

Super Title

Master Thesis

David-Matthias Sichau

Supervisors:
Pitt Hild,

GroupName

24. September 2014, Zürich

Inhaltsverzeichnis

1	Einleitung	1
2	Theorie und Problemlage	2
2.1	Transfer	2
2.1.1	Historischer Überblick	2
2.1.2	Kritiken	4
2.1.3	Lobato und Sievert 2002	4
2.1.4	Elemente von Transfer	5
2.1.5	Konsequenzen für den Unterricht	6
2.1.6	Zusammenfassung zu Transfer	7
2.2	Kompetenz	7
2.2.1	Bildungsreformen	8
2.2.2	Definition von Kompetenz	8
2.2.3	Kompetenz und Transfer	10
2.3	Kompetenz des skalenbasiertes Messens	11
2.4	Forschungsfrage	11
3	Untersuchungsanlage	12
3.1	Anforderungen	12
3.2	Umsetzung	12
3.3	Durchführung	13
3.3.1	Vorbereitung	13
3.3.2	Durchführung	14
3.3.3	Nachbereitung	14
4	Ergebnisse	16
4.1	Kodierung	16
4.1.1	Items	16
4.1.2	Qualitätsstandards	17
4.1.3	Niveau	18
4.2	Fragebogen	19
4.3	Unterschiede zwischen den Klassen	20
4.4	Korrelation der Niveaus des skalenbasierten Messens	21
4.5	Rasch-Analyse	23
4.5.1	Parameterschätzung	23
4.5.2	Modellkontrolle des Rasch-Modells	24
4.5.3	Unterschied in den Qualitätsstandards	26
5	Diskussion	29
6	Ausblick	30

Abstrakt

Code erhältlich auf:

GitHub

<http://sichau.eu>

Toller Abstrakt

1 Einleitung

Im Rahmen der ExKoNawi hands-on Testaufgaben soll untersucht werden inwiefern die Qualitätsstufen von verschiedenen Test zur Erhebung der Kompetenz des Messens korrelieren.

Dabei soll insbesondere untersucht werden, ob der fachliche Kontext oder der inhaltliche Kontext ausschlaggebend ist. In anderen Wort hängt die Leistung der SuS davon ab, ob die Kompetenz des Messens im gleichen Fach gemessen wird, oder hängt dies davon ab was für eine Messung durchgeführt wird.

Diese Frage ist besonders von Interesse, wenn es darum geht Test auszuwählen. Ist es möglich die Kompetenz des Messens mit einem einzelnen Test zu erheben oder sind mehrerer Tests notwendig um das erreichte Kompetenz Niveau ausreichend zu beschreiben.

2 Theorie und Problemlage

Für die Beantwortung und Entwicklung der Forschungsfrage ist es wichtig den Begriff des Kontextes zu definieren. Zu Beginn wird der Begriff des Kontextes aus dem Sichtwinkel des Transfers untersucht. In einem zweiten Teil wird der Kontext basierend auf dem Kompetenzbegriff analysiert. Zuletzt werden die Erkenntnisse gesammelt und die genaue Forschungsfrage dieser Masterarbeit definiert.

2.1 Transfer

Es wird erwartet, dass in der Schule vermitteltes Wissen universell aufgerufen werden kann und dass das Gelernte auf andere Kontexte angewendet werden kann. Dieses universell verfügbare Wissen ist eng mit dem Begriff des Transfers verknüpft. **Greeno1996** definierten Transfer als „the process of applying knowlege in new situations“. Aber auch innerhalb der schulischen Bildung gibt es einen Transfer zwischen den verschiedenen Fächern. So wird von Schülerinnen und Schüler erwartet, dass Fähigkeiten, Lösungsstrategien, Konzepte und anderes Wissen auf andere Fächer übertragen werden soll und dort abgerufen werden kann.

2.1.1 Historischer Überblick

Um den Begriff des Kontextes im Zusammenhang mit Transfer zuverorten, soll zuerst ein historischer Überblick über den Begriff des Transfers gegeben werden.

Woodworth 1901

Eines der ersten Experimente zu Transfer wurde von **Woodworth1901** gemacht. Dabei mussten Probanden die Grösse von Rechtecken schätzen. Nachdem die Personen sich durch Wiederholungen verbessert hatten, wurde ihnen zwei neue Test Sets gegeben. In einem gab es neue Rechtecke, welche im ursprünglichen Set nicht enthalten waren. Die zweite Gruppe bekam Sets bei denen andere Formen enthalten waren (z. B. Kreise und Dreiecke). Die zweite Testgruppe machte ähnlich viel Fehler, wie vor dem Training mit den Rechtecken. Daraus schloss **Woodworth1901** dass kein Transfer stattgefunden haben kann.

Ein ähnliches Resultat auf universitärem Niveau konnte von **Renkl1994** gezeigt werden. Er konnte zeigen, dass Nichtökonom eine simulierte Firma besser führten, als Studierende der Betriebswissenschaften kurz vor ihrem Abschluss. Diese Resultate führen zu dem Schluss, dass Transfer nur sehr schwierig erreicht werden kann, und wenn oft nur unter sehr ähnlichen Bedingungen. Die **Woodworth1901** Theorie zu Transfer basiert auf der Idee von identischen Elementen (**Pea2013b**). In diesem Theorie-Verständnis entsteht Transfer, wenn Wissen auf zwei verschiedenen Aufgaben, welche jedoch identische Merk-

male/Elemente besitzen, angewendet wird. Dieses Transfers Verständnis basiert und stützt das Reiz-Reaktions-Modell des Lernens (**Detterman1993**; **Mietzel2007**).

Ferguson 1956

Eine alternative Theorie zum Transfer wurde von **Ferguson1956** entwickelt. Fergusons Theorie basiert darauf, dass die Intelligenz einer Person sich auf deren Transferleistung auswirkt. So findet nach **Ferguson1956** bei dem Lernen permanent ein Transfer statt, da jede Lernaufgabe von der anderen unterschiedlich ist und daher Transfer stattfinden muss. Im Unterschied zu **Woodworth1901** betrachtet **Ferguson1956** Transfer als einen kontinuierlichen Prozess, welcher durch Lernen verbessert werden kann. Wichtig ist jedoch zu beachten, dass **Ferguson1956** Theorie nur Nah-Transfer beschreibt. Unter Nah-Transfer wird Transfer zwischen sehr ähnlichen Situationen definiert.

