
RUN R -e "install.packages('jsonlite')" > /install_jsonlite.log

# Expose port 80 for the web server
EXPOSE 80
```

Kommunikation Moodle → R

- R ist für CAT-spezifischen Berechnungen verantwortlich
- Kommunikation über **REST-API** mittels **JSON**

Input Moodle → R

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Itempool


```
{
  "courseID": [
    "2"
  ],
  "itempool": {
    "items": [
      {
        "cluster": "A",
        "content_area": [
          "I",
          "9",
          "25",
          "A",
          "KI"
        ],
        "dbID": "67",
        "diff": [
          "0.57"
        ],
        "disc": [
          "1.52"
        ],
        "enemys": [
        ],
        "ID": "I0925AKa"
      },
```


Testsettings


```
"settings": {
  "content_areas": {
    "distribution": [
      "KI:0.5;HI:0.5"
    ],
    "enabled": [
      "1"
    ]
  },
  "criteria_adaptive": [
    "MI"
  ],
  "criteria_not_adaptive": [
    "random"
  ],
  "exposure": {
    "enabled": [
      1
    ],
    "nitems_exposure": [
      "4"
    ]
  },
  "maxItems": [
    "12"
```


Testverlauf


```
    "itemID": [
      "U0206VHa",
      "U0206AHa"
    ],
    "itemtime": [
      21,
      4
    ],
    "scoredResponse": [
      0,
      0
    ],
    "timeout": [
      false
    ]
  },
  "testID": [
    "2"
  ]
}
```

Output R → Moodle

Fähigkeitsschätzung + nächstes Item

```
{
"personID":["2"], // user id
"terminated":[false], // is test terminated
"theta":[-1.3127], // ability measure
"SE":[0.3918], // standard error of ability measure
"nextItem":["B0720VHb"] // next item id
}
```

KKS: Ziele

1. **Geringer Initialaufwand**
2. **Stabile Berichtsmetrik**
3. **Kontinuierliche Verbesserung** Itemparameterschätzungen
4. **Itempoolvergrößerung** während Testbetrieb
5. Identifikation **problematischer Items**

plus

6. **Kontrolle** von **Itempositionseffekten**

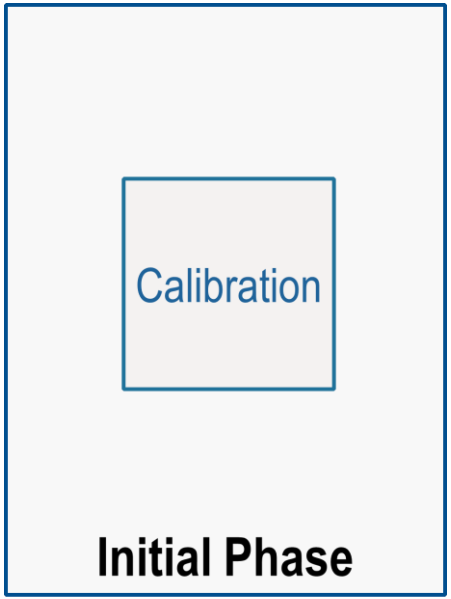
