Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Entwicklung einer Applikation zur Erhebung und Visualisierung psychometrischer Kennzahlen aus Lernstandskontrollen in Ilias

Mark Czepan

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Christoph Stach
Betreuer/in:	Dipl. Päd. Jan Vanvinkenroye
Beginn am:	1. April 2021
Beendet am:	1. Oktober 2021

Kurzfassung

Daten gehören heutzutage zu den wertvollsten Ressourcen unserer Zeit. Sie werden in Massen gesammelt, um wertvolle Informationen zu extrahieren und zu monetarisieren. Durch die große und immer weiter wachsende Anzahl an Portalen und Werkzeugen, die der Universität und ihren Mitarbeitenden zur Verfügung stehen, steigt auch die Zahl der gewonnen Daten. Diese bleiben jedoch oft ungesehen und ungenutzt und verschenken somit ihr Potenzial, sinnvoll weiterverarbeitet zu werden. Deshalb haben sich die Technischen Informations- und Kommunikationsdienste der Universität Stuttgart dazu entschieden, dieses Themenfeld aufzugreifen und im Rahmen einer Bachelorarbeit zu bearbeiten.
Hierbei wurde sich für die Weiterverarbeitung von Daten aus Lernstandskontrollen entschieden. Diese Daten liegen sowohl in Ilias als auch in EvaExam vor und bilden somit die Grundlage für die Auswertung durch die psychometrische Testtheorie.
Durch diese Methode sollen die vorhandenen Daten sinnvoll aufbereitet und für Lehrende bereitgestellt werden. Dies soll die Lehrenden dazu motivieren, ihre Prüfungen an den richtigen Stellen zu verbessern und dazu dienen, die Lehre im digitalen Zeitalter weiter zu voranzutreiben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In der heutigen Zeit fallen immer und überall Daten in elektronischer Form an. Sie gelten als wertvolle Ressource unserer Zeit. Die größten Unternehmen der Welt verdienen ihr Geld mit Daten.[1] Dabei sind die Handlungsfelder für die Verarbeitung nahezu unbegrenzt, da überall Daten anfallen. Der Computer spielt eine zentrale Rolle in jedem Bereich des täglichen Lebens, dadurch sind diese leicht elektronisch zu erfassen und zu speichern. So ist er das zentrale Werkzeug für die Durchführung digitaler Lehre an Universitäten. Dieser Umstand wird durch die weltweite Coronoakrise verschärft und führt zu einem Ausbau der digitalen Lehre.[2] Die Herausforderungen, die durch die Digitalisierung von Lehrinhalten entstehen, können aber auch als Chance begriffen werden. Durch die steigende Zahl an Daten entsteht so die Möglichkeit, diese auszuwerten und Wissen zu generieren, welches ohne nicht möglich gewesen wäre. Jedoch bedarf es für die Analyse der Daten an Wissen und Kompetenz, damit diese Chance ergriffen werden kann.

Dabei besteht digitale Lehre nicht nur aus Abhalten von Seminaren oder Bereitstellen von Kursinhalten über das Internet. Ein wichtiger Bestandteil der digitalen Lehre besteht auch in der Erstellung, Durchführung und Auswertung von Lernstandskontrollen, beispielsweise in Form von Onlineklausuren. Diese können heutzutage einfach über verschiedene Plattformen erstellt und durchgeführt werden. [3] Diese elektronischen Methoden der Lernstandskontrolle führen zu einer großen Zahl an Daten, die weiterverarbeitet werden können. Da die Technischen Informations- und Kommunikationsdienste der Universität Stuttgart (TIK) zuständig für die Verwaltung und Instandhaltung dieser Dienste sind, wurde beschlossen, eine Anwendung zu entwickeln, die den Lehrenden dabei helfen soll, ihre Lernstandskontrollen besser zu analysieren. Dabei wird im Rahmen dieser Arbeit ergründet, wie Lernstandskontrollen für Lehrende sinnvoll ausgewertet und aufbereitet werden können.

Es existieren bereits verschiedene Theorien und Techniken zur Auswertung und Analyse von Lernstandskontrollen, die sogenannte Testtheorie. Diese Testtheorien lassen sich in die Klassische Testtheorie und die Item-Response-Theorie einteilen und bilden die Grundlage für diese Arbeit. Sie basieren auf einer wissenschaftlichen Grundlage und fusen auf stochastischen Verfahren. Dies erschwert die Interpretation der Ergebnisse, da hier theoretische Grundlagen vorausgesetzt werden, [Lie69] die bei unterschiedlichen Lehrpersonen unterschiedlich ausgeprägt sind. Dies ist dem Umstand geschuldet, dass die Menge an Lehrpersonen an der Universität eine heterogen Masse bildet, welche viele verschiedene Lehrhintergründe besitzt.

[1] https://de.statista.com/infografik/25062/wertvollste-unternehmen-der-welt-nach-marktkapitalisierung/

[2] https://hochschulforumdigitalisierung.de/de/dossiers/digitale-pruefungen

[3] Gilt für die Universität Stuttgart.

Aus diesem Grund beschäftigt sich diese Arbeit auch damit, wie die ausgewerteten Daten den Lehrpersonen gut vermittelt und dargestellt werden können ohne, dabei ein Vorwissen von den Lehrenden zu verlangen. Nichtsdestotrotz stellt die Testtheorie einen unerlässlichen Teil für die Testkonstruktion dar, da nur ein gut entworfener Test genau das testet, was getestet werden soll [Lie69].

1.1 Ziele der Arbeit

Das Ziel der Arbeit besteht darin, den Lehrenden eine Möglichkeit an die Hand zu geben, die von Ihnen erstellten Klausuren besser analysieren und auswerten zu können. Die Auswertung der Klausuren erfolgt dabei durch die psychometrische Testtheorie, welche sich in die Klassische Testtheorie und die Item-Response-Theorie einteilen lässt. Ein wichtiger Teil der Arbeit besteht auch darin, die Klausuren mithilfe einer deskriptivstatistischen Evaluation der Items auszuwerten. In dieser Arbeit wird dazu eine Anwendung erstellt, welche die oben genannten Methoden implementiert und für die Lehrenden verständlich aufbereitet und visualisiert. Dabei wird versucht, die breite Masse der Lehrenden zu erreichen und gleichzeitig den größtmöglichen Informationsgehalt in der Darstellung der Daten zu bewahren.

1.2 Aufbau der Arbeit

Zunächst wird in Kapitel 2 ein typischer Anwendungsfall beschrieben und wie die Arbeit daran anknüpfen kann. In diesem Kapitel werden auch die Anforderungen an die Anwendung formuliert. In Kapitel 3 werden die mathematischen Grundlagen, auf denen die Itemanalyse und die Testtheorien basieren, näher beleuchtet. Daraufhin wird in Kapitel 4 auf das Konzept hinter der Arbeit näher eingegangen und die Entscheidungsfindung für die Anwendung begründet. Hierbei wird auch der für diese Arbeit erstellte Fragebogen näher vorgestellt. In Kapitel 5 wird auf die Implementierung der Arbeit eingegangen. Kapitel 6 evaluiert die Umsetzung und vorgestellten Theorien dieser Arbeit. Den Schluss bildet Kapitel 7, das eine Zusammenfassung der Arbeit und einen Ausblick auf weitere Möglichkeiten gibt.

2 Anwedungsfall

Dieses Kapitel beschreibt einen typischen Anwendungsfall, auf den sich diese Arbeit bezieht. Sie soll dabei helfen, die Klausurauswertung und -analyse für Lehrende zu verbessern. Der Anwendungsfall ist dabei sehr spezifisch, da sich die Arbeit mit der Entwicklung einer Applikation beschäftigt, die für die Lehrenden der Universität Stuttgart bereitgestellt wird. Durch die breite Masse an Lehrenden aus verschiedenen Disziplinen besteht die Herausforderung insbesondere darin, die erarbeiteten Informationen bestmöglich aufzuarbeiten. Dabei soll es gelingen, die Ergebnisse gehaltvoll, verständnisvoll und kompakt aufzubereiten und darzustellen, damit alle Lehrenden davon profitieren können. Die so erhaltenen Daten sollen ohne Mehraufwand für die Lehrenden zur Verfügung gestellt werden. Im folgenden Abschnitt 2.1 wird erläutert, wie die Anwendung in den Workflow der Lehrenden eingebunden werden soll. In Abschnitt 2.3 werden mögliche Anwendungsbereiche aufgezeigt und in Abschnitt 2.2 die Anforderungen spezifiziert.

2.1 Klausuranalyse als Teil des Workflows

In diesem Abschnitt wird erläutert, wie der typische Ablauf beim Erstellen einer Klausur ist und an welcher Stelle die Arbeit dabei helfen soll, diesen zu verbessern. Dabei wird herausgearbeitet, welche Aufgaben an die Anwendung gestellt werden. In Abschnitt 2.2 werden diese Aufgaben in Anforderungen umformuliert, die an die Anwendung gestellt werden.

Klausuren werden an der Universität Stuttgart in der Regel in Eigenregie, in Ilias[1] oder in EvaExam[2] von den Lehrenden erstellt. Dabei bieten diese Verfahren nur wenig bis gar keine Möglichkeiten, die Klausuren oder deren Aufgaben bezüglich der verschiedenen Analyseverfahren der Testtheorie zu bewerten und deren Metriken zu erfassen.
Die Lehrenden greifen bei der Erstellung ihrer Klausuren oft auf einen Pool aus Aufgaben zurück. Aus diesen Aufgaben entstehen die Klausuren auf der jeweiligen Plattform. Nach der Auswertung bekommen die Lehrenden in EvaExam nur wenige und unübersichtliche Informationen zu ihren Klausuren.[3] In Ilias besteht die Möglichkeit einen Fragepool zu erstellen und diesem Kategorien wie Schwierigkeit zuzuordnen. Ilias führt jedoch lediglich eine Korrektur der Fragen durch und keine weitreichendere Analyse.[4] Sobald die Lehrenden eigene Klausuren erstellen, müssen sie diese Erhebungen selbst tätigen. Für die Erstellung von weiteren Klausuren können die Lehrenden nur auf unübersichtliche und wenige Informationen zurückgreifen, um ihre Aufgaben und somit ihre Klausuren zu verbessern. An dieser Stelle soll die Arbeit eine Möglichkeit bieten,

[1] Plattform der Universität Suttgart: https://ilias3.uni-stuttgart.de
[2] Plattform der Universität Suttgart: https://scanklausuren.uni-stuttgart.de
[3] Dieser Umstand wurde dem TIK von Lehrenden zurückgemeldet
[4] Siehe dazu Ilias der Universität Stuttgart

ohne Mehraufwand für die Lehrenden bessere Auswertungen für ihre Klausuren bereitzustellen (**Ag1**). Präziser bedeutet dies, dass die Lehrenden keinen Zugriff auf die eingesetzten Methoden haben sollen und sie keine Anpassung der Parameter vornehmen dürfen. Dies verhindert, dass die Lehrenden grobe Fehler begehen und so die Ergebnisse des Programms verfälschen (**Ag2**). Außerdem soll die Anwendung einfach gestaltet werden und dennoch die Möglichkeit bestehen, dass jede Lehrperson daraus einen Nutzen ziehen kann (**Ag3**). Dabei soll der Workflow für die Klausurerstellung möglichst unverändert bleiben. Die Lehrenden sollen bei der Auswahl ihrer Aufgaben für die Klausuren unterstützt werden. Die ausgewerteten Daten sollen daher für jede Lehrperson lesbar und interpretierbar sein sowie nur die wichtigsten Informationen enthalten (**Ag4**). Diese Informationen sollen nur mithilfe der Anwendung für jede Lehrperson lesbar und interpretierbar sein, darf jedoch keine Lehrperson ausschließen (**Ag5**). Diese Anforderung ist besonders herausfordernd, da sie den Lehrenden neue Modelle und Methoden näher bringen soll. Beispielsweise haben Institute mit naturwissenschaftlichen Hintergrund bessere Kenntnisse im Bereich der Mathematik und Statistik als Intitute mit einem geisteswisschenschaftlichen Hintergrund. Die Auswertung soll jedoch für Lehrende aller Fachrichtungen leicht interpretierbar sein.
Der Vorteil bei der Art der Daten und ihrer Verarbeitung besteht darin, dass vorhandene Methoden und Modelle genutzt werden können. Des Weiteren werden die benötigten Daten von Ilias und EvaExam bereits gut strukturiert bereitgestellt. Dies erleichtert die Weiterverarbeitung der Daten und die Eingliederung in eine eigene Datenstruktur.

Hieraus lassen sich die Anforderungen an die Anwendung spezifizieren, die im folgenden Abschnitt 2.2 genauer beschrieben werden. In Abschnitt 2.3 werden daraus drei mögliche Anwendungsbereiche für die Anwendung abgeleitet. Daraufhin wird eine bestimmte Plattform zur konkreten Implementierung der Anwendung bestimmt.

2.2 Anforderungen

Aus dem oben aufgeführten Abschnitt 2.1 lassen sich nun die Anforderungen ableiten. Die folgenden fünf Anforderungen werden an das Programm gestellt.

A1 (aus Ag1) **Zugänglichkeit** Die Anwendung soll für die Lehrenden der Universität Stuttgart leicht zugänglich und erreichbar sein. Das bedeutet, sie soll auf einer Universitäts-Website gehostet oder auf Ilias oder EvaExam genutzt werden können. Dies soll ohne Mehraufwand für die Lehrenden geschehen.
A2 (aus Ag2) **Benutzerfreundlichkeit** Die Anwendung soll möglichst einfach gehalten werden. Beispielsweise bedeutet dies, keine Konfigurationsmöglichkeiten von Parametern zuzulassen, da dies zur Verfälschung der Ergebnisse führen kann. Die Lehrenden sollen ohne viele Klicks oder Wartezeit zu ihren Ergebnissen kommen.
A3 (aus Ag3) **Reduziertheit** Die Anwendung soll nur die nötigsten Informationen enthalten. Sie soll die Lehrenden dabei nicht mit einer Informationslast überfordern, sondern ihnen eine Hilfestellung bieten. Dabei soll jedoch die Informationstiefe gewahrt bleiben.
A4 (aus Ag4) **Übersichtlichkeit** Die Anwendung soll die Informationen übersichtlich darstellen. Dabei soll die Lehrperson selbst auswählen dürfen, welche Information für sie von Bedeutung ist.

A5 (aus Ag5) **Praktikabilität** Die Anwendung soll möglichst einfach bedienbar sein. Die Lehrenden sollen dabei jegliche Information, die sie erhalten mithilfe des eigenen Kenntnisstandes oder der Anwendung lesen und interpretieren können.

2.3 Mögliche Anwendungsbreiche

In diesem Abschnitt werden mögliche Einsatzszenarien der Anwendung auf einer Plattform beschrieben. Anschließend werden diese diskutiert. Unter Berücksichtigung der Anforderungen soll entschieden werden, auf welcher Plattform die Anwendung implementiert werden soll.

Ilias Ilias ist eine freie Software, die an der Universität Stuttgart als Lernmanagmentsystem eingesetzt wird. Sie bietet den Lehrenden zahlreiche Möglichkeiten, Studierenden Lerninhalte zur Verfügung zu stellen. Darunter fallen einerseits das Bereitstellen von Lehrmaterial, zum Beispiel in Form von Vorlesungsfolien, (Vorlesungs-)Videos, Programmierübungen oder Übungsblättern. Andererseits besteht in Ilias auch die Möglichkeit, die Studierenden in Form von Lernstandskontrollen abzufragen. Daran kann die Anwendung anknüpfen und den Lehrenden zusätzliche Möglichkeiten bieten, ihre Leistungstest nach den in Kapitel 3 genannten Verfahren analysieren zu lassen.
Ilias steht dabei allen Lehrenden bereit. Diese haben zum Großteil auch schon Erfahrungen damit, da es zum Lehralltag der Universität Stuttgart gehört.

EvaExam EvaExam ist eine von der Universität Stuttgart lizensierte Software zur Erstellung und Auswertung von Klausuren. Diese können über ein Onlinewerkzeug erstellt werden. Klausuren können als Onlineprüfung oderin Papierform durchgeführt werden. Anschließend werden diese durch EvaExam ausgewertet. Dabei bietet EvaExam einen Gesamtbericht an, welcher die Prüfung mithilfe von Methoden der deskriptiven Itemanalyse analysiert. Daran könnte die Anwendung anknüpfen und die gegebenen Methoden übersichtlicher darstellen und um nicht implementierte Methoden und Modelle erweitern.
EvaExam bietet den Vorteil, dass es explizit für Klausuren genutzt wird. Die Anwendung würde in dieser Umgebung optimal eingebunden sein, da die Lehrenden die Plattform nur zur Klausurerstellung nutzen.