Judd 1908

Eine der grundlegenden Studien zu Fern-Transfer, bei welchem erworbenes Wissen auf Kontexte angewendet werde soll welche sich deutlich vom Kontext, unter welchem das Wissen erworben wurde, unterschieden, wurde von **judd1908** gemacht. Im Vergleich zu **Woodworth1901** geht Judd davon aus, dass der Unterschied zwischen den beiden Situationen nicht nur abhängig von der Ähnlichkeit und den Unterschieden zwischen den beiden Situationen ist, sondern auch davon abhängt, wie die erste Situation gelernt wurde. Um dies zu belegen führte Judd eine sehr bekannte Studie durch. Bei dieser wurden Kinder genommen, welche mit einem Dart auf eine Zielscheibe unter Wasser werfen sollten. Beide Gruppen bekamen zu beginn die Möglichkeit dies zu trainieren. Später wurde das Werfen wiederholt, wobei die Position der Zielscheibe jedoch unterschiedlich war, und untersucht welche Gruppe besser war. Eine der beiden trainierten Gruppen wurde während sie die Situation A trainierten erklärt, warum die Scheibe so schwierig zu treffen war. Indem ihnen zusätzlich zum Training noch das Prinzip der Lichtbrechung erklärt wurde. Die Gruppe welche die Erklärung bekommen hatte schnitt unter der neuen Situation deutlich besser ab, als die andere Gruppe. **judd1908** erklärte dieses damit, dass die einen wussten, welches Prinzip sie auch bei der zweiten Situation anwenden können. Die Schülerinnen und Schüler welchen das Prinzip nicht erklärt wurden, haben gelernt ihren Wurf auf die erste Situation anzuwenden, konnten dieses Wissen jedoch nicht generalisieren, da dies spezifisch für die Situation erworben wurde.

Im Vergleich zu **Woodworth1901** beinhaltet die Theorie von **judd1908** einen kognitivistisches Verständnis des Lernens. Da die Lernenden ein immer besseres Verständnis der Welt um sich selbst konstruieren und so neue Situation basierend auf ihrer internen Repräsentation der Welt lösen können. **Detterman1993** kritisiert an dieser Studie die Verwendung von Transfer. So erklärt **judd1908** einem Teil der Personen das zugrunde liegende Prinzip. Dies ist laut **Detterman1993** jedoch äquivalent, wie wenn man den Personen sagen würde, dass sie dieses Prinzip verwenden sollen. Was dann identisch wäre, wie wenn man einer Anleitung folgen würde.

Gick und Holyoak 1980

Eine weitere bedeutende Studie zu Transfer wurde von **Gick1980** durchgeführt. Dabei wurde untersucht, unter welchen Bedingungen Lernende Analogien verwenden, um strukturell ähnliche Probleme zu lösen. Ein Beispiel Problem welches sie den Lernenden gaben war, wie kann ein Tumor mit Strahlung zerstört werden, ohne dass gesundes Gewebe geschädigt wird. Dieses Problem wurde erstmals von **Duncker1945** verwendet. Dieses Problem kann gelöst werden, indem man mehrere Strahlen verwendet, welche sie nur im Tumor überlagern. Bevor sie dieses Problem lösten bekommen die Lernenden eine Geschichte erzählt, bei welcher das gleiche Prinzip verwendet wird. In dieser Geschichte ging es darum ein Fort, welches von Minen umgeben ist zu erobern. Durch aufteilen der Angreifer in mehrere angreifende Gruppen, die unterschiedliche Wege gehen, wurde die Belastung auf die Minen reduziert und das Fort konnte erobert werden. Das Resultat dieser Studie zeigte, dass spontaner Transfer nur sehr selten stattfindet. Das Hören der Geschichte führt nicht zu einer höheren Wahrscheinlichkeit das zweite ähnliche Problem zu lösen, solange die Lernenden nicht auf die Ähnlichkeit aufmerksam gemacht werden.

2.1.2 Kritiken

Lave1988 kritisiert die verschieden hier vorgestellten Untersuchungen. Da bei allen angenommen wird, das Wissen automatisch generalisierbares Wissen erzeugt, welches auf verschiedene Situationen angewendet werden kann. Sie schlägt eine Alternative vor welche sie als “practice view“ bezeichnet. Bei dieser wird Wissen von Personen erworben, welche an speziellen Übungen teilnehmen und daraus nur Wissen entwickelt wird, welches auf diese spezifische Situation(Kontext) zutrifft.

Folgende Kritiken erhebt sie. So stellt sie die Frage was lernen die Teilnehmer der verschiedenen Studien überhaupt. So greift sie insbesondere die Annahme an, dass die Teilnehmer der Studien Kontext unabhängig lernen. Sie lernen immer Kontext spezifisch. Ein anderer Punkt welche Sie angreift ist, wer definiert die Ähnlichkeit der Probleme. Ist die Ähnlichkeit der Probleme für die Teilnehmer der Studie auch greifbar. Auch **Detterman1993** kritisiert die Studien. So sollten seiner Meinung alle Studien zu Transfer als Doppel-Blind Studien durchgeführt werden, da der Studienleiter unbewusst die Leistung der Probanden ändern kann. **Detterman1993** fordert daher:

No tranfer experiment should be carried out without using a double blind procedure, particularly experiments assessing general transfer **Detterman1993**.

2.1.3 Lobato und Sievert 2002

Nachdem einige historische Studien zu Transfer exemplarisch aufgezeigt wurden, soll eine aktuelle Studie zu Transfer welche auf die Kritiken eingeht gezeigt werden.