Itemtypen & Itemcluster

Items

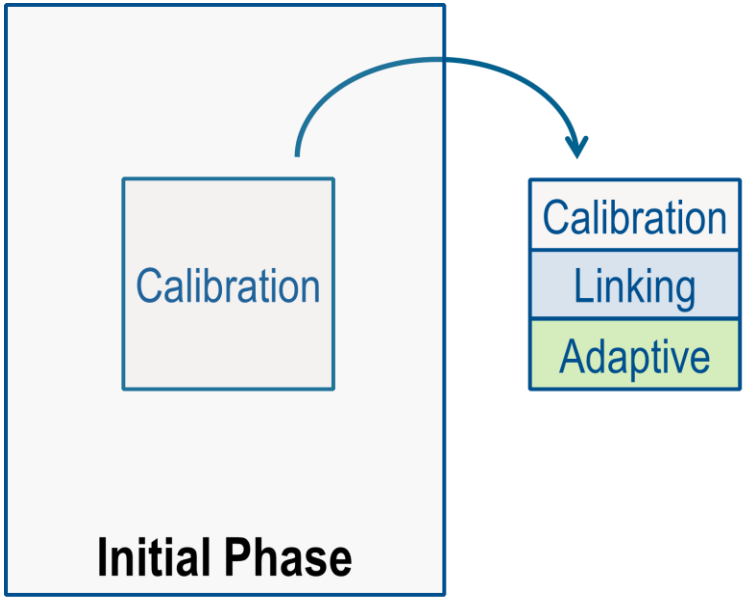
- Link Items
- Neue Items
- Adaptive Items

- **Cluster**
- Linkingcluster
- Kalibrierungscluster
- Adaptive Cluster

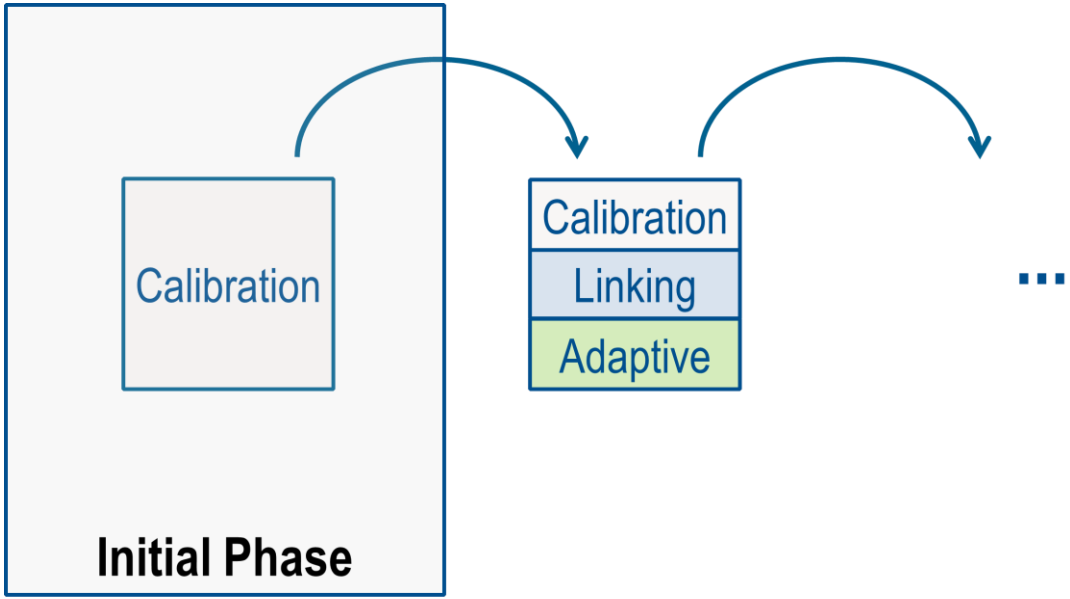
KKS: Phasen



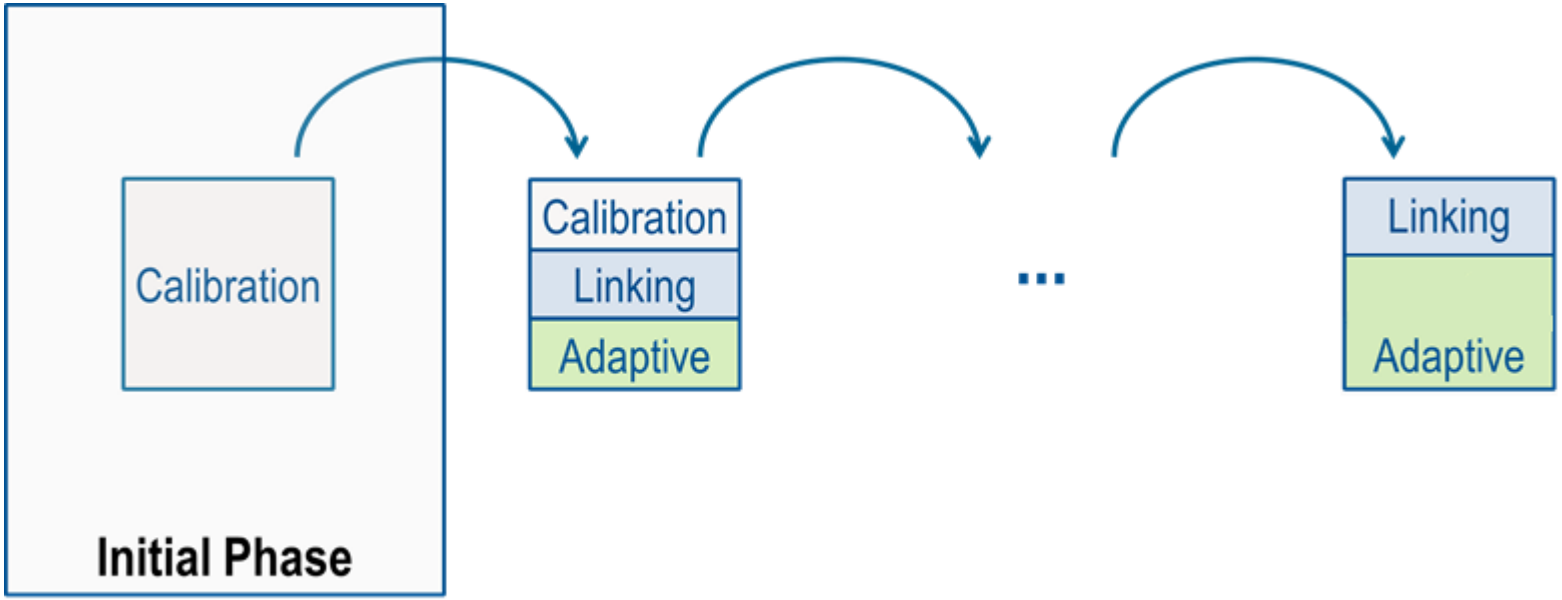
KKS: Phasen



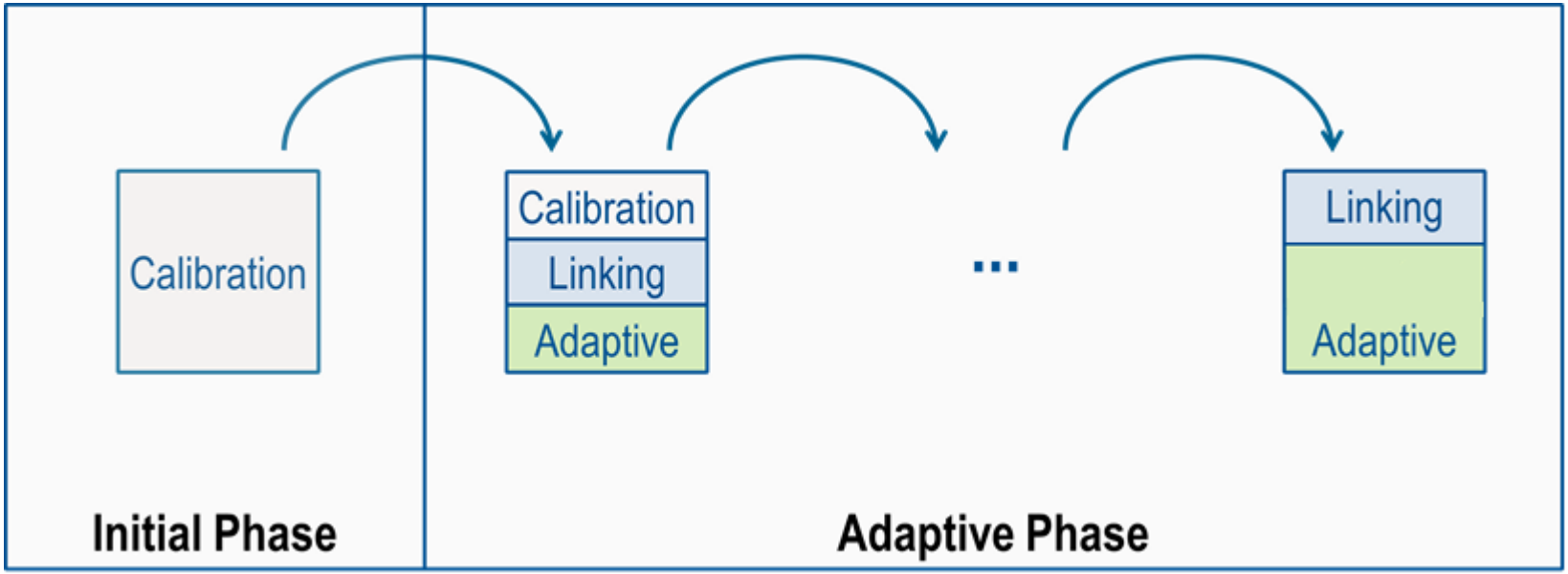
KKS: Phasen



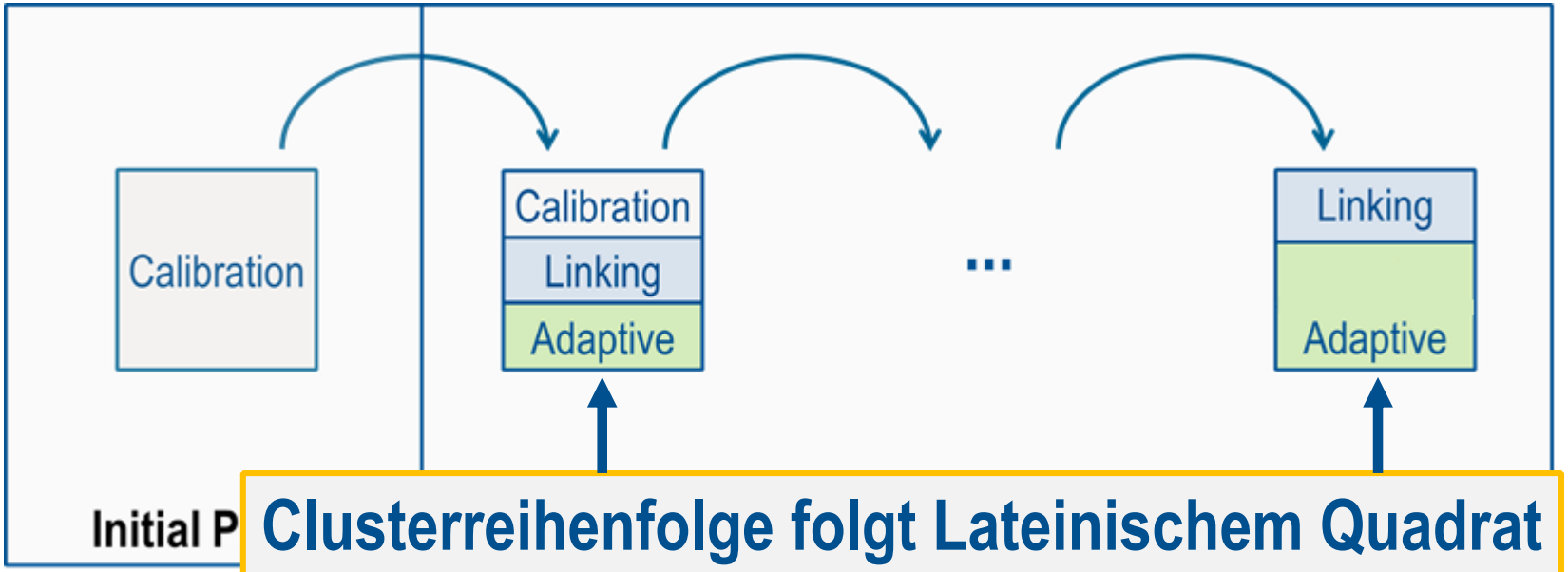
KKS: Phasen


Calibration
Initial Phase
Calibration
Linking
Adaptive
...
Linking
Adaptive

KKS: Phasen


Calibration
Calibration
Linking
Adaptive
...
Linking
Adaptive
Initial Phase
Adaptive Phase

KKS: Phasen



Clusterreihenfolge folgt Lateinischem Quadrat

- Alle Itemcluster auf allen Positionen präsentiert
- Gleicher mittlerer Effekt der Itemposition für alle Itemparameter
- Keine Unterschiede von IPE zwischen Testpersonen

Literatur

Literatur

Frey, A. (2023). Computerized adaptive testing and multistage testing. *In International Encyclopedia of Education* (Fourth Edition) (pp. 209–216). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10028-4

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42, 1-28. https://doi.org/10.18637/jss.v042.i08

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Psychology Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. MESA Press.

Nandakumar, R., & Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Research report). Newtown, PA: Law School Admission Council.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules. R package version* 4.1-4. https://CRAN.R-project.org/package=TAM

Van der Linden, W. J. (Hrsg.). (2016). Handbook of item response theory (Vol. 1). New York: CRC press.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In: W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8 17

Zwick, R., Thayer, D. T., & Wingersky, M. (1994) A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement*, 18, 121–140.