Eigene Anwendung Eine eigene Anwendung könnte beispielsweise in Form eines zu installierenden Programms oder einer Webanwendung implementiert werden. Dabei gibt es keine Methoden und Modelle, die bereits implementiert wurden.
Die Anwendung wäre somit eigenständig nutzbar und könnte flexibel eingesetzt werden. Jedoch gilt es zu beachten, dass die Lehrenden in diesem Fall ihre Ergebnisse manuell weiterverarbeiten müssen.

Im Folgenden werden die möglichen Plattformen bezüglich der Anforderungen evaluiert.

A1 Die Zugänglichkeit zu der Anwendung soll für die Lehrenden ohne viel Mehraufwand und so einfach wie möglich gestaltet sein. Dafür eignen sich Ilias und EvaExam am besten, da sie schon im Universitätskontext genutzt werden. Eine eigene Anwendung würde hier den

Nachteil haben, dass die Lehrenden ausschließlich zur Analyse ihrer Lernstandskontrollen die Anwendung wechseln müssten. Dies widerspricht der Anforderung.

A2 Zu dieser Anforderung zählt die automatisierte Durchführung der Analyse. Dies entfällt bei einer eigenen Anwendung. Hier müssten die Lehrenden manuell ihre Ergebnisse hochladen und verarbeiten, um so zu einem Ergebnis zu kommen. Ilias und EvaExam hingegen sind geschlossene Systeme, bei denen die Anwendung als Plugin verwirklicht werden kann. So kann die Anwendung automatisiert die Lernstandskontrollen analysieren und die Ergebnisse darstellen.

A3- A5 Diese Anforderungen haben keinen Einfluss auf die Wahl der Anwendungsumgebung und werden daher in dieser Diskussion nicht berücksichtigt.

Es wird deutlich, dass die Entwicklung einer eigenen Plattform mehr Nachteile als Vorteile gegenüber schon vorhandenen Plattformen birgt. Daher wird sich gegen die Entwicklung einer eigenen Plattform für die Anwendung entschieden. Ilias und EvaExam bieten fast die selben Vor- und Nachteile und damit auch die gleichen Möglichkeiten. Aus folgenden Gründen wurde sich in dieser Arbeit für EvaExam als Plattform entschieden:

1. EvaExam bietet den größeren Vorteil bezüglich der Zugänglichkeit. Die Lehrperson befindet sich hier nur für die Erstellung und Auswertung einer Klausur. Da Ilias zwar die Möglichkeit besitzt, Lernstandskontrollen durchzuführen, jedoch auch viele andere Funktionen hat, könnte die Anwendung unbeachtet bleiben.

2. EvaExam stellt ein SDK bereit, Ilias hingegen setzt auf eine freiere Implementierung durch PHP, HTML und CSS. Durch diese Hilfestellung durch EvaExam ist es sinnvoller, die Anwendung auf dieser Plattform zu implementieren.

Die Entscheidung zur Implementierung der Anwendung fällt nach Abschätzung der Erfüllbarkeit der Anforderungen auf EvaExam. Hier wird die Anwendung in Form eines Plug-Ins in die vorhandene Oberfläche integriert. So wird die einfache Nutzung der Anwendung für die Lehrenden garantiert.

3 Grundlagen

In diesem Kapitel werden die Grundlagen der verschiedenen Analyseverfahren erläutert. In Abschnitt 3.1 gibt es eine Einführung in die grundlegenden Ideen und verwendeten Termini. Die darauffolgenden Abschnitte behandeln in 3.2 die deskriptivstatistische Itemanalyse, in 3.3 die Klassische Testtheorie und in 3.4 die Item-Response-Theorie. In ihnen werden die verschiedenen Methoden und Modelle vorgestellt, auf die sich diese Arbeit bezieht.

3.1 Einführung

"Ein Test ist ein wissenschaftliches Routineverfahren zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung " [Lie69]. Dabei können Tests in *Leistungstests* und *Persönlichkeitstests* kategorisiert werden. Leistungstest sollen dabei die kognitive Leistungsfähigkeit messen, darunter fallen beispielsweise Klausuren. Persönlichkeitstests hingegen messen das *typische Verhalten* eines Probanden [MK20].
Ein Test sollte dabei folgende Hauptgütekriterien [Lie69] erfüllen:

1. Objektivität
2. Reliabilität
3. Validität

Nach Lienert [Lie69] gibt es zusätzlich auch Nebengütekriterien. Diese sind die Normierung, die Testökonomie, die Nützlichkeit und die Fairness eines Tests. Dazu kommen nach Moosbrugger und Kelava. [MK20] noch die Zumutbarkeit, die Unverfälschbarkeit und die Skalierung. Diese sind jedoch von allgemein-planerischer Natur und entsprechen nicht den Kriterien, die speziell für die Testtheorie gelten. Für wissenschaftliche Tests und Fragebögen sind die Hauptgütekriterien der *Objektivität*, *Reliabilität* und der *Validität* von besonderer Bedeutung [Lie69]. Die *Objektivität* beschreibt die Unabhängigkeit der Ausführung eines Tests. Dabei ist die *Objektivität* bei Klausuren bereits gegeben, da jeder Teilnehmende dieselben Voraussetzungen hat. Die *Reliabilität* beschreibt die Messgenauigkeit eines Tests. Der Test misst das Merkmal, welches er misst, fehlerfrei. Die *Validität* eine Tests gibt eine Aussage über die Gültigkeit eines Tests. Sie sagt aus, ob der Test das gewünschte Merkmal misst oder ein anderes [Lie69]. Die Schätzung der Validität ist numerisch nur schwer zu erfassen und würde den Rahmen dieser Arbeit sprengen, weshalb darauf nicht näher eingegangen wird. Trotzdem sind gut entworfene Tests auch zuträglich für die Validität [Lie69]. Dies kann durch die folgenden Methoden und Modelle erreicht werden.

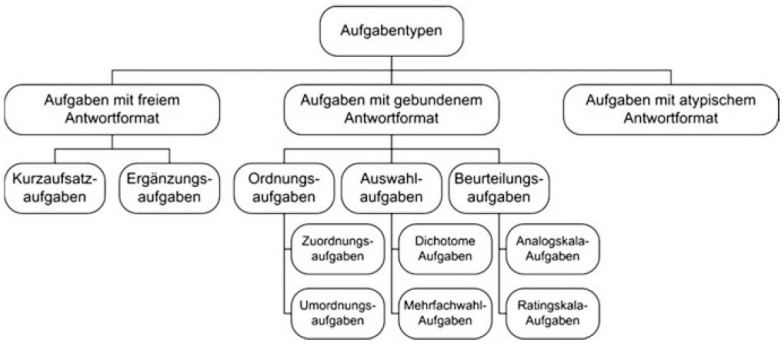


Abbildung 3.1: Aufgabentypen [MK20]

Proband	Item 1	Item 2	...	Item i	...	Item m	Zeilensumme
Proband 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,i}$	...	$x_{1,m}$	$\sum_{i=1}^{m} x_{1i} = x_1$
Proband 2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,i}$	...	$x_{2,m}$	$\sum_{i=1}^{m} x_{2i} = x_2$
...	...	...	...	...	...	...	...
Proband n	$x_{v,1}$	$x_{v,2}$	...	$x_{v,i}$	...	$x_{v,m}$	$\sum_{i=1}^{m} x_{vi} = x_v$
Spaltensummen	$\sum_{v=1}^{n} x_{v1}$	$\sum_{v=1}^{n} x_{v2}$	...	$\sum_{v=1}^{n} x_{vi}$	...	$\sum_{v=1}^{n} x_{vm}$	$\sum_{v=1}^{n} \sum_{i=1}^{n} x_{vi}$

Tabelle 3.1: Datenmatrix einer Stichprobe der Größe n über m Items

Ein Test besteht dabei aus mehreren Items. Items sind Bestandteile eines Tests, die eine Reaktion von Proband*innen erwarten. Die Items sollten dabei innerhalb des Tests homogen sein, da sie dasselbe Merkmal messen sollen [Ros96]. Ein Item kann dabei verschiedene Aufgabentypen annehmen. In Abbildung 3.1 sind die verschiedenen Formen von Aufgabentypen zu sehen.

Dabei spielt der Begriff der *Dichotomie* eine bedeutende Rolle in der Testtheorie. Dichotomie bezeichnet die Eigenschaft zwei annehmbarer Werte. Beispielsweise kann eine *dichotome Antwortvariable* die zwei Werte 0 (falsch) und 1 (richtig) annehmen [Ros96]. Dies ist beispielsweise bei Leistungstests der Fall, da hier eine Wissensabfrage stattfindet. Bei einem Persönlichkeitstest hingegen wird die Antwort nicht dichotom aufgefasst, da es kein *richtig* oder *falsch* gibt [MK20]. Ein Beispiel für eine nicht dichotome Antwortvariable wäre ein Multiple-Choice-Item, welches Teilpunkte nach der Menge der richtig beantworteten Felder vergibt. In dieser wird Arbeit jedoch von dichotomen Antwortvariablen ausgegangen, da dies andernfalls den Rahmen sprengen würde. Die Darstellung der Daten erfolgt dabei wie in Tabelle 3.1 zu sehen und bildet damit die Grundlage für die folgenden Methoden und Modelle.

3.2 Deskriptivstatistische Itemanalyse

Bei der deskriptivstatistischen Itemanalyse werden die erhobenen Daten einer deskriptivstatistischen Analyse unterzogen. Dabei spielt die erhobene Stichprobe eine große Rolle über die Aussagekraft der Daten. Dabei sind die Methoden der deskriptivstatistischen Itemanalyse stichprobenabhängig. In

den folgenden Abschnitten werden die einzelnen Methoden beschrieben und im letzten Abschnitt die Vor- und Nachteile der deskriptivstatistischen Itemanalyse aufgezeigt. Dabei gibt enthält Lienerts 'Testaufbau und Testanalyse' [Lie69] eine Zusammenstellung mehrere Verfahren, die mit den Verfahren aus Krauths 'Testkonstruktion und Testtheorie' [Kra95] und Moosbrugger& Kevalas 'Testtheorie und Fragebogenkonstuktion' [MK20] übereinstimmen. An den zu zitierenden Stellen wird daher auf Lienert [Lie69] Bezug genommen.

3.2.1 Schwierigkeitsindex

Der erste Schritt in der deskriptiven Itemanalyse besteht in der Berechnung der Schwierigkeiten der Items. Dabei nimmt die *Itemschwierigkeit* [Kra95] Werte von 0 bis 1 an. Diesen Wert wird mit 100 multipliziert[Lie69]. Daraus resultiert der *Schwierigkeitsindex* P_i für ein Item i.
Dieser lässt sich wie folgt berechnen:

$$P_i = \frac{\sum_{v=1}^{n} x_{vi}}{n * max(x_i)} * 100 \tag{3.1}$$

Ein Item ist somit leichter für hohe Werte und schwieriger für niedrige Werte.
Bei einem Leistungstest muss dabei zwischen zwei Testarten unterschieden werden, sogenannten Geschwindigkeits- und Niveautests [Lie69].

- Der Geschwindigkeitstest ist ein Leistungstest unter Zeitbegrenzung. Hierbei kann es dazu kommen, dass Fragen unbeantwortet oder unbearbeitet bleiben. Der *Schwierigkeitsindex* ist dann wie folgt definiert [Lie69]:

$$P_i = \frac{n_{richtig} + n_{falsch} + n_{ausgelassen}}{n_{beantwortet}} * 100 \tag{3.2}$$

- Der Niveautest hingegen gibt keine Zeitbegrenzung oder eine Zeitbegrenzung ohne Zeitdruck vor. Hierbei werden alle Fragen beantwortet. Der *Schwierigkeitsindex* ist dann wie folgt definiert [Lie69]:

$$P_i = \frac{n_{richtig}}{n} * 100 \tag{3.3}$$

3.2.2 Itemvarianz

Die *Itemvarianz* ist ein weiterer Teil der deskriptivstatistischen Itemanalyse. Sie beschreibt die Differenzierungsfähigkeit eines Items bezüglich der Stichprobe. Die *Itemvarianz* $Var(x_i)$ eines Items i wird nach [MK20] wie folgt definiert:

$$Var(x_i) = \frac{\sum_{v=1}^{n} (x_{vi} - \bar{x}_i)^2}{n} \tag{3.4}$$

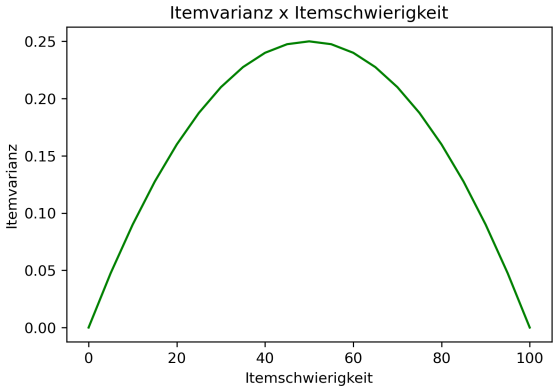


Abbildung 3.2: Itemschwierigkeit in Abhängigkeit der Itemvarianz

und lässt sich für dichotome Items wie folgt vereinfachen [Kra81] :

$$Var(x_i) = p_i * (1 - p_i) \tag{3.5}$$

Dabei lassen sich *Itemschwierigkeit* und *Itemvarianz* für dichotome Items in einen Zusammenhang bringen. Bei einer mittleren *Itemschwierigkeit* von 50 erreicht die *Itemvarianz* ihr Maximum von 0.25, die *Itemvarianz* fällt ab, umso schwieriger oder einfacher das Item wird. Abbildung 3.2 veranschaulicht dies.

3.2.3 Itemtrennschärfe

Die *Itemtrennschärfe* r_{it} gibt die Differenzierungsfähigkeit eines Items i bezüglich der Merkmalsausprägung einer Stichprobe an. Dabei wir die Korrelation r_{it} zwischen den Itemwerten x_{vi} der Stichprobe und den Testwerten x_v der Stichprobe ermittelt. Der Testwert x_v entspricht dabei der Zeilensumme x_{vi} eines/einer Proband*in v (siehe 3.1). Es gilt hierbei zu beachten, dass eine *part-whole-correction* ([McN62] nach [Lie69]) vorgenommen werden sollte, um die *Itemtrennschärfe* nicht zu überschätzen. Hierbei wird der Testwert ohne das zu korrelierende Item heranzgezogen. Daraus ergibt sich die *Itemtrennschärfe* $r_{it(i)}$ mit *part-whole-correction* nach Lienert [Lie69] wie folgt:

$$r_{it(i)} = r_{x_{vi}, x_{v(i)}} \tag{3.6}$$

Für dichtome Items kann die unkorrigierte *Itemtrennschärfe* wie folgt berechnet werden [Bor10]:

$$r_{it(i)} = \frac{\bar{x}_{v_1} - \bar{x}_{v_0}}{SD(x)} * \sqrt{\frac{n_0 n_1}{n(n-1)}} \tag{3.7}$$

mit $SD(x)$ als *Standardabweichung* der Testwerte und die *Mittelwerte* der Stichproben von x_{v_1} & x_{v_0}, die Item i richtig oder falsch gelöst haben.

Mit der Lösungswahrscheinlichkeit $p_i = \frac{P_i}{100}$ für Item i lässt sich aus Gleichung 3.7 die korrigierte *Itemtrennschärfe* wie folgt berechnen [Bor10]:

$$r_{it(i)} = \frac{r_{it}SD(x) - \sqrt{p_i(1-p_i)}}{\sqrt{SD(x)^2 + p_i(1-p_i) - 2r_{it}SD(x)\sqrt{p_i(1-p_i)}}} \tag{3.8}$$

Da die *Itemtrennschärfe* eine Korrelation ist, sind Werte zwischen -1 und 1 als Ergebnis möglich. Diese Werte können wie folgt interpretiert werden [MK20]:

r_{it} **nahe 1** Diese Items werden von Proband*innen mit hoher Merkmalsausprägung gelöst, von Proband*innen mit niedriger Merkmalsausprägung hingegen nicht. Hierbei gelten Trennschärfen mit einem Wert zwischen 0.4 und 0.7 als 'gut'.

r_{it} **nahe 0** Diese Items können nicht unterscheiden zwischen Proband*innen mit einer hohen und einer niedrigen Merkmalsausprägung. Das Item misst eine andere Merkmalsausprägung als die anderen Items.

r_{it} **nahe -1** Diese Items werden von Proband*innen mit niedriger Merkmalsausprägung gelöst, von Proband*innen mit hoher Merkmalsausprägung hingegen nicht. Das Item misst entweder eine andere Merkmalsausprägung als die restlichen Items oder wurde falsch formuliert.