Lobato2002a möchte einen Kritik-Punkt von **Lave1988** lösen. So kritisierte **Lave1988** dass der Untersucher festlegt was Transfer von Wissen ist. Daher legten **Lobato2002a** als Messung für Transfer fest, welche Ähnlichkeit der Proband selbst zwischen verschiedenen Situationen zieht. So untersuchten Sie einen Schüler, welcher eine Rollstuhlrampe erhöhen

sollte ohne die Steigung zu verändern. Der Schüler löste dieses Problem in dem er die Verhältnisse von Höhe zu Länge konstant hielt. Er verwendete dafür jedoch nicht die im Mathematik Unterricht gelernten Formeln. In den bisherigen Untersuchungen wäre daher angenommen worden, dass der Schüler keinen Transfer geleistet hat. Aufgrund der Interviews stellte sie jedoch fest, dass der Schüler sehr wohl Transfer geleistet hat, indem er das Konzept von konstanter Geschwindigkeit als Verhältnis von zurückgelegter Strecke zur Zeit auf dieses Problem angewendet hatte.

Lobato2002a konnten damit zeigen, dass wenn man die Ähnlichkeit zwischen zwei verschiedenen Situationen(Kontexten) nicht mit strukturellen Ähnlichkeiten oder Unterschieden beschrieben werden sollen. Sondern damit, wie der Lernende die Ähnlichkeiten zwischen den Situationen(Kontexten) wahr nimmt.

2.1.4 Elemente von Transfer

Nachdem ein Überblick über die historische Entwicklung von Transfer gegeben wurde, soll nun auf die grundlegenden Elemente, welche bei Transfer anzutreffen sind eingegangen werden.

Marini1995 definieren drei Elemente, welche zu einem Transfer führen. Das erste Element besteht aus Merkmale des Lernenden. Dieser hat sobald er eine Situation antrifft bereits ein bestimmtes prozedurales und deklaratives Wissen, welches er sich erarbeitet hat und abrufen kann. In einem bestimmten Kontext kann er einen Teil davon abrufen und anwenden (**Marini1995**). Dies führt dazu das lösungsrelevante Wissen von den Vorhanden und dem verarbeitbarem Wissen abhängt. Zusätzlich kann in einem bestimmten Kontext jedoch nicht alles Wissen abgerufen werden, da man mit tragem Wissen rechnen muss und auch der aktuellen Motivation des Lernenden.

Als zweites Element von Transfer gibt **Marini1995** die Merkmale einer Aufgabenstellung an. So hängt Transfer von der Ähnlichkeit der Aufgabe ab. Dabei gibt es jedoch einen Unterschied zwischen Novizen und Experten. Novizen vergleichen Aufgaben hauptsächlich aufgrund oberflächlicher Merkmale, wohingegen Experten sich auf die zugrunde liegenden Prinzipien fokussieren (**Marini1995**). Aufgrund dessen, haben Novizen oft Problem den Zusammenhang zwischen Aufgaben zu sehen und können daher keinen Transfer durchführen.

Das dritte Element ist der Kontext in den ein Problem eingebettet ist. Ein Beispiel dafür ist die Untersuchung von **Godden1975**. Dort lernten Taucher Wörter Unterwasser auswendig. Bei einer späteren Überprüfung konnten sie sich an mehr Wörter erinnern, wenn es Unterwasser wiederholt wurde im Vergleich zu einer Wiederholung auf dem Festland. Dieser Ortswechsel ist auch bei ausserschulischem Kontext gegeben. Aber auch innerhalb der Schule kann es zu unterschieden kommen. Ein Beispiel dafür liefert **Schoenfeld1988** so hatten Lernende keine Schwierigkeiten mit einer Divisionsaufgabe. Wenn die Aufgabe jedoch in einen Kontext gestellt wurde, wie z.B. in eine Textaufgabe eingebettet wurde, scheiterten die meisten der Lernenden.

Erst durch die Berücksichtigung aller drei Elemente lässt sich Transfer ganzheitlich Betrachten. Nicht wie **Woodworth1901**, welcher nur den Aspekt der Aufgabenmerkmale genauer untersucht hat. Erst neuere Arbeiten berücksichtigen alle Elemente und insbeson-

dere den Kontext (**Lobato2002a**; **Detterman1993**; **Greeno1996**)

2.1.5 Konsequenzen für den Unterricht

Wie **claxton1990** zeigte, darf jedoch nicht davon ausgegangen werden, dass in der Schule erworbenes Wissen ohne weiteres auf andere Alltags Probleme angewendet werden können. So sprach **Whitehead1929** von „trägem Wissen“ (inert knowledge), wenn Wissen vorhanden ist um ein Problem zu lösen, dieses jedoch nicht automatisch abgerufen werden kann. Dieses Wissen ist erst greifbar, wenn die Person angeregt wird dieses Wissen zu verwenden. Nach **Whitehead1929** entsteht träges Wissen oft unter schulischen oder universitären Bedingungen.

Detterman1993 geht sogar noch weiter und schliesst aus den Studien zu Transfer:

that, if you want people to learn something, teach it to them. Don't teach them something else and expect them to figure out what you really want them to do (**Detterman1993**).

Andere Autoren haben jedoch keine so pessimistische Sicht auf die Fähigkeit zu Transfer und geben Empfehlungen, wie schulischer Unterricht aussehen muss, welcher verhindert, dass träges Wissen entsteht und möglichst viel Transfer von Wissen stattfinden kann.

Überlernen von Fähigkeiten

Eine Möglichkeit, gute Transferleistung zu erreichen, ist das intensive einüben von Grundfertigkeiten, wie zum Beispiel in der Grundschule. **LaBerge1974** untersuchten dies bei der Fertigkeit des Lesens. Dabei wird das Üben nicht abgebrochen, wenn die Schülerinnen und Schüler die Fertigkeit subjektiv bereits können, sondern noch einige Zeit fortgesetzt. **LaBerge1974** haben dabei Schülerinnen und Schüler einen Text solange laut vorlesen lassen, bis sie keinen Fehler mehr machten und einen hohen Flüssigkeitsgrad aufwiesen. Dieses *überlernen* einer Fertigkeit fördert Transfer. So führt nach **Perkins1989** hochgradig eingeübte Fertigkeiten zu spontanem automatischem Transfer, ohne dass es längeren Nachdenkens bedarf. Der Grund dafür liegt darin, dass Routinen gebildet wurden, welche in einer neuen Situation helfen, die Aufmerksamkeit verstärkt auf neue Aspekte zu richten (**LaBerge1974**; **Mietzel2007**).

