

# 1 Einleitung

In unterschiedlichen Situationen des Alltags finden psychologische Tests Anwendung. Ziel ist es unter anderem, Aussagen über die Ausprägung bestimmter Persönlichkeitsmerkmale von Personen zu treffen. Gerade in der Psychologie ist es nicht einfach, die zu messenden Eigenschaften in Zahlen zu fassen, da es sich um latente, d.h. nicht beobachtbare Merkmale handelt. Aufschluss über die interessierenden Größen soll die Beantwortung mehrerer Aufgaben eines psychologischen Tests geben [Strobl, 2010]. Im Folgenden werden die Aufgaben eines solchen Tests immer mit Items bezeichnet.

Bei einem Intelligenztest wird beispielsweise erfasst, wie viele Items eine Testperson richtig gelöst hat. Als Ergebnis erhält die jeweilige Person eine Schätzung ihrer Fähigkeit [Strobl, 2010]. Das wohl bekannteste statistische Modell zur Auswertung der Ergebnisse solcher Intelligenztests ist das Rasch-Modell [Rasch, 1960]. Dieses ist ein Vertreter der probabilistischen Testtheorie bzw. Item-Response-Theorie (IRT). Die IRT umfasst eine Familie von mathematischen Messmodellen, welche postulieren, dass den beobachtbaren manifesten Daten (hier die Antworten auf Testitems) latente Variablen wie z.B. Eigenschaften oder Fähigkeiten der Personen zugrunde liegen, die das Testverhalten steuern [Rost und Spada, 1982].

## 1.1 Gegenstand der Arbeit

Geht man in einer Testsituation davon aus, dass das Rasch-Modell Gültigkeit besitzt, so ist die Wahrscheinlichkeit für die richtige Beantwortung eines Testitems für alle Personen mit derselben Fähigkeit exakt gleich. Falls dies nicht erfüllt ist und die Wahrscheinlichkeit für die richtige Beantwortung bestimmter Testitems für Personen verschiedene Subgruppen mit derselben Fähigkeit unterschiedlich ist, spricht man von „Differential

Item Functioning“ (DIF) [Osterlind und Everson, 2009]. Differential Item Functioning tritt beispielsweise dann auf, wenn ein Item für eine Gruppe eines der schwierigsten und für eine andere Gruppe eines der leichtesten Items darstellt. Differential Item Functioning heißt aber nicht einfach, dass ein Item für eine Gruppe schwerer zu lösen ist als für eine andere. Bestehen nämlich grundsätzliche Wissensunterschiede, z.B. zwischen Gruppen von Studenten, werden diese im gesamten Test besser bzw. schlechter abschneiden. DIF ist also vorhanden, falls ein Item für eine Gruppe wesentlich schwerer zu beantworten ist als für eine andere Gruppe, nachdem der allgemeine Wissensunterschied über die Thematik des Tests berücksichtigt wurde. Typische Variablen zur Untersuchung von Subgruppeneffekten sind Rasse, Religion und Geschlecht [Osterlind und Everson, 2009].

Die vorliegende Arbeit beschäftigt sich mit einer Erweiterung des klassischen Rasch-Modells zur Berücksichtigung des Differential Item Functioning. Dies wird durch die Hinzunahme sogenannter „itemmodifizierender Effekte“ erreicht. Hauptziel der Analysen ist es herauszufinden, für welche Items itemmodifizierende Effekte vorhanden sind, d.h. welche Items in verschiedenen Subgruppen unterschiedlich beantwortet werden. Des Weiteren ist von Interesse, welches die relevanten Subgruppen-Variablen sind, für die itemmodifizierende Effekte vorhanden sind.

Um bei der Schätzung der vorgestellten Modelle die gewünschte Variablenelektion zu erzielen und dem Problem der großen Anzahl zu schätzender Parameter vorzubeugen, ist eine gewöhnliche Maximum-Likelihood-Schätzung nicht umsetzbar. Inhalt der Arbeit ist die Schätzung der Modelle mithilfe von Boosting. Dies ist eine Möglichkeit, mit der regularisierte Maximum-Likelihood-Schätzungen durchgeführt werden können. Einen alternativen Ansatz durch penalisierte Maximum-Likelihood Schätzung untersuchen Tutz und Schauburger [2013]. Ein Großteil der theoretischen Aus-

fürungen in Kapitel 2, unter anderem die Einbettung der betrachteten Modelle in das Framework der generalisierten Regressionsmodelle, basiert auf den Vorarbeiten von Tutz und Schauberger [2013]. Die Stärke dieser Betrachtungsweise ist, dass die Variablen, für welche itemmodifizierende Effekte untersucht werden, nicht nur binär oder kategorial, sondern auch stetig sein können und, dass die Anzahl an Variablen beliebig groß sein kann.