3.2.4 Testwertermittlung und Testwertverteilung

Die einfachste Form der Testwertermittlung entspricht dem Wert x_v (vgl. Tabelle 3.1), also dem Wert der korrekt gelösten Aufgaben eines/einer Proband*in. Die andere Möglichkeit ist, die Punktevergabe der Lehrenden in die Teswertermittlung miteinzubeziehen. Die Testwertermittlung wird benötigt, damit die *Testwertverteilung* erstellt werden kann. Dabei kann die Testwertverteilung mithilfe von statistischen Maßen genauer untersucht werden.
Der *arithmetische Mittelwert* $\bar{x}$ ist wie folgt definiert:

$$\bar{x} = \frac{\sum_{v=1}^{n}\sum_{i=1}^{m} x_{vi}}{n} \tag{3.9}$$

Der *Median* entspricht dem Testwert, welcher die Stichprobenverteilung in zwei gleich große Hälften teilt [Lie69].
Der *Modalwert* beschreibt den häufigsten Testwert in der Testwertverteilung [Lie69].

Eine beispielhafte *Testwertverteilung* ist in Abbildung 3.3 zu sehen.

Die *Testwertvarianz* ist wie folgt definiert [Lie69]:

$$Var(x) = \frac{\sum_{v=1}^{n}(x_v - \bar{x})^2}{n-1} \tag{3.10}$$

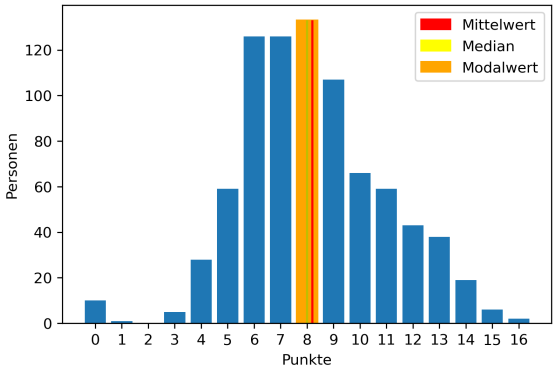


Abbildung 3.3: Beispielhafte Testwertverteilung

Daraus ergibt sich die *Standardabweichung* $SD(x)$ [Lie69] :

(3.11) $SD(x) = \sqrt{Var(x)}$

Die *Spannweite* entspricht [Lie69]:

(3.12) $Spannweite = x_{max} - x_{min}$

Der *Interquartilsabstand* $IQR(x)$ ist die Differenz zwischen dem oberen Quartil und dem unteren Quartil der *Testwertverteilung* [Lie69].

Sollte die *Testwertverteilung* keiner Normalverteilung entsprechen, so kann man *Schiefe* und *Exzess* berechnen. Diese werden wie folgt berechnet [Lie69]:

(3.13) $Schiefe = \frac{\sum_{v=1}^{n}(x_v - \bar{x})^3}{n * SD(x)^3}$

(3.14) $Exzess = \frac{\sum_{v=1}^{n}(x_v - \bar{x})^4}{n * SD(x)^4} - 3$

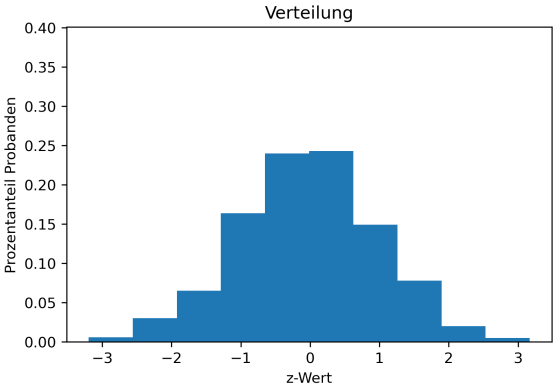


Abbildung 3.4: Beispielhafte z-Wertverteilung

Falls keine Normalverteilung vorliegt, gibt es verschiedene Methoden, um die *Testwertverteilung* dieser anzupassen. Dies kann sinnvoll sein, um verschiedene Stichproben miteinander zu vergleichen. Bei einer rechtsschiefen Verteilung ist es möglich die Testwerte zu *logarithmieren*, um so eine Normalverteilung darzustellen. Bei linksschiefen Verteilungen muss zuvor eine Spiegelung der Verteilung vorgenommen werden. Die gespiegelten Testwerte können nun *logharitmiert* und zurückgespiegelt werden. So kann auch hier eine Normalverteilung erreicht werden. Eine andere Möglichkeit besteht darin, die Skala anzupassen. Eine gängige Methode hierbei ist die *z-Standardisierung*. Hierbei beschreiben die *z-Werte* die Entfernung des Testwertes eines/einer Proband*in zum Mittelwert. Beispielsweise bedeutet ein *z-Wert* von drei, dass der/die Proband*in mit seinem/ihrem Testwert drei Standardabweichungen vom Mittelwert entfernt ist. Dabei lässt sich der *z-Wert* z_v eines/einer Proband*in v wie folgt berechnen [Lie69]:

$$z_v = \frac{x_i - \bar{x}}{SD(x)} \tag{3.15}$$

Diese Werte können nun wie in Abbildung 3.4 zu sehen dargestellt werden.

Dies sind die wichtigsten Verfahren der deskriptivstatistischen Itemanalyse. Items können nun mithilfe der *Itemschwierigkeit* , der *Itemvarianz* und der *Itemtrennschärfe* zu einem Test zusammengefasst werden. Dabei hilft die *Testwertverteilung* bei der Analyse einer Stichprobe.

3.2.5 Vor- und Nachteile

Die Vorteile der deskriptivstatistischen Itemanalyse liegen bei der Einfachheit der Umsetzung und der Verständlichkeit der daraus resultierenden Daten. Da ‘klassische’ Methoden der Statisitik verwendet werden, ist die Lesbarkeit der Daten für Lehrende einfach. Der Nachteil ist die Stichprobenabhängikeit der Methoden [Ros96]. Es kann also nie auf die Grundgesamtheit geschlossen werden, sondern nur auf die vorhandene Stichprobe.

3.3 Klassische Testtheorie

Die Klassische Testtheorie (KKT) kann im Wesentlichen als Messfehler-Theorie angesehen werden und wurde von Gulliksen [Gul50](zitiert nach [MK20]) und Lord & Novick [LBN68] begründet. Dabei ist die Annahme der KTT, dass der Testwert eines/einer Proband*in sich aus zwei Werten zusammensetzt. Der Messwert besteht dabei aus der wahren Merkmalsauspräung τ und einem zufälligen Messfehler ε, jedoch lässt sich nur die Messung x beobachten. Aus dieser Annahme lassen sich die vier Axiome ableiten, aus welchen dann die zwei wichtigsten Verfahren der KTT resultieren. Die *Reliabilität*, die beschreibt, wie messgenau ein Fragebogen misst und die Bestimmung von Konfidenzintervallen für einzelne Proband*innen [LBN68]. Da sich diese Arbeit auf die Entwicklung einer Anwendung für Lehrende bezieht, wird in diesem Abschnitt nicht näher auf Konfidenzintervalle eingegangen.

3.3.1 Axiome

Die verschiedenen Axiome lassen sich aus den Grundannahmen der KTT ableiten. Sie sind wie folgt definiert [LBN68]:

1. Existenzaxiom Der wahre Wert τ_{vi} eines/einer Proband*in v in einem Item i liegt im Erwartungswert der Messung x_{vi}.

$$\tau_{vi} = E(x_{vi}) \tag{3.16}$$

2. Verknüpfungsaxiom Eine Messung x_{vi} setzt sich aus dem wahren Wert τ_{vi} und dem Messfehler ε_{vi} zusammen.

$$x_{vi} = \tau_{vi} + \varepsilon_{vi} \tag{3.17}$$

3. Unabhängigkeitsaxiom Die Korrelation zwischen wahrem Wert τ_{vi} und Messfehler ε_{vi} ist null.

$$Corr(\tau_{vi}, \varepsilon_{vi}) = 0 \tag{3.18}$$

4. Zusatzannahmen Die Fehlerwerte zweier beliebiger Items i, j sind bei demselben/derselben Proband*in v unkorreliert.

$$Corr(\varepsilon_{vi}, \varepsilon_{vj}) = 0 \tag{3.19}$$

Die Fehlerwerte zweier beliebiger Proband*innen v, w sind bei demselben Item i unkorreliert.

$$Corr(\tau_{vi}, \tau_{wi}) = 0 \tag{3.20}$$

3.3.2 Reliabilität

Mit diesen Annahmen kann nun die wahre Varianz und die Fehlervarianz der Testwerte bestimmt werden und im weiteren Schritt die Reliabilität eines Test. Die Reliabilität beschreibt die Messgenauigkeit eines Tests.

Dabei ist die Varianz des wahren Wertes τ und zweier Testwerte x_p, x_q wie folgt definiert [LBN68]:

$$Var(\tau) = Cov(\tau_p, \tau_q) = Cov(x_p, x_q) \tag{3.21}$$

Die *Reliabilität* lässt sich dann wie folgt berechnen [LBN68]:

$$Rel = \frac{Var(\tau)}{Var(x)} = \frac{Cov(x_p, x_q)}{SD(x_p) * SD(x_q)} \tag{3.22}$$

Dabei nimmt die *Reliabilität* Werte im Bereich zwischen 0 und 1 an. Je höher der Wert, umso besser ist die Reliabilität.

Zur Schätzung der *Reliabilität* gibt es laut Lienert [Lie69], Krauth [Kra95] und Moosbrugger Kelava [MK20] in der Praxis vier Methoden, welche sich als Grundlagen in 'Statistical Theories of Mental Test Scores' [LBN68] wiederfinden.

Paralleltest-Reliabilität Für die Bestimmung der Paralleltest-Reliabilität sind zwei Testhälften nötig. Diese Testhälften müssen in Parallelform vorliegen. Das bedeutet, dass die beiden Testhälften dasselbe Merkmal mit derselben Genauigkeit erfassen.

Retest-Reliabilität Zu der Bestimmung der Retest-Reliabilität wird das Testverfahren an derselben Stichprobe zweimal durchgeführt und mit diesen beiden Testhälften wird dann die Reliabilität bestimmt.

Split-Half-Reliabilität Bei dieser Methode wird ein Test in zwei Testhälften geteilt. Diese Testhälften sollten möglichst parallel sein. Die Testhälften sollten – wenn möglich – dasselbe Merkmal messen und dieselbe Genauigkeit erfassen.

Interne Konsistenz Diese Methode der Reliabilitätsbestimmung ist eine Weiterführung der Split-Half-Methode. Hierbei werden die einzelnen Items als seperate Testteile aufgefasst. Dadurch kann durch die Zusammenhangsstruktur der Items auf die interne Konsistenz als Form der Reliabilität geschlossen werden.

Die Methoden der Paralleltest- und Retest-Reliabilität sind für diese Arbeit ungeeignet, da sie mehr als einen Test und eine Stichprobe zur Bestimmung der Reliabilität benötigen. Diese werden daher nicht näher erläutert. Die Bestimmung der Reliabiltät durch die Split-Half-Methode und durch die Interne Konsistenz eignen sich dagegen. Diese werden im Folgenden näher erläutert.

Für die Bestimmung der Reliabilität wird eine Formel benötigt. Die Testverlängerung wird dabei für die Methode der Split-Half-Reliabilität benötigt. Durch die Testverlängerung kann bestimmt werden, wie die Reliabilität gesteigert werden kann, wenn Tests in Parallelform hinzugenommen

werden.
Dabei ist die Testverlängerung durch die Spearman-Brown-Formel nach Lord & Novick [LBN68] und Lienert [Lie69] wie folgt definiert:

$$Rel(k,l) = \frac{k * Rel}{1 + (k-1) * Rel} \tag{3.23}$$

wobei k der Faktor der Testverlängerung und l die Länge des Tests bezeichnen.

Split-Half-Reliabilität

Bei dieser Methode wird ein Test x in die zwei Testhälften x_a und x_b halbiert. Diese sollten im besten Falle eine parallele Form vorweisen. Durch die Halbierung der Testhälften muss die Spearman-Brown-Formel angewandt werden, um die Reliabilität für den gesamten Test zu bestimmen. Die Formel der Reliabilität lautet dann wie folgt [LBN68]:

$$Rel(x) = \frac{2 * Corr(x_a, x_b)}{1 + Corr(x_a, x_b)} = \frac{2 * Rel(x_a)}{1 + Rel(x_a)} \tag{3.24}$$

Für die Methode der Testhalbierung kommen nach Lienert [Lie69] zwei Möglichkeiten in Frage.

Odd-Even Bei dieser Methode werden die Items abwechselnd den einzelnen Testhälften zugewiesen.

Item-Zwillinge Hier werden die Items mithilfe der deskriptivstatistischen Itemanalyse in die zwei Testhälften geteilt. Dabei werden Paare mit möglichst gleicher Schwierigkeit und Trennschärfe gebildet. Diese Paare werden dann in die jeweiligen Testhälten aufgeteilt.

Der Nachteil dieser Methode liegt in der Einteilung der beiden Testhälften. Es kann nie zu hundert Prozent sichergegangen werden, dass diese parallel sind.

Interne Konsistenz

Zur Bestimmung der internen Konsistenz wird meist Cronbachs α genutzt [LBN68] [Lie69]. Sie gilt nach Moosbrugger und Kelava [MK20] als heutiger Standard. Cronbachs α ist nach Lord & Novick [LBN68] wie folgt definiert:

$$Rel(x) = \alpha = \frac{m}{m-1} * (1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)}) \tag{3.25}$$

mit m als Anzahl der Items, $Var(x)$ als Varianz des gesamten Tests und $Var(x_i)$ als Varianz des i-ten Items.

Die Vorraussetzung für die Berechnung von Cronbachs α ist die Parallelform der Items. Dies bedeutet, alle Items sollten eine ähnliche Schwierigkeit aufweisen. Cronbachs α bietet hierbei den Vorteil, dass immer mindestens die untere Schranke für die Reliabilität berechnet wird [MK20] .

3.3.3 Vor- und Nachteile

Ein Nachteil in der KTT liegt in der Stichprobenabhängigkeit der genutzen Kenngrößen, da diese aus der deskriptivstatistischen Itemanalyse stammen. Dies stellt jedoch kein Problem dar, solange die Stichprobe groß genug ist. Auch die Annahmen der KTT, dass ein *wahrer Wert* und ein *Messfehler* existieren, sind nicht überprüfbar. Diese Werte können nicht direkt beobachtet werden. Dennoch wird die KTT schon lange in der Praxis angewandt und bietet den Vorteil der einfachen Umsetzbarkeit [MK20]. Ein weiterer Nachteil ist die Auslegung der *Reliabilität*. Da die Grundlage der Methoden nicht immer gegeben ist, ist auch die Berechenbarkeit der *Reliabilität* nicht immer gegeben. Dabei muss eine niedrige *Reliabilität* jedoch nicht bedeuten, dass der Test nicht reliabel ist, kann aber ein Hinweis darauf sein. Auch kann laut Hartig & Frey [HFJ] mithilfe der KTT auf die *Kriteriumsvalidität* geschlossen werden, jedoch nicht auf die *Konstruktvalidität*. Folglich kann mithilfe der KTT gezeigt werden, dass ein Test geeignet ist, um Merkmale zu testen, die auch außerhalb einer Testsituation verlangt werden. Durch den Test kann jedoch nicht direkt auf die Merkmalsausprägung einer Person geschlossen werden [HFJ].

3.4 Item-Response-Theorie

Der Item-Response-Theorie (IRT) liegt die Annahme zugrunde, dass aus den Antworten der Proband*innen Rückschlüsse auf deren Merkmalsausprägung geschlossen werden können. Die grundlegenden Ideen und Modelle der IRT wurden von Rasch [Ras60] (nach [SEM93]) und Birnbaum [Bir68] vorgestellt. Dabei wird in der IRT zwischen zwei Variablen unterschieden. *Manifeste Variablen* bezeichnen dabei das beobachtbare Antwortverhalten der Proband*innen, wohingegen *latente Variablen* die unbeobachtbare Merkmalsausprägung ξ bezeichnen. Die manifeste Variable ist dabei abhängig von der latenten Variable. Das Antwortverhalten eines/einer Proband*in ist also abhänging von seiner/ihrer Merkmalsausprägung. Dies bedeutet, Proband*innen hoher Merkmalsausprägung werden viele Items korrekt lösen, Proband*innen mit niedriger Merkmalsausprägung hingegen werden nur wenige Items korrekt lösen [Ras60] (nach [SEM93]) [Bir68]. Die Schätzung der Variablen wird mithilfe eines Maximum-Likelihood-Schätzers vorgenommen, welcher in Unterabschnitt 3.4.4 näher vorgestellt wird. Diese wird mithilfe eines Modellkonformitätstests auf die Modellpassung überprüft, welche in Unterabschnitt 3.4.6 vorgestellt wird. In Unterabschnitt 3.4.5 wird die Reliabilitätsschätzung durch die IRT vorgestellt.