Diese Erkenntnis deckt sich mit den Forderungen vom **Whitehead1929**, welcher bereits 1929 davor warnte, dass in der Schule träges Wissen entsteht. Daher soll in der Schule darauf geachtet werden, nicht zu viel in zu kurzer Zeit zu erreichen. So fordert er auch wenige Themen gebiete gründlich zu erarbeiten. Diese Forderung wurde auch von neueren Studien bestätigt (**Porter1989**; **Brophy1992a**; **Millar1999**).

Entkontextualisieren

Wie bereits vorher angesprochen, hängt Wissen sehr stark vom Kontext ab unter welchem es gelernt wurde (**Godden1975**; **Schoenfeld1988**). **Anderson1996** fordern daher, dass Wissen, so erworben werden soll, dass Lernende lernen, irrelevante Aspekte der Situation

vom Wissensinhalt zu trennen. Dieser Prozess wird als Entkontextualisieren bezeichnet. Dadurch verliert der Lernende die Assoziation einer Aufgabe mit einem bestimmten Kontext und allmählich tritt das zugrunde liegende Prinzip hervor **Perkins1989**. Entkontextualisieren von Wissen ist jedoch nicht ausreichend, zusätzlich müssen Lernende lernen, wann und wo welches Wissen angewendet werden muss **Wiggins1993**. Diese Erkenntnis deckt sich mit den Ergebnissen von **Gick1980**, bei denen die Teilnehmer einen höheren Transfer aufwiesen, wenn auf die Ähnlichkeit der Situationen hingewiesen wurden.

Problemorientierter Unterricht

Williams1992 untersuchte viele Lernsituationen im Medizinischen Studium auf ihre Möglichkeiten zu Transfer. Sie stellt fest, dass das Wissen, welches in Vorlesungen gelernt wurde, im klinischen Teil der Ausbildung vergessen ist. Ein Grund dafür sind, dass in vielen Lehrbüchern und Vorlesungen theoretisches Wissen losgelöst von Anwendungen dargestellt werden. So werden Fragen beantwortet, welche sich Lernende nicht stellen und daher von diesen nicht auf konkrete Problemsituationen angewendet werden können. Diese Erkenntnis gilt nicht nur für Mediziner, sondern wurde auch in anderen Fachdisziplinen nachgewiesen. So bedauert **Shuell1996**, dass zukünftige Lehrpersonen faktisches Wissen lernen, anstelle von anwendungsbezogenem Wissen. So lernen sie etwas *über* das Unterrichten jedoch nichts darüber, *wie* zu unterrichten ist.

Um diese Probleme zu vermeiden wurde problemorientierte Unterrichtsgelegenheiten entwickelt und untersucht (siehe unter anderem **Barrows1985**; **Michael1993**; **Shuell1996**; **Corte2003**; **Reusser2005**; **Fassler2007**; **Pea2013b**). Das Ziel dabei ist, dass Wissen in möglichst lebensnahen Kontexten zu erwerben. Dies führt dazu, dass bei der Anwendung der Kontext ähnlich zu dem Kontext ist, unter welchem das Wissen erworben wurde.

2.1.6 Zusammenfassung zu Transfer

Es wurde in dem letzten Abschnitt versucht einen Überblick über den Begriff des Transfers zu geben und zu zeigen wie der Begriff des Kontextes damit verknüpft ist. Zuerst wurde eine historische Übersicht, über die wichtigsten Untersuchungen zu Transfer gegeben, um den Wandel des Begriffes des Transfers aufzuzeigen. Als Vorbereitung für den nächsten Abschnitt wurde noch der Begriff des Transfers elementarisiert. Darauf aufbauend wurden die Konsequenzen für den Unterricht, welcher transferierbares Wissen fördern soll zusammengefasst. Im nächsten Abschnitt geht es um den Begriff der Kompetenz und deren Verknüpfung mit dem Begriff des Transfers und Kontext.

2.2 Kompetenz

Nachdem ein Überblick über den Transfer erarbeitet wurde, soll in diesem Abschnitt versucht werden die Konsequenzen aus der Betrachtung zu Transfer mit dem Kompetenzbegriff zu verknüpfen.

2.2.1 Bildungsreformen

In den letzten Jahrzehnten fand international ein Wandel in der Bildungspolitik statt. In der Vergangenheit wurde der Fokus auf den Input des Bildungssystems gelegt. In den letzten Jahren fand eine Erweiterung der Perspektive statt und auch der Output des Bildungssystems wurde beachtet. Das Ziel dabei ist, die Qualität des Bildungssystems fassbar zu machen und soll helfen die Ressourcen effektiv einzusetzen.

Dieser Perspektivwechsel wurde von den grossen Bildungsstudien (PISA (**PISA-KonsortiumDeutschland**)), TIMSS (**Martin2003**) und IGLU (**Bos2003**) in den letzten Jahren ausgelöst. Diese führten zu einem Wandel, sowohl in der Forschung als auch in der politischen Diskussion über das Bildungssystem. So wurden die Resultate des Bildungssystems in den Vordergrund gerückt. Insbesondere die Definition von Standards und deren Verankerung im gesamten Bildungssystem sind neu. Davor wurden Standards meistens durch strukturelle Vorgaben umgesetzt (Lehrpläne, Stundentafeln und Schulorganisation). Diese Vorgaben haben einen Einfluss auf den Input des Bildungssystems. Die Qualität des Bildungssystems wurde jedoch nur sehr gering über die Überprüfung der erreichten Ergebnisse (Leistung der Schülerinnen und Schüler, Übertrittsquoten und Abschlussprüfungen) überprüft. Es wurde implizit angenommen, dass der Input einen Einfluss auf das Ergebnis des Bildungssystems als ganzes hat.