## 1.2 Gleichmäßiges und ungleichmäßiges DIF

Im Zusammenhang mit itemmodifizierenden Effekten gilt es im Allgemeinen zwei Konzepte zu unterscheiden. Itemmodifizierende Effekte können entweder gleichmäßig oder ungleichmäßig vorliegen. Unter einem itemmodifizierenden Effekt versteht man den Unterschied der Wahrscheinlichkeit einer korrekten Antwort auf ein Testitem zwischen Personen verschiedener Subgruppen mit derselben Fähigkeit. Falls dieser Unterschied unabhängig von der Fähigkeit der Personen immer gleich ist, spricht man von „gleichmäßigem“ DIF. Ist dieser Unterschied nicht konstant, sondern von der Fähigkeit der Person abhängig, so spricht man von „ungleichmäßigem“ DIF [Osterlind und Everson, 2009]. Abbildung 1.1 visualisiert die beiden unterschiedlichen Effekte beispielhaft für den einfachen Vergleich zweier Gruppen. Die eingezeichneten Kurven ergeben sich bei Modellierung der Wahrscheinlichkeiten durch ein logistisches Regressionsmodell. In Abbildung 1.1 sind diese rein qualitativ zur Verdeutlichung der beschriebenen Effekte dargestellt.

In der linken Graphik in Abbildung 1.1 sieht man, dass die Wahrscheinlichkeit für eine korrekte Antwort in Gruppe 2 immer höher ist als in Gruppe 1. In der rechten Graphik hingegen schneiden sich die beiden Kurven. Unter den Personen mit geringeren Fähigkeiten ist die Wahrscheinlichkeit einer

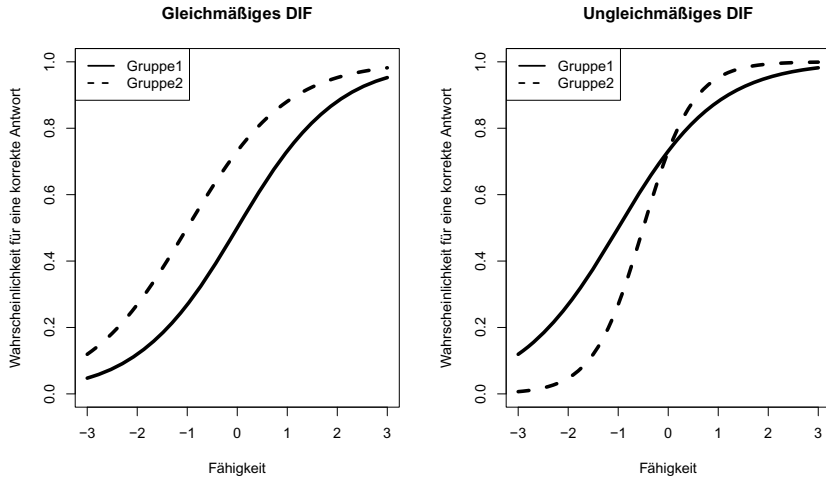


Abb. 1.1: Wahrscheinlichkeit für die richtige Beantwortung einer Frage in Abhängigkeit der Fähigkeit der Person. Unterschieden werden gleichmäßige Effekte (links) und ungleichmäßige Effekte (rechts).

richtigen Antwort in Gruppe 1 höher als in Gruppe 2. Unter den Personen mit höheren Fähigkeiten ist dies genau umgekehrt.

Anhand der Item-Response-Modelle, die Gegenstand der Arbeit sind (Kapitel 2), können nur gleichmäßige itemmodifizierende Effekte modelliert werden. Die Unterschiede zwischen den Subgruppen sind für alle Personen, unabhängig von ihrer Fähigkeit, immer gleich.

### 1.3 Aufbau der Arbeit

In Kapitel 2 werden die in der Arbeit betrachteten Modelle, insbesondere das Rasch-Modell mit itemmodifizierenden Effekten, vorgestellt. Entscheidend für die Schätzung der Modelle ist die Einbettung in das Framework

der generalisierten Regressionsmodelle.

Kapitel 3 führt in die Theorie des Boosting ein und erläutert im Besonderen die Vorgehensweise für das Rasch-Modell mit itemmodifizierenden Effekten. An entsprechenden Stellen wird auch auf die praktische Umsetzung mit statistischer Software eingegangen.

In Kapitel 4 werden alternative Schätzverfahren eingeführt, die sich ebenfalls zur Modellierung itemmodifizierender Effekte eignen.

Kapitel 5 beinhaltet eine Simulationsstudie, in der untersucht wird, wie gut die Boosting-Methode zur Modellierung relevanter itemmodifizierender Effekte geeignet ist.

Abschließend enthält Kapitel 6 zwei Anwendungsbeispiele, an denen die Schätzung mithilfe von Boosting praktisch umgesetzt wird. Anhand der Simulationsergebnisse aus Kapitel 5 kann Rückschluss auf die Güte der Schätzung gezogen werden.

Alle Analysen, die in der Arbeit vorgestellt werden, wurden mit der Software **R** durchgeführt [R Core Team, 2013]. Anhang B enthält eine Übersicht der erstellten Ordnerstruktur, in der die erzeugten Source-Dateien („R“) und die Ergebnisse der Analysen („RData“) gespeichert wurden.

In den mathematischen Formeln und Ausdrücken der Arbeit sind Vektoren klein und fett markiert (z.B.  $\gamma$ ) und Matrizen groß und fett markiert (z.B.  $\mathbf{Z}$ ), um diese von Skalaren und Funktionen zu unterscheiden.

Boosting-Techniken zur Modellierung

itemmodifizierender Effekte

Eine Erweiterung klassischer Item-Response-Modelle

Berger, M.

2015, IX, 125 S. 45 Abb., Softcover

ISBN: 978-3-658-08704-3