In der IRT kann man zwischen zwei Modellannahmen unterscheiden, den Latent-Class-Modellen und den Latent-Trait-Modellen. Für diese Arbeit sind Latent-Class-Modelle eher ungeeignet, da sie laut Moosbrugger und Kelava [MK20] Aufschluss über die Unterscheidung verschiedener Personengruppen geben Im Rahmen dieser Arbeit werden zwei Latent-Trait-Modelle betrachtet, da diese zurzeit in der psychologischen Diagnostik die meiste Anwendung finden [MK20]. Dabei wird auf das Rasch-Modell [Ras60] (nach [Ros96] und[SEM93]) in Unterabschnitt 3.4.2 und das Birnbaum-Modell [Bir68] in Unterabschnitt 3.4.3 eingegangen.
Zunächst wird jedoch in Unterabschnitt 3.4.1 auf die Voraussetzungen eingegangen, die für alle IRT-Modelle gelten müssen.
Auch in der IRT dient die Datenstruktur aus Tabelle 3.1 als Grundlage für die Berechnungen.

3.4.1 Itemhomogenität

In der IRT gilt das Vorliegen der *Itemhomogenität* bezüglich ξ als hinreichende Bedingung. Das Antwortverhalten der Proband*innen darf also nur von ξ abhängen und von keinem anderen Einflüssen. Damit die Itemhomogenität erfüllt ist, müssen die manifesten Variablen die Bedingung der *lokalen stochastischen Unabhängigkeit* erfüllen. Damit diese Bedingung erfüllt ist, muss folgendes gelten: Die Antwortvariablen von Proband*innen mit der gleichen Merkmalsausprägung dürfen nicht miteinander korrelieren. Dies bedeutet, dass alle Proband*innen mit der gleichen Merkmalsausprägung ein Item immer gleich beantworten. Die Grundvoraussetzung zur Anwendung der IRT auf einen Test ist also die Unabhängigkeit der Items untereinander. Es darf keine Items geben, welche nur zu lösen sind, wenn ein anderes vorangestelltes Item richtig gelöst wurde [Ras60] (nach [SEM93]) [Bir68].
Ist dies nicht der Fall, kann zum Beispiel davon ausgegangen werden, dass das Item noch ein anderes Merkmal misst. Es können also von der manifesten Variable keine Rückschlüsse auf die latente Variable gezogen werden. Dies verletzt jedoch die Grundannahme der IRT und ist daher nicht zulässig.
Da die Original-Literatur zu Raschs 'Probabilistic models for some intelligence and attainment tests.' [Ras60] nicht zur Verfügung stand, die Methoden jedoch in Steyer et al. 'Messen und Testen' [SEM93], Rosts 'Testtheorie und Testkonstruktion' [Ros96] und Trevor et al. 'Applying The Rasch Model' [BYH20] übereinstimmen, wird von der Richtigkeit des vorgestellten Modells ausgegangen.

3.4.2 Rasch-Modell

Das Rasch-Modell ist das einfachste Modell der Latent-Trait-Modelle und wird auch als 1PL-Modell bezeichnet [Ras60] (nach [SEM93]) [MK20]. Es nimmt für alle Items die gleiche logistische itemcharackteristische Funktion (IC-Funktion) an. Dabei beschreibt das Modell die Antwortwahrscheinlichkeit $P(x_i = 1)$ für ein Item i in Abhängigkeit des Itemschwierigkeitsparameter σ (Itemparameter). Die Itemschwierigkeit ist dabei der Wert der für $P(x_i = 1) = 0.5$ gilt. Somit können σ und ξ auf derselben Achse abgebildet werden. Dabei gilt zu beachten, dass σ sich deutlich von der Itemschwierigkeit der deskritptivstatistischen Itemanalyse unterscheidet.
Die Modellgleichung besitzt dabei die folgende Modellgleichung [Ras60] (nach [SEM93]):

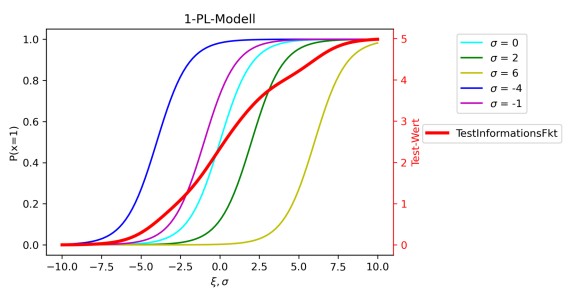
$$P(x_{vi}) = \frac{exp(x_{vi}(\xi_v - \sigma_i))}{1 + exp(\xi_v - \sigma_i)} \tag{3.26}$$

Für richtig gelöste Aufgaben gilt daher [Ras60] (nach[SEM93]):

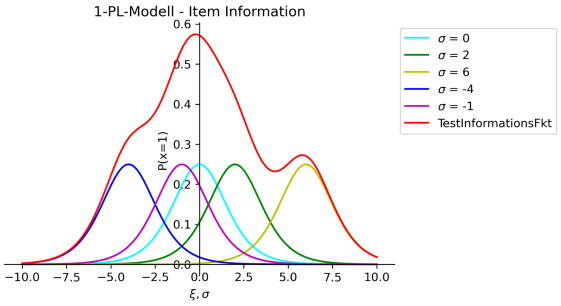
$$P(x_{vi} = 1) = \frac{exp(\xi_v - \sigma_i)}{1 + exp(\xi_v - \sigma_i)} \tag{3.27}$$

Für falsch gelöste Aufgaben gilt [Ras60] (nach [SEM93]):

$$P(x_{vi} = 0) = \frac{1}{1 + exp(\xi_v - \sigma_i)} \tag{3.28}$$



(a) Logistische-IC-Funktionen mit Test-Antwortfunktion



(b) Item-Informationsfunktionen und Test-Informationsfunktion

Abbildung 3.5: 1PL-Modell

Mithilfe dieser Gleichungen kann weiterführend die *Test-Antwortfunktion* erstellt werden. Für diese Funktion werden lediglich die schon berechneten Itemfunktionen miteinander aufsummiert [Ras60] (nach [BYH20]):

$$I_v = \sum_{i=1}^{m} P(x_{vi} = 1) \tag{3.29}$$

Dabei beschreibt die Funktion die Anzahl der korrekt beantworteten Items in Abhängigkeit der Merkmalsausprägung.
In Abbildung 3.5a sind beispielthaft drei Items mit verschiedenen Schwierigkeitsgraden sowie die *Test-Antwortfunktion* zu sehen.
Zur Veranschaulichung: Ein Item mit $\sigma_i = 0.5$ wird von einem/einer Proband*in mit $\xi = 0.5$ in 50% der Fälle gelöst. Je höher die Merkmalsausprägung ist, umso wahrscheinlicher kann das Item von dem/der Proband*in gelöst werden. Je niedriger die Merkmalsausprägung ist, umso unwahrscheinlicher kann das Item von dem/der Proband*in gelöst werden. Ein/Eine Proband*in mit einer Merkmalsausprägung von -2.5 würde den Test in Abbildung 3.5a im Schnitt mit einem Punkt abschließen.

Für das 1PL-Modell kann weiterführend die *Item-Informationsfunktion I* bestimmt werden und daraus resultierend die *Test-Informationsfunktion*. Dabei ist die *Item-Informationsfunktion* wie folgt definiert [Ras60] (nach [BYH20] :

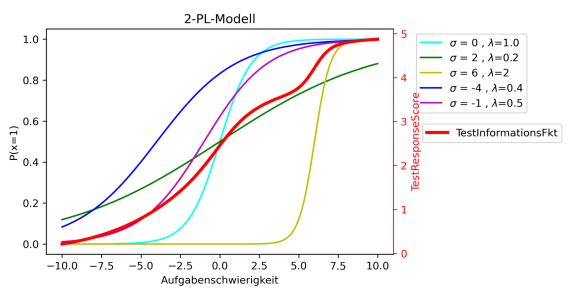
$$I_{iv} = P(x_{iv} = 1) * P(x_{iv} = 0) = \frac{exp(\xi_v - \sigma_i)}{(1 + exp(\xi_v - \sigma_i))^2} \tag{3.30}$$

Sie beschreibt die Aussagekraft eines Items in Abhängigkeit zur Merkmalsausprägung.

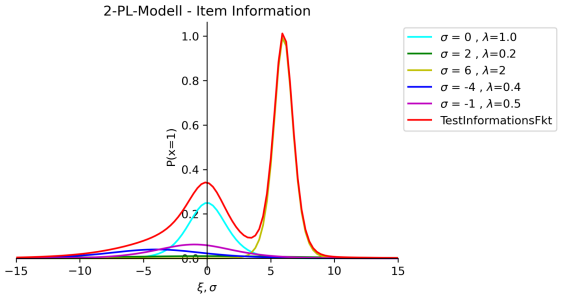
Die *Test-Informationsfunktion* hingegen beschreibt die Aussagekraft eines Tests in Abhängigkeit zur Merkmalsausprägung. Sie wird gebildet, indem die *Item-Informationsfunktionen* aufsummiert werden [Ras60] (nach [BYH20]):

$$I_v = \sum_{i=1}^{m} I_{iv} \tag{3.31}$$

In Abbildung 3.5b ist ein Beispiel dieser beiden Funktionen zu sehen.



(a) Logistische-IC-Funktionen mit Test-Antwortfunktion



(b) Item-Informationsfunktionen und Test-Informationsfunktion

Abbildung 3.6: 2PL-Modell

3.4.3 Birnbaum-Modell

Birnbaum [Bir68] formuliert zwei verschiedene Modelle . Das 2PL-Modell enthält zusätzlich zum Personenparameter ξ und Itemparameter σ den Diskriminationsparameter λ. Dieser ist vergleichbar mit der Itemtrennschärfe der deskriptivstatistischen Itemanalyse. Im 3PL-Modell kommt zusätzlich noch der Rateparameter ρ hinzu, welcher bei Multiple-Choice-Tests dem Umstand des Ratens gerecht wird [Bir68].

Die logistische IC-Funktion des 2PL-Modell ist dabei wie folgt definiert [Bir68]:

$$P(x_{vi} = 1) = \frac{exp(\lambda_i(\xi_v - \sigma_i))}{1 + exp(\lambda_i(\xi_i - \sigma_i))} \tag{3.32}$$

In Abbildung 3.6a sind beispielhaft 2PL-IC-Funktionen zu sehen.

Und die logistische IC-Funktion des 3PL-Modells ist wie folgt definiert [Bir68]:

$$P(x_{vi} = 1) = \rho_i + (1 - \rho_i)\frac{exp(\lambda_i(\xi_v - \sigma_i))}{1 + exp(\lambda_i(\xi_i - \sigma_i))} \tag{3.33}$$

In Abbildung 3.7 sind beispielhaft 3PL-IC-Funktionen zu sehen.

Im 2PL-Modell und 3PL-Modell können die *Test-Antwortfunktion* (3.29), die *Item-Informationsfunktion* (3.30) und die *Test-Informationsfunktion* (3.31) analog zum 1PL-Modell berechnet werden [Bir68]. Diese werden für das 2PL-Modell auch in Abbildung 3.6b dargestellt.

3.4.4 Maximum-Likelihood (ML)

Für die Schätzung der Parameter gibt es verschiedene ML-Methoden. In dieser Arbeit wird der Marginal-Maximum-Likelihood-Schätzter (MML) genutzt. Dieser hat den Vorteil, dass er für kleine Stichproben schon gute Ergebnisse liefert und stabil läuft [DL70]. Ein weiterer Vorteil des MML ist die Schätzung der Parameter für die verschiedenen Modelle.

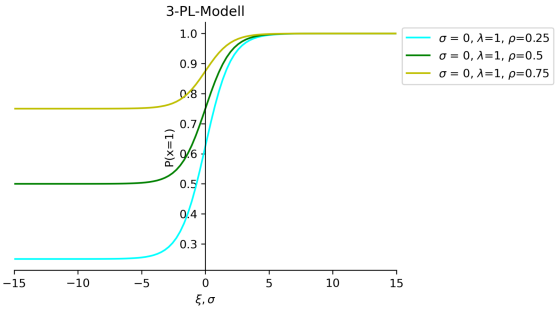


Abbildung 3.7: Logistische-IC-Funktionen

Es reicht einen Schätzer für die vorgeschlagenen Modelle aus. Für diesen wird im 1PL-Modell $\lambda = 1$ und $\rho = \frac{1}{n}$ und im 2PL-Modell $\rho = \frac{1}{n}$ gesetzt, wobei n die Anzahl der Antwortmöglichkeiten für die Items in einem Test beschreibt.
Außerdem schätzt der MML zunächst nur die Itemparameter und keine Personenparameter. Dies resultiert in einer zusätzlichen Zeitersparnis, da die Personenparameter nicht geschätzt werden. Da diese Arbeit sich jedoch mit einer Entwicklung einer Anwendung für Lehrende beschäftigt, sind diese Parameter zu vernachlässigen.
Dabei ist der MML-Schätzer nach Bock & Lieberman [DL70] wie folgt definiert:

$$mL(\iota) = \prod_{v=1}^{n} P(Y_v = y_v; \iota) = \prod_{v=1}^{n} \int_{\theta} P(Y_v = y_v|\theta; \iota) f(\theta) d\theta = P(Y; \iota) \tag{3.34}$$

In Bock & Aitkin [BA81] wird ein Erwartungs-Maximierungs-Algorithmus (EM-Algorithmus) beschrieben, welcher das Problem auch für eine große Zahl an Items berechenbarer macht. Dabei kann die Idee des EM-Algorithmus für den MML in zwei Schritten umschrieben werden. Im E-Schritt wird der Erwartungswert des marginal Log-Likelihood berechnet. Im M-Schritt wird dieser Erwartungswert, bezogen auf die Zielfunktion, maximiert. Diese Schritte werden so oft ausgeführt, bis sie ein vorher festgelegtes Kriterium erreichen. Diesem EM-Algorithmus liegen heutzutage die meisten Implementierungen des MML-Verfahrens zugrunde [MK20].

Es wird beim MML von einer Normalverteilung ausgegangen, doch auch bei einer Verletzung der Verteilung bleiben die Schätzwerte fast identisch zu der konkurrierenden Conditional-Maximum-Likelihood (CML) [Ros96]. Der CML liegt dabei ein anderes Verfahren zu Grunde, welches jedoch nicht von der Normalverteilung ausgeht.

3.4.5 Reliabilität

Anschließend an die Schätzung der Itemparameter kann eine Reliabilitätsschätzung vorgenommen werden. Diese ist für Modelle der IRT genauso wie in der KTT definiert, also als Verhältnis der wahren Varianz und der Varianz der Testwerte [And88] (nach [MK20]) .

Mithilfe einer *Expected A Posteriori* (EAP) - Schätzung kann der bedingte Erwartungswert für eine latente Personenvariable geschätzt werden. Eine Zufallsziehung der latenten Personenvariablen aus der *a posteriori* Verteilung bezeichnet dabei sogenannte *Plausible Values* (PV). [MK20]. Damit kann nun die sogenannte Reliabilität geschätzt werden [AWW97]:

$$Rel(\hat{\theta}) = \frac{Var[E(\theta_v|Y_v = y_v)]}{Var(\theta)} \tag{3.35}$$

Da im MML-Schätzer eine Normalverteilung angenommen wird, werden fixierte Werten für die Varianz und den Mittelwert der Verteilung genutzt. Daraus resultiert:

1. die Gleichheit der geschätzten EAP-Reliabilität und der Varianz des EAP-Schätzers.
2. die Varianz der PVs ist gleich die wahre Varianz der latenten Variblen.

Beide dieser Werte können anschließend als Nenner verwendet werden. Für die Anwendung wird die EAP/PV- Reliabilität genutzt, weshalb letzteres als Nenner verwendet wird [MK20].

Das Ergebnis der Reliabilitätschätzung kann laut Rost [Ros96] mit Cronbachs α (3.25) verglichen werden.