Neu ist, dass die Steuerung des Bildungssystems vermehrt über den Output erfolgen soll. So soll die Leistung des Bildungssystems messbar gemacht werden und objektiv vergleichbar. Mit den bisherigen Leistungserhebungen auf Klassen oder Schulstufe, lässt sich der Output des Schulsystems nicht akkurat beschreiben, da festgelegte Messstandards fehlten. So wurden im Zuge der Entwicklung von Bildungsstandards kompetenzbezogene Niveaus eingeführt, welche einen Aufschluss über die erreichten Kompetenzen eines Schülers oder Schülerin geben soll (**Oelkers2008**).

Diese Bemühungen führten in vielen Ländern zur Entwicklung von neuen Bildungsstandards (**Berner2006**). In der Schweiz wurde dies von der **EDK Schweizer Konferenz der Kantonalen Erziehungsdirektoren** unter dem Title „Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat)“ angestossen. In Deutschland wurde neue Bildungsstandards von der **Kultusministerkonferenz2004** verabschiedet.

Im englischsprachigen Raum fanden diese Diskussionen bereits früher statt. So wurde in Neuseeland bereits zu Beginn der 1990er Jahren ein „Outcome-based“ Curriculum verabschiedet (**McGee1996**). Auch in Australien wurde ein ähnliches Bildungskonzept 2000, unter dem Name „outcomes-based education (OBE)“, umgesetzt (**Killen2000**). Auch in England wurde zu Beginn des Jahrtausends Bildungsreformen gefordert (**Millar1999**), welche dann um 2005 umgesetzt wurden (**Huber2006**).

2.2.2 Definition von Kompetenz

Der Begriff der Kompetenz wird im Moment sowohl fachlich als auch politisch sehr stark diskutiert. So spricht **Weinert2001b** von einer Inflation des Kompetenzbegriffes. Der Begriff der Kompetenz, welcher in den Bildungsstandards (sowohl der Schweiz als auch von Deutschland) verwendet wird, basiert auf der Arbeit von **Klieme2004**. **Klieme2004**

unterscheidet verschiedene Varianten des Kompetenzbegriffes:

1. Kompetenz als kognitive Leistungsdisposition, welche es Personen erlaubt unterschiedliche Aufgaben zu lösen.
2. Kompetenz als kontextspezifische kognitive Leistungsdisposition, welche sich auf spezifische Kontexte bezieht. Dieser Kompetenzbegriff wird oft mit Kenntnisse, Routinen oder Fertigkeiten bezeichnet.
3. Kompetenz als motivationaler Orientierungen, welche notwendig ist um eine Aufgabe zu bewältigen.
4. Handlungskompetenz, als Integration der vor-gängigen Kompetenzbegriffe, im Bezug auf die Anforderungen eines genau definierten Handlungskontextes.
5. Metakompetenzen als Strategiewissen oder Motivation, welche die Anwendung und den Erwerb anderer Kompetenzen erleichtert.
6. Schlüsselkompetenzen als generalisierbare kontextspezifische kognitive Leistungsdispositionen. Das heisst Kompetenzen, welche auf viele verschiedene Situationen angewendet werden können, wie zum Beispiel mathematische oder sprachliche Kenntnisse.

Abgrenzung von Kompetenz und Intelligenz

Problematisch an der Definition des Kompetenzbegriffes ist, dass der Kompetenzbegriff schwierig von der Definition der allgemeinen Intelligenz zu unterscheiden ist, insbesondere der erste Punkt. **Weinert2001b** empfiehlt daher eine Einschränkung des Kompetenzbegriffes. So sollen Kompetenzen auf einen eingeschränkten Raum von Kontexten und Situationen bezogen werden und allgemeine intellektuellen Fähigkeiten ausgeschlossen werden. Dies begründet **Weinert2001b** damit, dass allgemeine intellektuelle Fähigkeiten eine Grundausstattung des Menschen sind und nicht erworben werden können und daher nur sehr begrenzt trainiert werden können. Zusätzlich schränkt er den Kompetenzbegriff weiter ein, indem er affektive und motivationale Aspekte nicht einbezieht. Der Begriff der Kompetenz soll daher auf spezifische Kenntnisse angewendet werden, welche notwendig sind, um genau definierte Ziele zu erreichen. Dies führt zu einer Abgrenzung zum Intelligenzkonzept, da damit Fähigkeiten assoziiert werden, welche ohne spezifisches Vorwissen auf neue Problemstellungen angewendet werden sollen. Daher ist der Begriff der Kompetenz stärker mit spezifischen Kontexten verbunden, während die Intelligenz sich generalisieren lässt. Dies führt jedoch zu einem weiteren Problem, da bei breiteren Kontexten die Abgrenzung zwischen Kompetenz und Intelligenz schwieriger wird (**Hartig2006**).

Ein weiterer Unterschied zwischen dem Kompetenz- und dem Intelligenzkonzept beruht auf der Lernbarkeit. So fordert **Baumert2001**, dass Kompetenzen „prinzipiell erlernbare, mehr oder minder bereichsspezifische Kenntnisse und Strategien“ sind. Dies bedeutet, dass Kompetenzen durch schulischen Unterricht gefördert und erweitert werden können. Daher sind dies Leistungen, welche durch den Schulbesuch verbessert werden sollten und sind daher für das Bildungsmonitoring von Interesse. Intelligenz wird hingegen als relativ stabil betrachtet, da Intelligenz hauptsächlich von genetischen Faktoren abhängt (**Shakeshaft2013**). Daher sollte theoretisch der Schulbesuch keine direkte Verbesserung der Intelligenzleistung zur Folge haben. Dies führt auch zu einem weiteren Unterschied

zwischen der Intelligenz und der Kompetenz. So kann die Kompetenz in einem bestimmten Bereich bei null liegen, da die Erfahrungen, welche zu dem Erwerb der Kompetenz führen noch nicht gemacht wurden. Bei der Intelligenzleistung ist dies nicht möglich, da jeder Mensch sich diese Grundfertigkeiten irgendwo angeeignet haben sollte.