3.4.6 Modellkonformität

Trotz einer guten Schätzung der Variablen ist die Modellkonformität nicht zwangsläufig gegeben. Diese ist beim 1PL-Modell strenger und erfordert zusätzlich die sogenannte Rasch-Homogenität. Daher sollte überprüft werden, ob das berechnete Modell gut auf die Daten der Stichprobe passt. Dazu gibt es mehrere Methoden. Für diese Arbeit wird die Methode der *standardized root mean square root of squared residuals* (SRMSR) nach Maydeu-Olivaras [May13] gewählt. Diese Methode ist leicht zu interpretieren, da hier einen Zahlenwert zurückgeben wird. Außerdem ist sie schon für kleine Stichprobengrößen geeignet und liefert dort gute Ergebnisse [May13]. Sie ist nach Maydeu-Olivaras [May13] mit der Korrelation $r_{i,j}$ und der erwarteten Korrelation $p_{i,j}$ für ein Paar von Items i und j wie folgt definiert:

$$SRMSR = \sqrt{\sum_{i<j} \frac{(r_{ij} - p_{ij})^2}{n(n-1)/2}} \tag{3.36}$$

Sollte das Ergebnis ≤ 0.05 betragen, hat das berechnete Modell eine gute Passung. Bis 0.1 ist die Modellpassung in Ordnung, sollte aber mit dem Wissen des Wertes interpretiert werden. Jeder Wert > 0.1 lässt auf ein schlechtes Modell schließen und ist daher nicht geeignet [May13].
Um die Modellkonformität zu verbessern, können Items, die stark abweichen, aussortiert werden. Anschließend kann noch einmal die Modellkonformität überprüft werden [Ros96] [MK20].

3.4.7 Vor- und Nachteile

Der Vorteil bei den vorgeschlagenen Methoden ist de Stichprobenunabhängigkeit der Ergebnisse [Ros96]. Diese kommt durch die gewählte ML-Methode zustande, da diese nur Itemparameter schätzt. Durch die Angabe der Modellkonformität ist zusätzlicher Interpretationsspielraum gegeben. Dieser kann sehr hilfreich beim Einschätzen der Modelle sein, ist jedoch auch ein weiterer Wert, den man interpretieren muss. Die IRT bietet eine bessere Analyse einzelner Items als die deskriptivstatistische Itemanalyse. Jedoch basieren die IRT-Modelle auf ungewohnteren Methoden und sind daher für den Großteil der Lehrenden schwieriger zu lesen und zu interpretieren. Auch die Berechnung der Reliabilität durch die IRT ist gut, da es dadurch einen Vergleichswert zur Reliabilität der KTT gibt und so die berechneten Reliabilitäten besser eingeordnet werden können. Jedoch ist das EAP/PV Schätzverfahren nur bedingt geeignet, da bei der EAP Schätzung die Varianz oft unterschätzt wird, was dazu führen kann, dass die Reliabilität auch unterschätzt wird [MK20].

4 Konzept

In diesem Kapitel wird das Konzept der Anwendung besprochen. Dabei soll ergründet werden, welche der vorgestellten Methoden aus Kapitel 3 in die finale Anwendung kommen. Da diese nach Anforderung **A3** (2.2) nur die nötigsten Information enthalten soll, werden nicht alle der vorgestellten Methoden eingebunden. Des Weiteren werden mögliche Arten der Visualisierung besprochen und aufgezeigt, welche Art der Visualisierung sich für die verschiedenen Methoden eignet. Am Ende dieses Kapitel wird der für diese Arbeit erstellte Fragebogen vorgestellt. Dieser soll erörtern, welche Kennwerte, Modelle und deren Arten der Visualisierung für die Lehrenden von Wichtigkeit sind.
Des Weiteren soll für die Lehrenden eine Hilfestellung innerhalb der Anwendung geboten werden. Diese soll dabei helfen, die dargestellten Modelle und Daten besser zu interpretieren.
In dieser Arbeit wird für alle Aufgaben von einer *dichotomen Antwortvariablen* ausgegangen. Dies hat den Vorteil der besseren Interpretierbarkeit für die Lehrenden. Auch würden insbesondere die Modelle der IRT erheblich unzugänglicher machen, die Aussagekraft jedoch nur minimal steigern.

4.1 Umsetzung deskriptivstatistische Itemanalyse

Da die Anwendung nach **A2** 2.2 benutzerfreundlich gestaltet werden soll, wird dort der Niveautest als Methode implementiert. Dies wird damit begründet, da die Anwendung für jede Lehrperson gleich sein soll. Es gibt Lehrende, die als Mittel der Testkonstruktion Zeitdruck verwenden. Daher muss sich die Anwendung für die Berechnung des Schwierigkeitsindex nach dem Niveautest orientieren, da sonst unbeantwortete Fragen möglicherweise zu leicht eingestuft werden. Bei der Berechnung des Schwierigkeitsindex werden daher falsch beantwortete Fragen als unbeantwortete Fragen interpretiert. Ebenfalls wird eine Ratekorrektur der Antworten ausgeschlossen. Falsche Antworten würden hierbei als geraten interpretiert werden und müssten in die Formel für P_i (3.1) miteinfließen. Allerdings werden für die Berechnung Metadaten der Items benötigt, die EvaExam nicht liefert. Des Weiteren spricht die Erläuterung und deren Auswahl der beiden Testformen gegen **A2** (2.2). Daher wird eine Befragung der Lehrenden in der Anwendung ausgeschlossen. Für die Berechnung der *Itemschwierigkeit* wird nach dem Niveautest vorgegangen (3.3) Da die *Itemschwierigkeit* eine der wichtigsten Kennzahlen der deskriptivstatistischen Itemanalyse ist, wird diese Kennzahl in dem Fragebogen (7) zur Diskussion gestellt.
Des Weiteren wird die *Itemvarianz* (3.2.2), die *Itemtrennschärfe* (3.2.3) und die Verteilungen aus (3.2.4) im Fragebogen verhandelt.

4.1.1 Visualisierung

Für die Visualisierung der deskriptivstatistischen Itemanalyse werden in dieser Arbeit “klassische Methoden” verwendet. Die Auswahl der gewählten Visualisierung basiert größtenteils auf den vorgestellten Methoden aus ‘Deskriptive Statistik verstehen’ von Schendera [Sch15], welches sich an ‘Visual and statistical thinking : displays of evidence for making decisions’ von Tufte [Tuf97] als Grundlage bedient. Darin werden Methoden der Visualisierung der deskriptiven Statistik beschrieben und Vor -und Nachteile der einzelnen Methoden aufgelistet. Auch sind einige eigene Ideen zur Visualisierung dabei entstanden, die ebenfalls in dem Fragebogen (7) Platz gefunden haben.

4.2 Umsetzung Klassische Testtheorie

In der KTT spielt in dieser Arbeit nur die Berechnung der Reliabilität eine Rolle. Es soll abgefragt werden, ob diese für die Lehrenden eine bedeutsamen Kennwert darstellt. Es gilt dabei zu beachten, diesen für die Lehrenden richtig einzuordnen, da sonst Fehlschlüsse aus diesem Wert gezogen werden können. Dies soll mithilfe eines Hilfstextes umgesetzt werden. Deshalb wird diesem der Reliabilitätswert aus der IRT entgegengestellt, was den Interpretationsspielraum einschränken soll. Eine umfangreiche Visualisierung der Werte ist deshalb weniger sinnvoll.

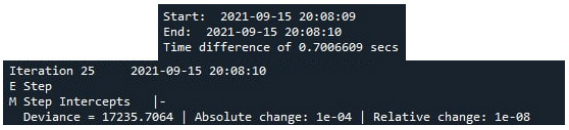
4.3 Umsetzung Item-Response-Theorie

Für die Umsetzung der IRT beschränkt sich die Arbeit auf das 1PL- und das 2PL- Modell. Der Grund dafür liegt einerseits in der Berechnungsdauer durch den EM-Algorithmus aus Unterabschnitt 3.4.4. Dieser benötigt deutlich mehr Schritte für das 3PL- Modell. Dadurch muss der Lehrende zu lange auf die Ergebnisse warten, was wegen Anforderung **A2** (2.2) jedoch ausgeschlossen wird. Das 1PL- und 2PL-Modell hingegen sind in kurzer Zeit berechnet, zu sehen am Beispiel in Abbildung 4.1. Des Weiteren ist die Interpretation eines 3PL-Modells schwerer als die der beiden anderen Modelle. Diese beiden Faktoren führen dazu, dass das 3PL-Modell in der Anwendung nicht dargestellt wird. Außerdem steht neben der Reliabilität der KTT auch eine Reliabilitätsschätzung der IRT zur Verfügung (4.3).

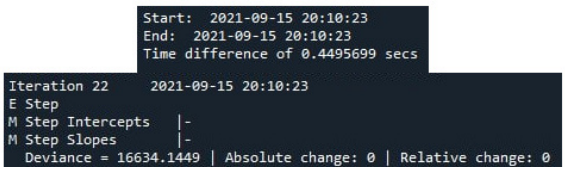
Auch bereitgestellt werden soll der Wert der Modelkonformität nach Unterabschnitt 3.4.6. Durch diesen Wert und einem zugehörigen Text sollen die Lehrenden den Wahrheitsgehalt der gebotenen Modelle besser einschätzen können.

4.3.1 Visualisierung

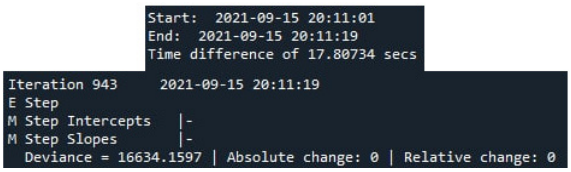
In der IRT sind die Möglichkeiten der Visualisierung begrenzt. Die typischen Darstellungsformen sind in den Abbildungen 3.5a, 3.5b, 3.6a, 3.6b und 3.7 zu sehen. Es besteht allerdings die Möglichkeit, die Darstellung über eine Wright-Map zu realisieren. Diese ist laut Trevor et al. [BYH20] eine gängige Methode der Visualisierung. Diese Art der Visualiserung benötigt jedoch die Personenwerte



(a) 1PL-Modell



(b) 2PL-Modell



(c) 3PL-Modell

Abbildung 4.1: Abbildung zeigt die Anzahl der Schritte im EM-Algorithmus für einen beispielhaften Datensatz von 16 Items und 828 Personen

der Modelle. Des Weiteren ist die Wright-Map schwerer zu lesen als die Abbildungen in 3.5a und 3.6a und würde eine ausführlichere Einführung benötigen, was jedoch gegen Anforderung **A3** (2.2) spricht. Deshalb steht diese Form der Visualisierung nicht zur Diskussion für den Fragebogen. Andere Arten der Darstellung würden die Informationsmenge begrenzen und sind daher nicht geläufig.
Es besteht die Möglichkeit, die Schwierigkeit des 1PL-Modell zu erfassen. Dies ist jedoch schon beim 2PL-Modell nicht mehr möglich. Dadurch wären die beiden Modelle schwerer miteinander zu vergleichen. Auch sind sie nicht mit der Itemschwierigkeit der deskriptivstatistischen Itemanalyse zu vergleichen, da sie einen anderen Wertebereich besitzen. Deshalb werden die Modelle als IC-Funktion dargestellt. Diese haben besonders für das 2PL-Modell eine hohe Aussagekraft und sind dadurch besser mit dem 1PL-Modell zu vergleichen.
Gleiches gilt für die *Test-Antwortfunktion* (3.29), die *Item-Informationsfunktion* (3.30) und die *Test-Informationsfunktion* (3.31). Eine andere Darstellungsform für diese Funktionen ist nicht gängig, da ihr Aussagegehalt nur über die Darstellung der IC-Funktion voll ausgeschöpft werden kann.

4.4 Fragebogen

Der Fragebogen (Anhang 7) dient der Evaluierung der angewendeten Visualisierungsmethoden und Werte. Dabei besteht der Fragebogen aus vier Teilen:

1. Einstieg
2. Aufgaben
3. KTT
4. IRT

Zu Beginn werden die Lehrenden nach ihrer aktuellen Testevaluation befragt, also ob sie die Fragen der Klausur evaluieren und für wie wichtig sie dies erachten. Danach folgen die in Kapitel 3 vorgestellten Verfahren. Zur Verständlichkeit für die Lehrenden werden die Methoden, Metriken und

Modelle zu Beginn kurz genauer erläutert und beschrieben. Des Weiteren werden für den Fragebogen prototypische Visualisierungen erstellt. Diese dienen dazu, die Lehrenden nach ihren bevorzugten Darstellungsformen zu befragen. Auch soll eine Freitextfrage zusätzliche Verbesserungswünsche und Vorschläge abfragen. In Abschnitt 6.1 wird der Fragebogen evaluiert und dargestellt, welche Methoden in der Anwendung Platz finden.

5 Implementierung

In diesem Kapitel wird zunächst erläutert, wie in dieser Arbeit bei der Implementierung der einzelnen Methoden und Modelle vorgegangen wurde. Im Anschluss wird beschrieben, wie die Anwendung in EvaExam integriert wird. Dabei werden die getroffenen Entscheidungen näher begründet und die Vorteile der verwendeten Bibliotheken herausgestellt.

5.1 Implementierung der Grundlagen

Die Implementierung der Grundlagen der Anwendung erfolgt in Python 3 [1]. In dieser Arbeit wird der Ansatz verfolgt, die Methoden in einzelne Module einzuteilen und zu implementieren. Dabei ist jedes Modul eine Python Script Datei. Die selbst geschriebenen Module sind dabei ausreichend getestet worden. Die Module sind wie folgt aufgeteilt:

Builder Anpassung an die interne Datenstruktur

Itemanalyse Methoden der deskriptiven Itemanalyse

KTT Berechnung der Reliabilität nach Vorbild der KTT

IRT Berechnung der Modelle der IRT

Visualisierung Darstellung der Ergebnisse der Itemanalyse, der KTT und der IRT-Modelle

Handler Verwaltet Daten und führt andere Module aus

Die grundlegende Datenstruktur des Programms stellt Tabelle 3.1 dar. Diese wird innerhalb des Programms als *numpy.array* [2] implementiert. *Numpy* bietet darüber hinaus den Vorteil, Verfahren der Statistik implementiert zu haben, die über *numpy.arrays* ausgeführt werden können. Jede Funktion jedes Moduls soll dabei mindestens diese grundlegende Datenstruktur in Form einer *numpy* Datenmatrix als Eingabe bekommen, ausgenommen dabei ist das Modul des Buildermoduls und des Visualisierungsmoduls.

Handler

Das Handlermodul stellt den Mittelpunkt des Programms dar. Es steuert den Datenfluss, führt die einzelnen Module aus und gibt die Visualisierung an EvaExam weiter.

[1] https://www.python.org/
[2] https://numpy.org/

Builder

Dieses Modul besitzt die Aufgabe, die Datenstruktur, die es als Eingabe bekommt, in ein *numpy.array* umzuwandeln. Zunächst werden in diesem Programm die Daten mit *pandas* [3] ausgelesen, da die Daten in CSV-Dateien vorliegen. Der Builder ist darauf ausgelegt, CSV-Dateien zu lesen, da diese von EvaExam exportiert werden. Zunächst wird die CSV-Datei mithilfe des *pandas*-Pakets ausgelesen und anschließend in ein *numpy.array* umgewandelt. Dabei werden überflüssige Metadaten entfernt. Jedoch werden wichtige Metadaten, wie beispielsweise Aufgabennamen, in einem extra Array gespeichert, um diese bei Bedarf zu verwenden.
Dies als Modul zu implementieren bietet den Vorteil, dass das Programm dadurch gut erweiterbar bleibt. Sollte ein neues Format als Eingabe dienen, muss das Buildermodul nur um die entsprechende Funktion erweitert werden.

Itemanalyse

Für die Implementierung der Methoden der deskriptivstatistischen Itemanalyse (3.2) wird für jede Funktion eine eigene Funktion implementiert. Diese greifen bei Bedarf auch gegenseitig aufeinander zu. Bei der Implementierung ist *numpy* dabei von großem Wert, durch die bereitgestellten Funktionen, wie zum Beispiel bei der Berechnung der Korrelation oder des Mittelwerts über eine Matrix und deren Dimensionen.
Da dieses Modul größtenteils aus selbst geschriebenem Code besteht, ist hierbei die Testung des Moduls von besonderer Bedeutung.

KTT

In diesem Modul sind die Funktionen zur Reliabilitätsschätzung (3.3.2) implementiert. Hier wurde für die *Split-half* Reliabilität (3.3.2), die *Odd-Even* Methode und die der *Item-Zwillinge* verwendet. Für die Wahl der *Item-Zwillinge* werden die Daten zur Berechnung der Itemschwierigkeit des Itemanalysemoduls verwendet.
Des Weiteren wurde Cronbachs α nach 3.24 implementiert.

IRT

Im IRT-Modul wird zur Hilfe der MML-Schätzung (3.4.4) und der EAP/PV-Reliabilität (3.4.5) der *Test Analysis Modules* (TAM) [4] verwendet. Diese Bibliothek steht allerdings nur in R [5] zur Verfügung, da R für statistische Berechnungen verwendet wird. In Python stehen keine vergleichbaren Bibliothken zur Verfügung, weshalb über R auf TAM zurückgegriffen wird. Dies wird durch die rpy2-Bibliothek [6] ermöglicht, welches R in Python ausführen lässt. Die so durch TAM berechneten und nach Modellkonformität überprüften Modelle werden anschließend wieder in eine Python

[3] https://pandas.pydata.org/
[4] https://cran.r-project.org/web/packages/TAM/TAM.pdf
[5] https://www.r-project.org/about.html
[6] https://rpy2.github.io/doc/latest/html/index.html

Datenstruktur überführt.
Die Nutzung dieser Bibliotheken erspart eine eigene Implementierung, welche auch eine ausreichende Testabdeckung erfordert. Dies sprengt jedoch den Umfang der Arbeit, weshalb auf die vorhandenen Bibliotheken zurückgegriffen wird.

Visualisierung

Itemanalyse

Zunächst wurden in diesem Modul die in Abschnitt 4.1 gewählten Kennzahlen mithilfe der *matplotlib* Bibliothek [7] visualisiert. Damit können aus *numpy.arrays* protypische Visualisierungen erstellt werden, welche gut für den Fragebogen geeignet sind. Zu sehen sind diese Visualisierungen im Anhang 7.
Für die Visualisierung in der endgültigen Anwendung wird eine interaktive Darstellungsweise gesucht, um **A4** aus Abschnitt 2.2 gerecht zu werden. Für die finale Visualisierung wird daher die *bokeh*-Bibliothek [8] verwendet. Diese Bibliothek ermöglicht es, einfach interaktive Visualisierungen zu erstellen. Diese werden in Form eines HTML-Strings zurückgegeben und werden daraufhin in EvaExam eingebunden. Die Entscheidung, welche Prototypen dabei weiterentwickelt werden, wird in Abschnitt 6.1 der Evaluation des Fragebogens ausgeführt.

Für *Itemschwierigkeit*, *Itemvarianz* und die *Trennschärfe* (3.2) werden jeweils drei verlinkte Grafiken erstellt. Dabei können die Nutzer*innen einzelne oder mehrere Aufgaben in den jeweiligen Grafiken markieren, welche ihnen dann in den jeweils anderen Grafiken hervorgehoben dargestellt werden. Dabei wurden die *Itemschwierigkeit*, die *Itemvarianz* und die *Trennschärfe* als Balkendiagramme dargestellt. Jede Grafiken besitzt Tooltips mit genaueren Informationen zu den Aufgaben.
Des Weiteren wurden zwei Formen der Testwertverteilung (3.2.4) mithilfe von *bokeh* visualisiert. Dabei wird die Testwertverteilung einerseits mit vergebener Punktzahl und andererseits mit der Anzahl der richtig beantworteten Fragen visualisiert. Die Verteilungen werden in Form eines Histograms im selben Graph dargestellt. Der Nutzer kann dabei entscheiden welche Verteilung ihm angezeigt wird. Herbei ist auch die gleichzeitige Darstellung möglich.

Reliabilität

Für die Reliabilität der KTT und der IRT werden zwei Werte auf einer farbigen Skala abgebildet. Diese sollen dabei helfen, die Reliabilität besser interpretieren zu können (rot = schlecht, grün = gut). Zusätzlich werden Tooltips hinzugefügt, um den genauen Wert der Reliabilität anzeigen zu lassen.

[7] https://matplotlib.org/
[8] https://docs.bokeh.org/en/latest/docs/dev_guide.html

IRT

Für die Modelle der IRT (3.4) wurden jeweils zwei Graphen erstellt. In diesen Graphen werden die IC-Funktionen der jeweiligen Modelle dargestellt. Über dem jeweiligen Graphen wird die Modellkonformität des Modells in Form des SRMSR-Werts und einer Farbe dargestellt, wobei die Farbe angibt, wie gut die Modellpassung ist (grün:gut - gelb: ausreichend - rot:schlecht). Dabei sind Tooltips für jede Aufgabe gegeben, um die genauen Werte und Aufgaben zu identifizieren. Auch werden die *Test-Informationsfunktionen* der beiden Modelle in einer gemeinsamen Grafik dargestellt. Dabei ist auswählbar, welche *Test-Informationsfunktionen* angezeigt werden soll. Hierbei kann zwischen dem 1PL-Modell und dem 2PL-Modell gewählt werden.

5.2 EvaExam

In diesem Abschnitt wird erläutert, wie die Daten aus EvaExam exportiert und anschließend in ausgewerteter und visualisierter Form in EvaExam dargestellt werden. Der zugehörige Datenfluss wird in Abbildung 5.1 abgebildet.

EvaExam verfügt über ein SDK, welches erlaubt die Daten per CSV-Datei zu exportieren. Um den in Abbildung 5.1 beschriebenen Datenfluss auszuführen, wird in EvaExam ein Button auf der Evaluationsseite der Lehrperson implementiert. Dieser fügt sich in die Oberfläche zu den bereits vorhandenen Auswertungsmethoden von EvaExam ein. Beim Bestätigen des Buttons durch die Lehrperson wird eine CSV-Datei mit den Rohdaten von EvaExam auf den Server exportiert. Daraufhin wird der Handler durch EvaExam ausgeführt. Der Handler greift dann auf die von EvaExam exportierten Daten zu und führt die weiteren Schritte aus. Sobald der Handler die HTML-Datei des Visualisierungsmoduls erhält, kann diese von EvaExam aufgerufen und dargestellt werden. Diese wird dann innerhalb von EvaExam dargestellt und kann bei Bedarf wieder geschlossen werden.

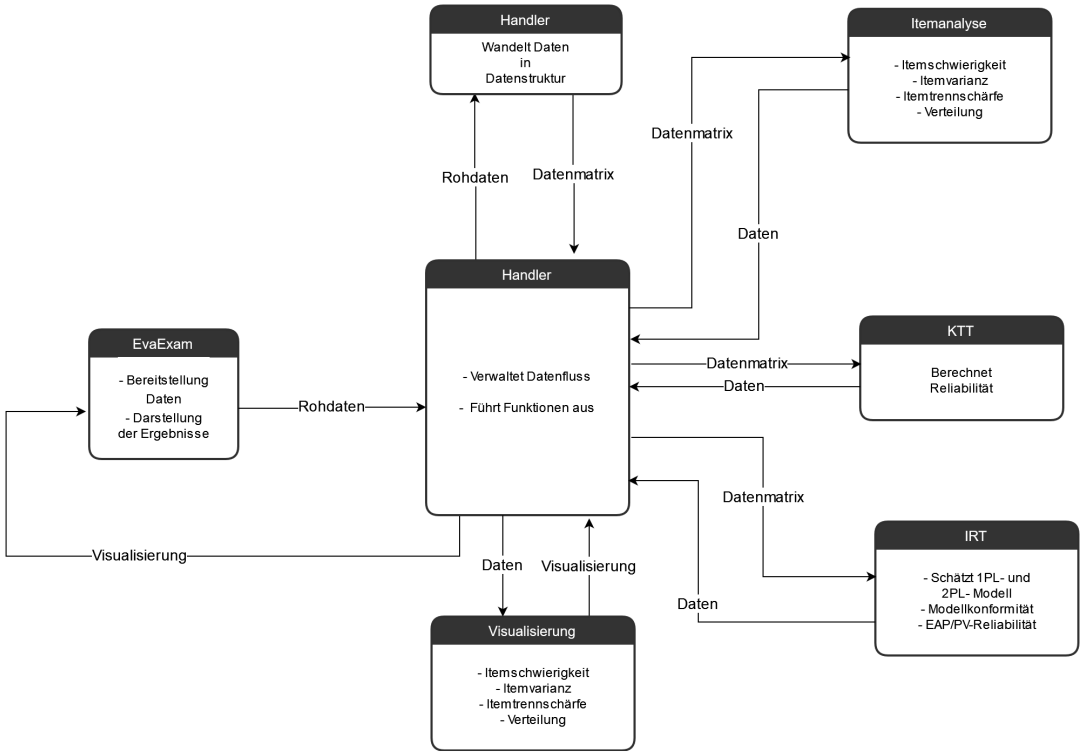


Abbildung 5.1: Datenfluss

6 Evaluation

Zunächst wird in diesem Kapitel der Fragebogen in Abschnitt 6.1 evaluiert. In Abschnitt 6.2 wird auf die in Abschnitt 2.2 gestellten Anforderungen Bezug genommen und untersucht, welche in dieser Arbeit erfüllt wurden. Im letzten Abschnitt 6.3 werden die angewandten Methoden evaluiert und gegenüber gestellt.

6.1 Fragebogenevaluation

Der Fragebogen wurde an sieben aktive EvaExam Nutzer*innen[1] verschickt. Dabei haben alle Nutzer*innen den Fragebogen bearbeitet. Bei einer aktiven Nutzer*innen Zahl von 38 Lehrpersonen entspricht das rund 18%. Dabei wurde die Stichprobe mit Lehrpersonen besetzt, die diverse Lehrhintergründe[2] besitzen. Zu finden sind die ausführlichen Ergebnisse der Umfrage in Anhang 7.

Einstieg

Der Großteil der Lehrenden gibt an, die vorhandene Evaluation durch EvaExam in ihren Workflow integriert zu haben. Zudem sind alle Lehrenden daran interessiert, ihre Aufgaben mithilfe von zusätzlichen Informationen zu verbessern. Dabei spielt für die Lehrenden eine detaillierte Analyse ihrer Klausuren eine wichtigere Rolle als die Analyse über die Fähigkeit der Studierenden. Die Umfrage ergibt zudem, dass die Lehrenden auch ein Interesse an analysierten Daten über Studierende besitzen.
Die erhobenen Daten zeigen, dass die Lehrenden einer Verbesserung ihrer Klausuren durch eine detaillierte Analyse der Klausurdaten zugeneigt sind. Darüber hinaus stellt sich heraus, dass bei Lehrpersonen auch Interesse an detaillierten Informationen über das Können ihrer Studierenden besteht. Es zeigt sich weiterhin, dass die Lehrenden verschiedene Einzelinteressen besitzen und Beispielsweise auch an Notenstatistiken interessiert sind. In den folgenden Abschnitten wird näher auf die einzelnen Teile des Fragebogens eingegangen und die Antworten der Befragten evaluiert.

[1] Als aktiv wurden alle Nutzer*innen gezählt, welche in den letzten 12 Monaten eine Klausur über EvaExam erstellt haben. Die Zählung geschieht durch interne Daten, welche aus Datenschutzgründen nicht veröffentlicht werden dürfen.

[2] Lehrpersonen aus verschiedenen Instituten mit geisteswissenschaftlichem und naturwissenschaftlichem Bezug.

Aufgaben

Dieser Abschnitt befragt die Lehrenden bezüglich der Darstellung der Metriken der deskriptivstatistischen Itemanalyse aus Abschnitt 3.2. Dabei stellt sich heraus, dass das Interesse an der Verteilung der Punkte und die Information über die Itemschwierigkeit besonders interessant für die Lehrenden sind. Jedoch sind die Varianz und die Trennschärfe der Items nicht zuvernachvlässigende Kenngrößen für die Lehrenden. Die restlichen vorgeschlagenen Ansätze sind zu vernachlässigen, da sie keinen ausreichenden Zuspruch durch die Lehrenden finden. Auch wichtig ist die genauere Beschreibung der angewandten Methoden. Die offene Frage 2.2 des Fragebogens zeigt auf, dass die Lehrenden genauer wissen möchten, wie genau diese Metriken zustande kommen.
Durch die Befragung der Lehrenden ist es möglich, abzuleiten, welche Metriken und Darstellungsformen einen besonderen Stellenwert für diese besitzen. Auch wird daraus abgeleitet, die finale Visualisierung mithilfe von Balkendiagrammen zu realisieren. Dies wird dabei auch von Schendera [Sch15] empfohlen . Diese sollen insbesondere bei der *Itemschwierigkeit*, der *Itemvarianz* und der *Itemtrennschärfe* verlinkt werden, da zwischen diesen Methoden eine Abhängigkeit besteht. Des Weiteren sollen in der finalen Anwendung die Methoden der Analyse genauer beschrieben werden, damit die Lehrenden die für sie wichtigen Informationen besser herausfiltern können.
In der finalen Anwendung stehen außerdem zwei Verteilungen in einem Graphen zur Verfügung:

1. Verteilung über Punkte
2. Verteilung über richtig beantwortete Fragen

Hierdurch soll eine mögliche Diskrepanz dargestellt werden können. Eine Verteilung über die Noten ist nicht möglich, da EvaExam dazu keine Metadaten liefert.

KTT

Die Reliabilität ist laut Fragebogenauswertung für jede Lehrperson von Bedeutung und sollte daher in der finalen Anwendung dargestellt werden.

IRT

Die Meinung der Lehrenden für die Darstellung des 1PL- und 2PL-Modells sind über die Darstellungsformen ähnlich verteilt. Dabei sticht die Item-Informationskurve des 2PL-Modells hervor. Diese scheint für die Lehrenden einen besonders hohen Informationsgehalt zu liefern.
Aus den Antworten der Fragen 4.1 und 4.2 des Fragebogens kann man auf eine allgemeine Schwierigkeit über die Lesbarkeit der vorgestellten Modelle schließen. Daraus kann man ableiten, dass auch hier eine genaue Beschreibung der Modelle durch einen Hilfstext von Bedeutung ist. Dennoch sollen in der fertigen Anwendung die Modelle dargestellt werden. Deshalb werden dort die IC-Funktionen und die *Test-Informationsfunktion* des 1PL- und 2PL-Modells dargestellt.

6.2 Anforderungsevaluation

Für die Anforderungsevaluation wird sich auf die in Abschnitt 2.2 für die Anwendung definierten Anforderungen bezogen.

A1 Zugänglichkeit Diese Anforderung wird erfüllt. Die Anwendung ist einfach über EvaExam zu erreichen und reiht sich in die bereitgestellten Auswertungsmethoden als Anwendung ein. Dadurch kann sie von jeder Lehrperson, welche EvaExam zur Klausurerstellung nutzt, verwendet werden.

A2 Benutzerfreundlichkeit Die Lehrenden haben die Möglichkeit, einfach und schnell zu der Anwendung zu gelangen und ohne Mehraufwand die Ergebnisse der Auswertung einzusehen. Dabei stehen ihnen keine Konfigurationsmöglichkeiten zur Verfügung, wie es die Anforderung verlangt. Daher wird diese Anforderung erfüllt.

A3 Reduziertheit Die Lehrenden haben Zugriff auf die in Abschnitt 6.1 gewünschten Methoden und Modelle. Die Schaubilder zeigen nur den von der Lehrperson ausgewählten Inhalt an. Es können auch Informationen verborgen werden, um die das Gewünschte darzustellen. Daher wird diese Anforderung erfüllt.

A4 Übersichtlichkeit Die Lehrenden können auswählen, welche Modelle und Metriken sie einsehen möchten. Dabei wurden die Visualisierungsmethoden nach den präferierten Möglichkeiten der Lehrenden gewählt (siehe dazu Abschnitt 6.1). Auch besteht die Möglichkeit, sich innerhalb der Graphen Informationen hervorheben zu lassen, um so wichtige Informationen einfacher herauszufiltern. Damit wird auch diese Anforderung erfüllt.

A5 Praktikabilität Die Anwendung ist intuitiv bedienbar und erfordert keinen Vorwissen. Dabei bietet die Anwendung alles, was die Lehrenden benötigen, um die Modelle und Metriken korrekt zu interpretieren. Dennoch kann diese Anforderung nicht als hundertprozentig erfüllt angesehen werden. Um dies zu überprüfen, bedarf es einer weiteren Befragung der Lehrenden, die im Rahmen dieser Arbeit nicht möglich ist.

Einen zusätzlichen Vorteil bietet die Anwendung durch ihren modularen Aufbau. Dadurch ist es leichter, diese in andere Plattformen wie beispielsweise Ilias zusätzlich zu integrieren. Dies entspricht zwar keiner Anforderung, sollte jedoch nicht unerwähnt bleiben, da **A1** durch das Ausrollen auf mehreren Plattformen stärker davon profitiert als eine feste Implementierung.

6.3 Evaluation der Theorie

Im Folgenden werden die in Kapitel 3 vorgestellten Methoden der Testtheorien miteinander verglichen.