Des Weiteren gibt es bei Erstellung von Leistungsmessungen Unterschiede. Bei Intelligenztests werden bestimmte Primärfaktoren verwendet, z.B. dreidimensionales Denken, Gedächtnis, usw., welche Unterschiede zwischen einzelnen Personen aufzeigen sollen. Kompetenzen hingegen werden durch die Anforderungen definiert (**Rychen**). In anderen Worten, Kompetenzen werden durch die relevanten Aufgaben definiert, welche von den Untersuchten gelöst werden sollen. So werden in HarmoS Kompetenzen in einem dreidimensionalen Modell definiert, Themengebiete, Kompetenzaspekt und Kompetenzniveau (**KonsortiumHarmonSNaturwissenschaften+2010**). Die Kompetenzaspekte werden nicht wie bei der Intelligenz über psychischen Prozesse definiert, sondern aus spezifischen Anforderungen in spezifischen Kontexten abgeleitet. Diese Unterschiede führen dann auch zu einer unterschiedlichen Konstruktion von Leistungstests. Intelligenztests sollten so konstruiert sein, dass möglichst wenig Vorwissen für das Lösen von Aufgaben notwendig ist. Tests, welche auf Kompetenzen abzielen, wie z.B. PISA, wurden mit dem Ziel entwickelt, Aufgaben in realitätsnahen Kontexten zu stellen.

Trotz all dieser Unterschiede werden typischerweise hohe Korrelation zwischen Intelligenz- und Kompetenzleistungen festgestellt. So fand **Rindermann2006** meistens eine sehr hohe Korrelation (>0.7) zwischen Kompetenzen und anderen Massen für kognitive Fähigkeiten. Es ist aber mit den Resultaten von PISA nicht möglich festzustellen ob dies daran liegt, dass Schüler und Schülerinnen eine höhere Kompetenz haben, weil sie intelligent sind, oder ob sowohl Intelligenz als auch Kompetenz durch die Schulbildung geprägt werden (**Hartig2006**).

2.2.3 Kompetenz und Transfer

Der Kompetenzbegriff wird in der Literatur sehr unterschiedlich definiert (**Klieme2004; Weinert2001b**). Dennoch ist die Definition des Kompetenzbegriffes für internationale Studien wie PISA (**PISA-KonsortiumDeutschland2004**), TIMSS (**Martin2003**) und IGLU (**Bos2003**) elementar.

Interessant ist, dass trotz der Einschränkung des Kompetenzbegriffes auf spezifische Kontexte, immer noch davon ausgegangen wird, dass die Kompetenz generalisierbar ist und teilweise auf andere Situationen übertragen werden kann **Hartig2006**.

Im Abschnitt 2.1.5 wurde herausgearbeitet, welche Eigenschaften von Unterricht zu besserer Transferleistung führen kann. Auch **Lersch2007** gibt Vorschläge wie kompetenzfördernder Unterricht gestaltet sein sollte. So fordert er, dass der Unterricht „viel stärker von den erforderlichen Lernprozessen und -gelegenheiten her konzipiert werden müsste und eben nicht nur von einer kontinuierlichen Abfolge von Inhalten“. Dies deckt sich mit der Forderung von **Mietzel2007** für das Entkontextualisieren von Unterricht um Transferleistung zu fördern. Auch die Problemorientierung von Lerngelegenheiten wird von **Lersch2007** für kompetenzfördernden Unterricht als wichtig gehalten, insbesondere fordert er, dass „systematische Wissensvermittlung [...] um variable Anwendungssituationen“ ergänzt werden soll-

ten. Zusätzlich fordert er, dass realistische Lernsituationen angeboten werden sollten, in anderen Worten: die Lerngelegenheiten sollten mit dem Ziel der Kompetenz übereinstimmen, da der Erwerb der Kompetenz ja kontextspezifische erfolgt (**Klieme2004**).

2.3 Kompetenz des skalenbasiertes Messens

Im Rahmen von ExKoNawi hands-on Experimentiertests wurde ein Modell entwickelt um verschiedene hands-on Kompetenzen von Schülerinnen und Schüler auf der Sekundarstufe I in der Schweiz zu messen werden (**Metzger2013**). Einer der Kompetenzen welche mit ExKoNawi hands-on Experimentiertests gemessen werden soll, ist die Kompetenz des *skalenbasiertes Messen* (**Gut2013a**). Die Definition dieser Kompetenz basiert auf der Arbeit von **Munier2013**. In dieser Kompetenz geht nach (**Gut2013a**) darum „quantitative Grössen mit gegebenen Messinstrument genau [zu] messen“. Bei dieser Kompetenz gibt es drei Teilbereiche die eine wichtige Rolle spielen. Zum einen müssen die Schülerinnen und Schüler entscheiden, welches Messinstrument besser für eine Messung geeignet ist. Ein weiterer Teilaspekt ist, dass sie die Messung mehrmals wiederholen um eine genauere Abschätzung des Resultates bekommen. Zusätzlich zu diesen Aspekten müssen sie auch das Messinstrument korrekt verwenden (**Munier2013; Gut2013a**).

Ein wichtiger Aspekt der Kompetenz des skalenbasierten Messens ist, dass diese Kompetenz ohne einen inhaltlichen oder fachlichen Kontext definiert wurde. Was bedeuten sollte, dass das Kompetenzniveau, welches ein Schüler oder eine Schülerin erreichen könnte unabhängig des fachlichen oder inhaltlichen Kontextes sein müsste, in welcher die Messung der Kompetenz stattfinden sollte.

2.4 Forschungsfrage

Dieses Modell führt nun daher zur Frage, ist das erreichbare Kompetenzniveau von Schülerinnen und Schüler tatsächlich unabhängig des inhaltlichen und fachlichen Kontextes? Insbesondere auch aus dem Aspekt, dass sich kontextspezifische Kompetenzen und Transferleistungen grundsätzlich nicht gegenseitig ausschliessen. Guter kompetenzorientierter Unterricht unterstützt hingegen sogar die Fähigkeiten das Wissen zu transferieren. Diese Erkenntnis führt nun jedoch zu der Frage:

Ist eine gewisse Kompetenz (hier skalenbasiertes Messen) von Lernenden auf der Sekundarstufe I in unterschiedlichen Kontexten gleich verfügbar?

Diese Frage verknüpft sehr stark den Begriff des Transfers mit dem Kompetenzbegriff. Im Bezug auf den Transferbegriff müssen Lernende eine Transferleistung erbringen, da sie diese gewisse Kompetenz auf verschiedene Kontexte anwenden müssen. Um diese Frage zu beantworten soll in der vorliegenden Arbeit untersucht werden ob die erreichten Kompetenzniveaus des skalenbasierten Messens unabhängig des fachlichen oder inhaltlichen Kontextes sind.

3 Untersuchungsanlage

3.1 Anforderungen

Um die vorliegende Fragestellung zu beantworten, ist es notwendig mehrere Test zu verwenden, welche die Kompetenz des skalenbasierten Messens messen. Zusätzlich müssen die Tests die Kompetenz des skalenbasierten Messens unter verschiedenen Kontexten messen. Es wurden zwei existierende Test aus dem ExKoNawi Projekt verwendet. Der eine war aus dem Fachbereich Chemie, bei welchem eine Temperatur gemessen werden musste. Der zweite Test war aus dem Fachbereich Physik, bei dem eine Kraft gemessen wurde. Zusätzlich wurde ein dritter Test neu entwickelt, bei welchem eine Temperatur Messung im Fach Physik durchgeführt wurde. Die Testerstellung wird genauer in **Sichau2015** detailliert beschrieben. Der dritte Test wurde so entworfen, damit einmal der inhaltliche Kontext verändert werden kann (Kraftmessung versus Temperaturmessung), bei gleichem fachlichen Kontext und zum anderen der fachliche Kontext verändert werden kann ohne den inhaltlichen Kontext zu verändern.

3.2 Umsetzung

Die Tests wurden zusammen mit einem Fragebogen an vier Klassen der Sek 1 A durchgeführt. In jeder Klasse wurden vier Gruppen gebildet, welche die Tests in unterschiedlicher Reihenfolge durchführten. Dafür gab es zwei Gründe. Zum einen war nur Material für 11 Tests verfügbar. Daher konnten die Tests nicht in voller Klassenstärke durchgeführt werden. Dies führte zur Bildung von zwei Gruppen, bei welcher eine zuerst den Fragebogen ausfüllte und die andere Gruppe den Fragebogen am Ende ausfüllte. Zusätzlich wurde noch der zweite und dritte Test in jeder Gruppe vertauscht um zu untersuchen ob Müdigkeit oder die Wiederholungen Einfluss auf die Test-Ergebnisse haben. Die Tabelle 3.1 gibt eine Übersicht über die Gruppeneinteilung der Schülerinnen und Schüler innerhalb einer Klasse an.

Gruppe FABC	Gruppe FACB	Gruppe ABCF	Gruppe ACBF
Fragebogen (F)	Fragebogen (F)	Temperatur Physik 305 (A)	Temperatur Physik 305 (A)
Temperatur Physik 305 (A)	Temperatur Physik 305 (A)	Kraft Physik (B)	Temperatur Chemie (C)
Kraft Physik 301 (B)	Temperatur Chemie 201 (C)	Temperatur Chemie 201 (C)	Kraft Physik 301 (B)
Temperatur Chemie 201 (C)	Kraft Physik 301 (B)	Fragebogen (F)	Fragebogen (F)

Tabelle 3.1: Aufteilung der Gruppen, innerhalb einer Klasse

Die Namen der Gruppen aus Tabelle 3.1 wurden auch für die Kodierung der Tests verwendet, sodass jeder Test einer Gruppe zuordenbar ist.

Die vier Klassen waren alle von derselben Schulstufe (7. Schuljahr) jedoch in verschiedenen Gemeinden. Die Klassen in Glattbrugg hatten beide dieselbe Lehrperson, alle anderen Klassen hatten eine unterschiedliche Lehrpersonen. Einen Überblick über die wichtigsten Daten zu den einzelnen Klassen befindet sich in Tabelle 3.2.

	Klasse 1	Klasse 2	Klasse 3	Klasse 4
Ort	Glattbrugg	Glattbrugg	Stadt Zürich	Stadt Schaffhausen
Anzahl SuS	15	13 (+1 nur einen Test)	22	22
Datum der Durchführung	6.11.14	6.11.14	12.11.14	11.12.14
Uhrzeit	8:20-10:00	10:20-12:00	10:20-12:05	13:15-14:45
Versuchsleiter	Pitt Hild und David Sichau	Pitt Hild und David Sichau	Pitt Hild und David Sichau	Martina Minges und David Sichau

Tabelle 3.2: Aufteilung der Gruppen, innerhalb einer Klasse

Alle Klassen wurden für die Durchführung in zwei Gruppen aufgeteilt. Zum einen konnten so die Schülerinnen und Schüler mit mehr Abstand positioniert werden um die Ablenkung zu reduzieren. Andererseits konnten so die Schülerinnen und Schüler, welcher der Videoaufnahme nicht zugestimmt wurden in ein Zimmer gesetzt werden, wo keine Videoaufnahme stattgefunden hat. Die Erlaubnis zur Videoaufnahme wurde im Vorhinein zur Durchführung von den Klassenpersonen organisiert und eingesammelt.

3.3 Durchführung

3.3.1 Vorbereitung

Für die Durchführung in den einzelnen Klassen wurden alle Tests in Boxen vorbereitet, sodass zwischen den Tests nur die Boxen ausgetauscht werden mussten. In jeder Box waren alle Materialien, welche für die Durchführung des Versuches notwendig waren vorbereitet, sodass die Schülerinnen und Schüler alle notwendigen Materialien in dieser Box finden konnten.

Zusätzlich wurden die Auswertungsbögen in der richtigen Reihenfolge und bereits mit einer Kodierung versehen, in einem Schnellhefter bereitgestellt. Ein für die Durchführung vorbereiteter Klassenraum ist im Bild 3.1 ersichtlich.

Im Bild 3.1 sieht man auch gut, wie die Kamera für die Videoaufwertung aufgestellt wurde. Die Videoaufnahme wurde bevor die Schülerinnen und Schüler den Klassenraum betreten haben gestartet, um die Ablenkung durch die Kamera möglichst gering zu halten.