Vergleich Deskriptivstatistische Itemanalyse und IRT

Zu Beginn wird die *Itemschwierigkeit* der deskriptistatistischen Itemanaylse mit der *Itemschwierigkeit* der IRT verglichen. Dabei kann man die *Itemschwierigkeit* in der IRT nur mit dem 1PL-Modell als Wert festsetzen. Dieser ist jedoch schwer zu vergleichen mit der *Itemschwierigkeit* der deskriptivstatistischen Itemanalyse. Die *Itemschwierigkeit* eines Items des 1PL-Modell ist dabei jedoch ein relativer Wert, der in Abhängigkeit zu den anderen geschätzten *Itemschwierigkeiten* der anderen Items des Modells steht. In der deskriptivstatistischen Itemanalyse ist die *Itemschwierigkeit* dagegen ein absoluter Wert. Dieser liegt immer im Bereich von 0 bis 100. Aus diesem Grund sind die *Itemschwierigkeiten* der beiden Methoden nicht miteinander zu vergleichen. Jedoch ergänzen die Methoden sich gegenseitig und können Unstimmigkeiten in den berechneten Werten aufzeigen.
Ein weiterer Vergleichswert ist die *Itemtrennschärfe*. Diese kann in der IRT mittels des 2PL-Modells dargestellt werden. Dabei zeigen sich jedoch die gleichen Probleme und Vorteile wie beim Vergleich mit der *Itemschwierigkeit* auf. Daher sind sie weniger als konkurriende Werte, denn als sich ergänzende Methoden zu verstehen.

KTT und IRT

Die in dieser Arbeit genutzte EAP/PV-Reliabilität lässt sich mit Cronbachs α vergleichen. Dabei gilt zu beachten, dass die verschiedenen Reliabilitätsschätzungen auf verschiedenen Annahmen beruhen. Cronbachs α gewichtet alle Items gleich, wohingegen die EAP/PV-Reliabilität aus einem IRT-Modell abgeleitet wird, welches den Items verschiedene Gewichtungen zuordnet. Das führt dazu, dass Cronbachs α Tests unterschätzt, aber dadurch auch eine untere Schranke für die Reliabilität darstellt. Die EAP/PV-Reliabilität muss immer gemeinsam mit dem Modell der IRT betrachtet werden. Dies führt auch innerhalb der IRT beim 2PL-Modell zu einer höheren Reliabilität, da dem 1PL-Modell strengere Annahmen unterliegen. Dennoch sind beide Theorien nur schwer miteinander zu vergleichen, da sie anderen Annahmen unterliegen.

Insgesamt lässt sich feststellen, dass die theoretischen Grundlagen der verschiedenen Ansätze nur schwer zu vergleichen sind. Die Methoden der IRT sollten daher als Ergänzung zur deskriptivstatistischen Itemanalyse und der KTT betrachtet werden.

7 Zusammenfassung und Ausblick

Die gesammelten Datenmengen steigen in der heutigen Zeit rasant an. Dadurch gibt es auch immer mehr ungenutzte Daten, welche nicht weiterverarbeitet werden. Auch an der Universität Stuttgart gibt es noch viele dieser ungenutzten Datenquellen. So auch die Daten im Rahmen der Testauswertung. In dieser Arbeit wurden die Möglichkeiten und Implementierbarkeit verschiedener Ansätze der Testtheorie aufgezeigt. Es stellte sich heraus, dass Lehrende großes Interesse an einer umfangreicheren Auswertung ihrer Klausuren haben und sie gleichzeitig auch daran interessiert sind, Informationen über die Studierenden zu erfahren. Jedoch besteht hierbei die Schwierigkeit, den Lehrenden die Ergebnisse gut zu vermitteln und ihnen Möglichkeiten der Interpretation aufzuzeigen. Dies ist insbesondere bei den Modellen der Item-Response-Theorie der Fall, da hier für einen Großteil der Lehrenden ungewohnte Methoden zum Einsatz kommen. Auch spielen hier verschiedene Parameter eine Rolle, die die Anleitung zur Interpretation in reiner Textform deutlich erschweren.
Dennoch konnte gezeigt werden, dass noch sehr viel Potenzial in dieser Art von Daten steckt. Durch die schon gegebenen Ansätze der Testtheorie sind die Werkzeuge für die Analyse gegeben, was die Arbeit an der Anwendung deutlich erleichtert hat.

Abschließend werden noch mögliche Erweiterungen und Verbesserungen der Anwendung aufgezeigt. Aus den gewonnenen Erkenntnissen geht hervor, dass die Vermittlung von Informationen eine nicht zu unterschätzende Hürde darstellt. Besonders neue Modelle sind für die Lehrenden nur schwer zu fassen. Daher ist es besonders sinnvoll zu ermitteln, wie man den Lehrenden die gewonnenen Informationen am besten vermittelt. Dies wird auch ein wichtiger Bestandteil des ‘Partnerschaft für innovative E-Prüfungen Projektverbund der baden-württembergischen Universitäten’ (PePP) Projektes, welches das Ziel verfolgt, das ungenutzte Potenzial von Klausuren für Lehrenden zu erschließen. Das Projekt findet parallel an allen baden-württembergischen Universitäten statt.
Darüber hinaus besteht die Möglichkeit, die Lehrenden aktiv in den Auswertungsprozess miteinzubeziehen, indem man beispielsweise die Lehrenden aktiv Parameter verändern lässt und somit die Möglichkeit besteht, die Modelle anzupassen. In der Item-Response-Theorie kann so aktiv eine Anpassung an die Modelle vorgenommen werden, um so die Modellkonformität zu steigern. Hierbei spielt jedoch die oben genannte Herausforderung eine wichtige Bedeutung, weshalb dies nicht ohne weiteres zu implementieren ist.
Möglich ist es auch, die Anwendung auf mehrere Plattformen wie Ilias zu erweitern oder eine Importmöglichkeit für selbst erstellte Klausuren zu bieten. Auch wäre eine Exportfunktion der Daten oder das (automatisierte) Hinzufügen von Metadaten für die einzelnen Fragen vorstellbar, um den zukünftigen Workflow der Klausurerstellung für die Lehrenden zu verbessern.
Auch wäre eine Auswertung über mehrere Klausuren denkbar. Dies bietet sich besonders für stichprobenabhängige Verfahren an oder um eine Übersicht über verschiedene Klausuren und Jahre zu erstellen.

Ein weiterer Schritt wäre die Ausweitung der Anwendung auf Studierende. Doch auch hier gilt es zu beachten, die Ergebnisse für die Studierenden richtig einzuordnen und gut darzustellen, da diese noch über weniger Wissen als Lehrende verfügen.

Literaturverzeichnis

[And88] D. Andrich, Hrsg. *Rasch models for measurement (Vol. 68)*. Newbury Park, CA: Sage, 1988. DOI: 10.4135/9781412985598 (zitiert auf S. 31).

[AWW97] R. J. Adams, M. Wilson, M. Wu. „Multilevel Item Response Models: An Approach to Errors in Variables Regression“. In: *Journal of Educational and Behavioral Statistics* 22.1 (1997), S. 47–76. DOI: 10.3102/10769986022001047 (zitiert auf S. 32).

[BA81] R. Bock, M. Aitkin, Hrsg. *Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm.* English. 1981, S. 443–459. DOI: 10.1007/BF02293801 (zitiert auf S. 31).

[Bir68] A. Birnbaum. „Some latent trait models.“ In: *Statistical theories of mental test scores*. Hrsg. von F. Lord, M. Novick. Reading: Addison-Wesley, 1968, S. 395–479 (zitiert auf S. 27, 28, 30).

[Bor10] J. Bortz. *Statistik für Human- und Sozialwissenschaftler*. Hrsg. von C. Schuster. 7., vollständig überarbeitete und erweiterte Auflage. SpringerLink Bücher. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010 (zitiert auf S. 20, 21).

[BYH20] T. G. Bond, Z. Yan, M. Heene. *Applying the Rasch model: fundamental measurement in the human sciences*. English. Fourth edition. New York, NY: Routledge, 2020 (zitiert auf S. 28, 29, 36).

[DL70] R. Darrell Bock, M. Lieberman, Hrsg. *Fitting a response model forn dichotomously scored items.* English. 1970, S. 179–197. DOI: 10.1007/BF02291262 (zitiert auf S. 30, 31).

[Gul50] H. Gulliksen. *Theory of mental tests*. John Wiley Sons Inc., 1950. DOI: 10.1037/13240-000 (zitiert auf S. 24).

[HFJ] J. Hartig, A. Frey, N. Jude. „Validität“. In: *Testtheorie und Fragebogenkonstruktion*. Hrsg. von H. Moosbbrugger, A. Kevala. 2., aktualisierte und überarbeitete Auflage. Heidelberg: Springer (zitiert auf S. 27).

[Kra81] H. T. Kranz. *Einführung in die klassische Testtheorie*. Fachbuchh für Psychologie, Verlag-Abt., 1981 (zitiert auf S. 20).

[Kra95] J. Krauth. *Testkonstruktion und Testtheorie*. Weinheim: Beltz, Psychologie Verl.-Union, 1995 (zitiert auf S. 19, 25).

[LBN68] F. Lord, A. Birnbaum, M. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley series in behavioral sciences: Quantitative methods. Charlotte, N.C.: Information Age Pub, 2008, Erstauflage: 1968. URL: https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=470245 (zitiert auf S. 24–26).

[Lie69] G. A. Lienert. *Testaufbau und Testanalyse*. ger. 3. Auflage. Weinheim: Beltz, 1969 (zitiert auf S. 11, 12, 17, 19–23, 25, 26).

[May13] A. Maydeu-Olivares. „Goodness-of-Fit Assessment of Item Response Theory Models“. In: *Measurement: Interdisciplinary Research and Perspectives* 11.3 (2013), S. 71–101. DOI: 10.1080/15366367.2013.831680 (zitiert auf S. 32).

[McN62] Q. McNemar. *Psychological Statistics*. 3. Auflage. New Yort, 1962 (zitiert auf S. 20).

[MK20] H. Moosbrugger, A. Kelava, Hrsg. *Testtheorie und Fragebogenkonstruktion*. 3., aktualisierte und überarbeitete Auflage. Berlin ; Heidelberg: Springer, 2020. ISBN: 978-3-662-61532-4. DOI: 10.1007/978-3-662-61532-4_2 (zitiert auf S. 17–19, 21, 24–28, 31–33).

[Ras60] G. Rasch, Hrsg. *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: The Danish Institute for Educational Research, 1960 (zitiert auf S. 27–29).

[Ros96] J. Rost. *Lehrbuch Testtheorie - Testkonstruktion. 2., vollst. überarb. u. erw. Aufl.* Bern u.a.: Huber, 2004, Erstauflage 1996. ISBN: 3-456-83964-2 (zitiert auf S. 18, 23, 27, 28, 31–33).

[Sch15] C. F. Schendera. *Deskriptive Statistik verstehen*. Stuttgart, Deutschland: UVK Verlag, 2015. Kap. Für das Auge Tabellen und Grafiken, S. 393. eprint: https://elibrary.utb.de/doi/pdf/10.36198/9783838539690. URL: https://elibrary.utb.de/doi/abs/10.36198/9783838539690 (zitiert auf S. 36, 46).

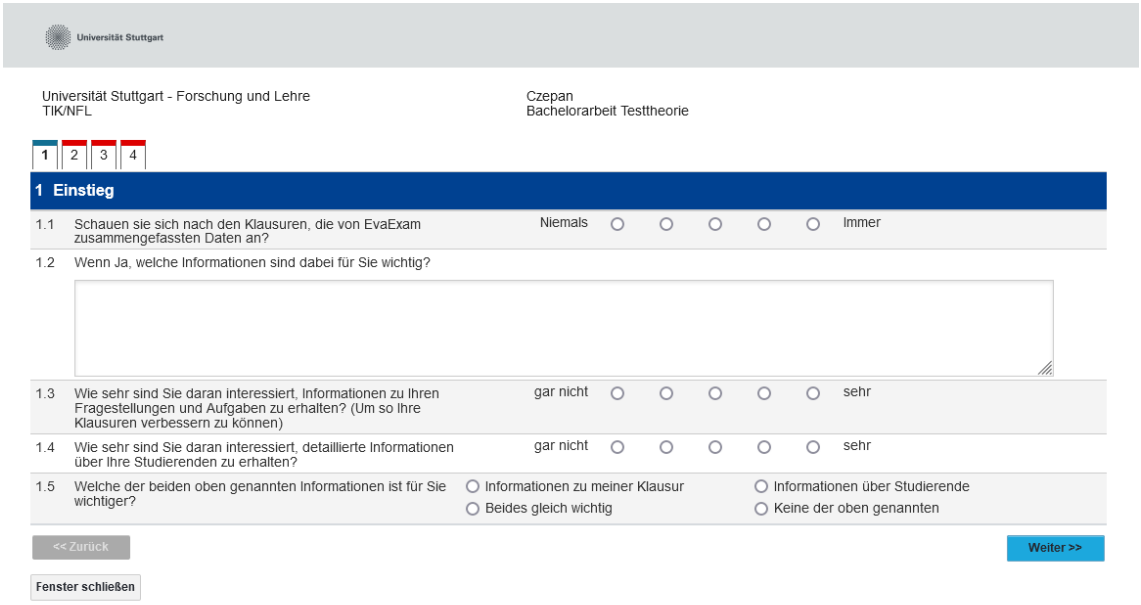
[SEM93] R. [Steyer, M. [Eid, A.-K. Mayer, Hrsg. *Messen und Testen*. Deutsch. Berlin ; Heidelberg [u.a.]: Springer, 1993, XIII, 397 Seiten. ISBN: 3540561692 (zitiert auf S. 27, 28).

[Tuf97] E. R. Tufte. eng. Cheshire, Conn: Graphics Press, 1997 (zitiert auf S. 36).

Alle URLs wurden zuletzt am 30. 09. 2021 geprüft.

Anhang

Fragebogen



Universität Stuttgart

Universität Stuttgart - Forschung und Lehre
TIK/NFL

Czepan
Bachelorarbeit Testtheorie

1 2 3 4

1 Einstieg

1.1	Schauen sie sich nach den Klausuren, die von EvaExam zusammengefassten Daten an?	Niemals	○	○	○	○	○	Immer
1.2	Wenn Ja, welche Informationen sind dabei für Sie wichtig?							
1.3	Wie sehr sind Sie daran interessiert, Informationen zu Ihren Fragestellungen und Aufgaben zu erhalten? (Um so Ihre Klausuren verbessern zu können)	gar nicht	○	○	○	○	○	sehr
1.4	Wie sehr sind Sie daran interessiert, detaillierte Informationen über Ihre Studierenden zu erhalten?	gar nicht	○	○	○	○	○	sehr
1.5	Welche der beiden oben genannten Informationen ist für Sie wichtiger?	○ Informationen zu meiner Klausur ○ Beides gleich wichtig				○ Informationen über Studierende ○ Keine der oben genannten		

<< Zurück

Weiter >>

Fenster schließen

Universität Stuttgart

Universität Stuttgart - Forschung und Lehre
TIK/NFL

Czepan
Bachelorarbeit Testtheorie

1 **2** 3 4

2 Aufgaben

Begriffsklärung:

Merkmal: Das von in Ihrer Klausur geforderte Wissen, um diese zu lösen.

Schwierigkeit: Gibt die Schwierigkeitsgrad einer Aufgabe an. (100 leicht - 0 schwierig)

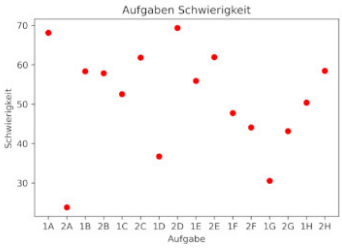
Varianz : Beschreibt die Differenzierungsfähigkeit einer Aufgabe (ist die Aufgabe mit der geforderten Merkmalsausprägung gut lösbar oder kann sie von jedem/keinem gelöst werden)

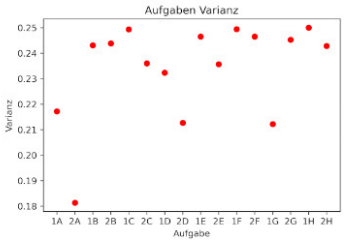
Trennschärfe: Zusammenhang zwischen Mermalsausprägung und Lösung der Aufgabe.
(**nahe 1** Aufgabe wird von Probanten mit hoher Merkmalsausprägung gelöst , mit niedriger nicht.
nahe 0: keine Aussage
nahe -1: Aufgabe wird von Probanten mit niedriger Merkmalsausprägung gelöst, mit hoher nicht.)

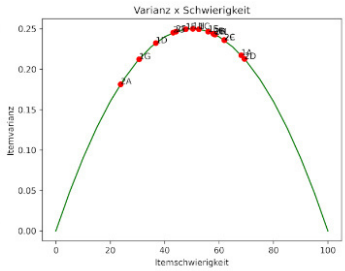
2.1 Welche der folgenden Darstellungsansätze erachten Sie als wichtig und/oder sinnvoll?
(Mehrfachnennung möglich)

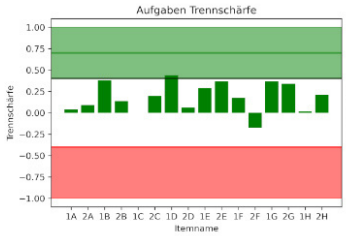
- ☐ Aufgaben Schwierigkeit
- ☐ Aufgaben Varianz
- ☐ Varianz x Schwierigkeit
- ☐ Aufgaben Trennschärfe (Balken)
- ☐ Aufgaben Trennschärfe (Punkte)
- ☐ Item Schwierigkeit x Trennschärfe
- ☐ Verteilung
- ☐ z-Wert Verteilung

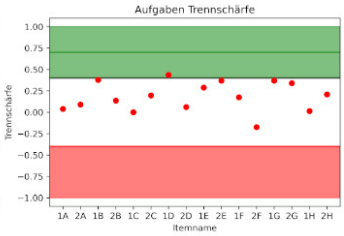
Jede Art der Darstellung soll mit Tooltips ausgestattet werden (hier nicht möglich).

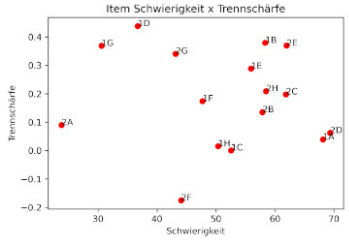


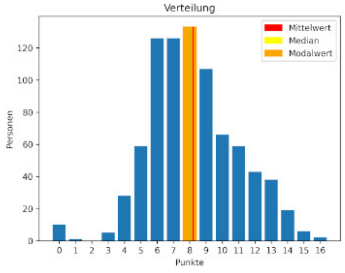


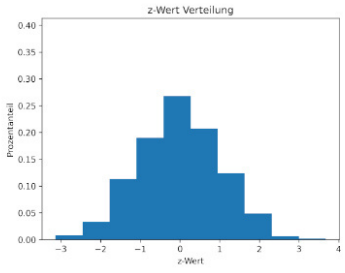










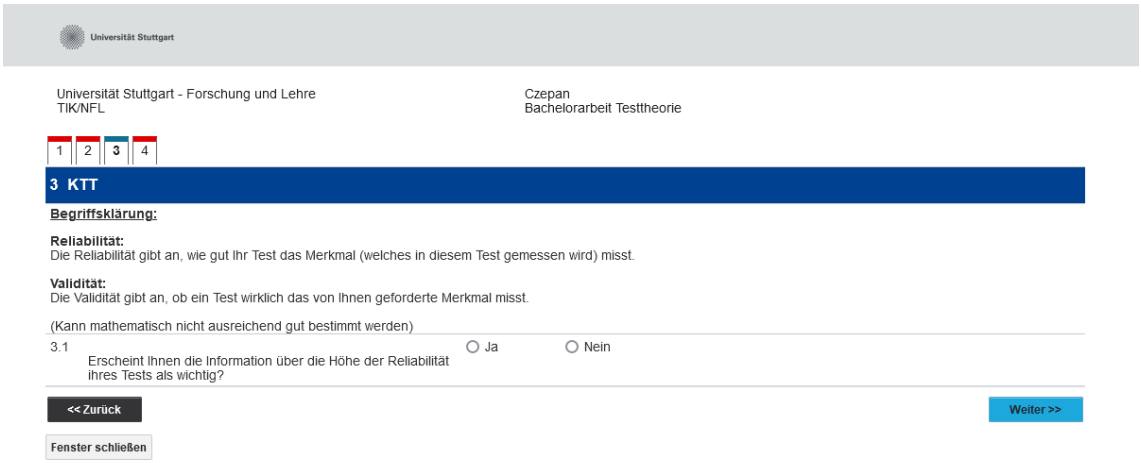


2.2 Wünsche oder Verbesserungsvorschläge bezüglich der Darstellung:

<< Zurück

Weiter >>

Fenster schließen



Universität Stuttgart

Universität Stuttgart - Forschung und Lehre
TIK/NFL

Czepan
Bachelorarbeit Testtheorie

1 2 3 4

3 KTT

Begriffsklärung:

Reliabilität:
Die Reliabilität gibt an, wie gut Ihr Test das Merkmal (welches in diesem Test gemessen wird) misst.

Validität:
Die Validität gibt an, ob ein Test wirklich das von Ihnen geforderte Merkmal misst.

(Kann mathematisch nicht ausreichend gut bestimmt werden)

3.1 Erscheint Ihnen die Information über die Höhe der Reliabilität ihres Tests als wichtig? ○ Ja ○ Nein

<< Zurück Weiter >>

Fenster schließen

Universität Stuttgart

Universität Stuttgart - Forschung und Lehre
TIK/NFL

Czepan
Bachelorarbeit Testtheorie

1 2 3 4

4 IRT

Begriffsklärung

1PL & 2PL sind Modelle aus der Testtheorie.

Die x-Achse beschreibt die benötigte Fähigkeit, eine Aufgabe zu lösen.
Die y-Achse beschreibt die Wahrscheinlichkeit, mit der die Aufgabe gelöst wird.

(Bsp. Frage 1 (1A&1B) : Bei einer Merkmalsausprägung von 0 ist die Warhscheinlichkeit, die Aufgabe zu lösen bei 50% | entspricht etwa einer Schwierigkeit von 50)

Das 2-PL Modell besitzt einen Parameter mehr. Dadurch kann zusätzlich so etwas wie die Trennschärfe einer Aufgabe beschrieben werden.

Test Response(2):
Die rote Kurve teilt sich die x-Achse, besitzt rechts jedoch eine eigene y-Achse.
Hierbei wird die Anzahl der richtig beantworteten Fragen in Relation mit der Merkmalsausprägung gesetzt.
(Beispiel (2A & 2B) : Bei einer Merkmalsausprägung von 0 löst ein Proband im Schnitt 2 Fragen richtig.)

Item Information (3) :
Beschreibt die Aussagekraft einer Aufgabe, bezogen auf die Merkmalsausprägung.
(Beispiel: Frage 1 (3A&3B) : Für Personen mit einer Merkmalsausprägung von 0 hat die Frage die höchste Aussagekraft.)

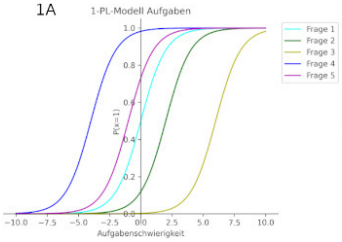
Test Information Function(3):
Beschreibt die Aussagekraft des Tests, bezogen auf die Merkmalsausprägung.

(Beispiel: 1PL-Modell-Item Information : Für Personen mit einer Merkmalsausprägung von 0 hat der Test die höchste Aussagekraft.)

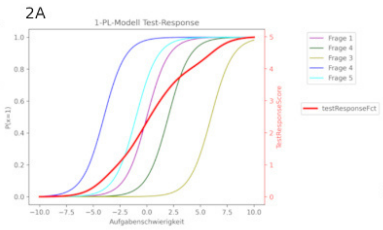
4.1 Welche der folgenden Darstellungsansätze erachten Sie als wichtig?
(Mehrfachnennung möglich)

- ☐ 1-PL-Modell Aufgaben
- ☐ 1-PL-Modell Test Response
- ☐ 1-PL-Modell Item Information
- ☐ 2-PL-Modell Aufgaben
- ☐ 2-PL-Modell Test Response
- ☐ 2-PL-Modell Item Information

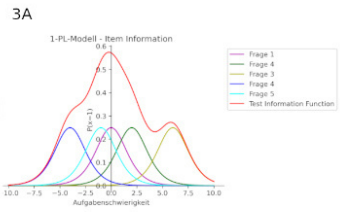
1A



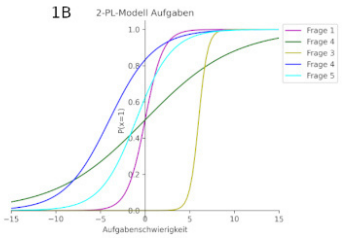
2A



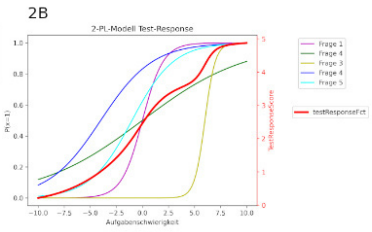
3A



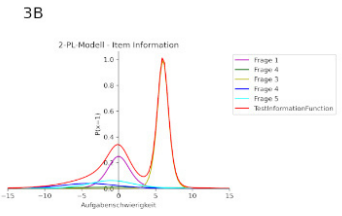
1B



2B



3B



4.2 Wünsche oder Verbesserungsvorschläge bezüglich der Darstellung:

<< Zurück

Fenster schließen

Fragebogen Auswertung

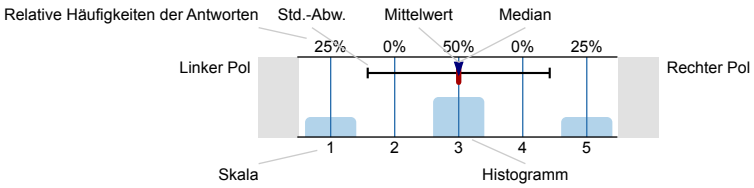
Czepan

Bachelorarbeit Testtheorie ()
Erfasste Fragebögen = 7

Auswertungsteil der geschlossenen Fragen

Legende

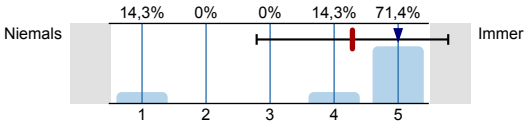
Fragetext



n=Anzahl
mw=Mittelwert
md=Median
s=Std.-Abw.
E.=Enthaltung

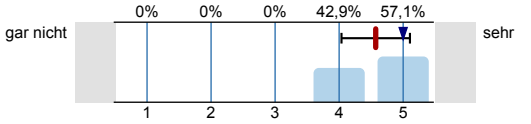
1. Einstieg

1.1) Schauen sie sich nach den Klausuren, die von EvaExam zusammengefassten Daten an?



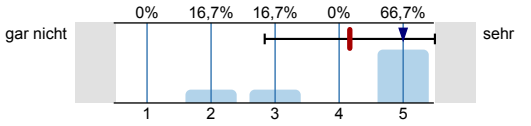
n=7
mw=4,3
md=5
s=1,5

1.3) Wie sehr sind Sie daran interessiert, Informationen zu Ihren Fragestellungen und Aufgaben zu erhalten? (Um so Ihre Klausuren verbessern zu können)



n=7
mw=4,6
md=5
s=0,5

1.4) Wie sehr sind Sie daran interessiert, detaillierte Informationen über Ihre Studierenden zu erhalten?



n=6
mw=4,2
md=5
s=1,3

1.5) Welche der beiden oben genannten Informationen ist für Sie wichtiger?

n=7

Informationen zu meiner Klausur	57.1%
Informationen über Studierende	0%
Beides gleich wichtig	42.9%
Keine der oben genannten	0%

2. Aufgaben

2.1) Welche der folgenden Darstellungsansätze erachten Sie als wichtig und/oder sinnvoll? (Mehrfachnennung möglich)

Jede Art der Darstellung soll mit Tooltips ausgestattet werden (hier nicht möglich).

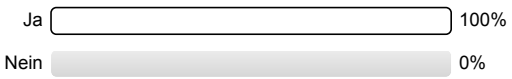
n=7

Aufgaben Schwierigkeit	71.4%
Aufgaben Varianz	57.1%
Varianz x Schwierigkeit	0%
Aufgaben Trennschärfe (Balken)	28.6%
Aufgaben Trennschärfe (Punkte)	57.1%
Item Schwierigkeit x Trennschärfe	28.6%
Verteilung	71.4%
z-Wert Verteilung	14.3%

3. KTT

3.1) Erscheint Ihnen die Information über die Höhe der Reliabilität ihres Tests als wichtig?

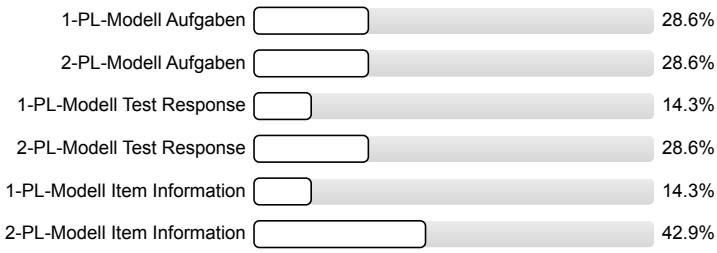
n=7



4. IRT

4.1) Welche der folgenden Darstellungsansätze erachten Sie als wichtig? (Mehrfachnennung möglich)

n=7



Profillinie

Teilbereich: TIK/NFL
Name der/des Lehrenden: Czepan
Titel der Lehrveranstaltung: (Name der Umfrage) Bachelorarbeit Testtheorie

Verwendete Werte in der Profillinie: Mittelwert

1. Einstieg

1.1)	Schauen sie sich nach den Klausuren, die von EvaExam zusammengefassten Daten an?	Niemals		Immer	n=7	mw=4,3	md=5,0	s=1,5
1.3)	Wie sehr sind Sie daran interessiert, Informationen zu Ihren Fragestellungen und Aufgaben zu erhalten? (Um so Ihre Klausuren	gar nicht		sehr	n=7	mw=4,6	md=5,0	s=0,5
1.4)	Wie sehr sind Sie daran interessiert, detaillierte Informationen über Ihre Studierenden zu erhalten?	gar nicht		sehr	n=6	mw=4,2	md=5,0	s=1,3

Auswertungsteil der offenen Fragen

1. Einstieg

1.2) Wenn Ja, welche Informationen sind dabei für Sie wichtig?

- - Distraktoranalyse
  - relative Lösungsquoten der Items
  - Veränderung der gesamten Gruppe, wenn einzelne Items aus der Bewertung gestrichen würden (fehlt momentan)
  - Personen die an Item x scheitern, lösen/scheitern an y (fehlt momentan)
- ... mache ich über SPSS. Dabei schaue ich mir v. a. an: Mittelwerte und Trennschärfen der Items.
- Die Notenstatistiken sowie die Punkteverteilung einzelner Aufgaben
- Durschnittliche Punkte je Aufgabe, Verteilung der Noten,
- Gesamtverteilung der Punkte
  durchschnittliche Punktezahl bei den Textaufgaben
- Statistik in welchen Semestern Aufgaben verwendet wurden und wie gut diese jeweils von Studierenden beantwortet wurden.
- Zahl der richtigen Antworten pro Frage (Schwierigkeitsgrad)
  Trennschärfe

2. Aufgaben

2.2) Wünsche oder Verbesserungsvorschläge bezüglich der Darstellung:

- Alle Darstellungen mit den Fragen auf der x-Achse als Balkendiagramme (übersichtlicher als mit Punkten)
- Ich würde Einspruch gegen die Interpretation der Itemschwierigkeit einlegen, wenn hier Lösungsquoten gemeint sind (KTT), oder wird hier tatsächlich ein Rasch-Modell zugrunde gelegt? Mit der Trennschärfe meinen Sie Item-Rest-Korrelationen oder gehen die Überlegungen in Richtung von einem Birnbaum-Modell?
  Mit der hier gegebenen Erläuterung der Varianz (s^2) würde ich eher überlegen, ob Sie nicht mit der SD eine leichtere Darstellung hätten?
- ich habe Verständnisprobleme mit Ihren Begrifflichkeiten, um hier sinnvolle Aussagen zu machen.
  Beispielsweise, verstehe ich nicht wie die "Schwierigkeit" gemessen wird. Das kann höchstens ich, sagen, wenn ich die Fragen entwickle.
  z-Wert wovon?
  Sorry

4. IRT

4.2) Wünsche oder Verbesserungsvorschläge bezüglich der Darstellung:

- zwei y-Achsen im einem X-y-Diagramm finde ich verwirrend.
- A) Ich wäre nicht unbedingt mit den gegebenen Verständnissen von Reliabilität und Validität einverstanden. Hier ist in meinen Augen einerseits die Präzision der Messung das geschicktere Verständnis, weil es die Brücke zur Signifikanz schlägt. Das Validitätsverständnis als Testeigenschaft ist eigentlich veraltet und wird abgelöst durch den Grad der Angemessenheit der Testwertinterpretation.

  B) Die Modelle gehen mit ziemlich harten Annahmen einher - diese als pauschal erfüllt anzusehen, ist problematisch. Dies führt im Zweifelsfall zu kritischen Missverständnissen bei der Ergebnisdeutung.

  C) Sollte ein Partial-Credit-Modell implementiert werden, brauchen die Benutzer eine zielführende Erklärung für die Interpretation invertierter Schwellen.
- ich könnte aus den Darstellungen nicht erkennen, welche Items "ungünstig" sind und daher bei der Bewertung nicht berücksichtigt werden sollten bzw. in Zukunft verändert werden sollten...

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Stuttgart, 01.10.2021

Ort, Datum, Unterschrift