Abbildung 3.1: Klassenzimmer für die Durchführung des ersten Durchganges vorbereitet.

3.3.2 Durchführung

Nachdem die Schülerinnen und Schüler in die beiden Räume aufgeteilt wurden, wurden Sie von den Versuchsleitern jeweils begrüßt. Die Begrüssung war stichwortartig vorbereitet, damit alle Klassen die gleichen Informationen erhielten und durch die Begrüssung die Testergebnisse nicht beeinflusst werden. Dabei wurde darauf hingewiesen, dass die Experimente keine Leistungskontrolle darstellt und alle Ergebnisse anonymisiert sind. Es wurde auch ein grober Überblick über den Ablauf gegeben. Im Raum in dem eine Videoaufnahme gemacht wurde, wurden die Schülerinnen und Schüler darüber informiert.

Nach der Begrüssung wurden die Schülerinnen und Schüler aufgefordert mit den Tests anzufangen. Während der Zeit in welcher die Tests durchgeführt wurden, gaben die Versuchsleiter jeweils kurze Zeit Informationen und forderten die Schülerinnen und Schüler auf ihre Ergebnisse zu verschriftlichen.

Nach dem ersten Test (nach 20 Minuten) wurde eine Pause von fünf Minuten durchgeführt, in dieser wurden die Boxen ausgetauscht, sodass alle Schülerinnen und Schüler den nächsten hands-on Experimentiertest vor sich hatten. Die Schülerinnen und Schüler wurden aufgefordert sich innerhalb des Klassenraumes zu bewegen. Nach dem zweiten Test wurde eine grosse Pause durchgeführt, in welcher die Schülerinnen und Schüler das Schulzimmer verlassen hatten. Nach dem dritten Test wurde wieder eine kurze fünf-minütige Pause durchgeführt. Während den Test wurden den Schülerinnen und Schülern nur Fragen bei Unklarheiten beantwortet, inhaltliche Fragen oder Fragen zum korrekten Vorgehen wurden nicht beantwortet.

3.3.3 Nachbereitung

Nachdem die Tests durchgeführt wurden, wurden die Auswertungsbögen eingesammelt und von David Sichau erst kodiert. Es wurde eine Zweitkodierung vor 15 % der Auswertungsbögen von Pitt Hild durchgeführt. Die 11 Auswertungsbögen zur Zweit-Kodierung wurden per zufällig (random generator) ausgewählt, um sicherzugehen das ein Bias ausge-

schlossen werden kann. Insgesamt wurden 72 Auswertungsbögen vollständig ausgefüllt.

Die Videoaufnahmen wurden geschnitten, sodass nur noch die einzelnen Tests sichtbar sind. Dies wurde gemacht um zu vermeiden, dass Aktionen der Schülerinnen und Schüler in der Pause einen Einfluss auf die Bewertung in der Test Situation haben. Insgesamt ist Material zu 8 Schülerinnen und Schüler verwertbar, da die andern zu weit entfernt sind und daher ihre Aktionen nicht beobachtbar waren.

4 Ergebnisse

4.1 Kodierung

Wie bereits geschrieben wurde die Erstkodierung von David Sichau durchgeführt. Es wurden eine Zweitkodierung von 15 % zufällig ausgewählten (per Random Generator) Auswertungsbögen von Pitt Hild durchgeführt. Es wurden dabei die gleichen Kodierschemata verwendet, welche sich im Anhang der Arbeit befinden.

4.1.1 Items

Es gab insgesamt elf Items welche nach dem Kodierschemata kodiert wurden.

Die Items wurden auf Interrater-Reliabilität untersucht. Dafür wurden die prozedurale Übereinstimmung p_0 und zusätzlich noch das ungewichtete Cohen's Kappa κ als Zufalls-korrigierter Koeffizient berechnet. Bei einem Teil der Datensätze war dies mathematisch nicht möglich (Division durch 0), daher können nicht für alle Items ein Cohen's Kappa angegeben werden. In Tabelle 4.1 sind alle Ergebnisse zusammengefasst.

Code erhältlich auf:

GitHub

<http://git.io/mk9z-Q>

Item	201		301		301	
	p_0	κ	p_0	κ	p_0	κ
1.1	1	1	0.91	0.74	0.91	0.79
1.2	0.91	0.81	1	/	1	1
2.1	0.81	0.67	0.81	0.74	1	1
3.1	1	1	0.91	0.81	1	1
3.2	1	/	1	1	0.91	0.82
4.1	0.91	0.79	0.81	0.65	0.91	0.81
4.2	0.91	0.62	0.91	0.79	0.91	0.74
4.3	1	/	1	/	1	/
4.4	1	/	1	/	1	/
5.1	1	/	1	/	1	/
5.2	0.91	/	1	1	0.91	0.78

Tabelle 4.1: Übereinstimmung der Kodierungen für die einzelnen Items (p_0) und Cohens Kappa κ . Für die drei Tests 201 (Chemie-Temperatur), 301 (Physik Kraft) und 305 (Physik Temperatur)

4.1.2 Qualitätsstandards

Aus den elf Items wurden fünf Qualitätsstandards entwickelt (**Hild2014a**). Es gibt bedingte und unbedingte Qualitätsstandards. Bei den bedingten Qualitätsstandards ist für das Erreichen notwendig, dass sowohl die Bedingung erfüllt ist, als auch dass der vorgängige Qualitätsstandard erfüllt ist. Die unbedingten Qualitätsstandards werden in dieser Arbeit mit Q1 bis Q5 bezeichnet. Die bedingten Qualitätsstandards werden mit QS1 bis QS5 bezeichnet.

Qualitätsstandard 1

Im Qualitätsstandard 1 geht es um das korrekte und präzise messen. Dieser Qualitätsstandard wird nur erreicht wenn Item 1.1 (Richtige Tendenz des Resultates) und Item 1.2 (Ist das Resultat vollständig und korrekt) zusammen mindestens 1 ergeben.

Qualitätsstandard 2

Bei Qualitätsstandard 2 wird die Dokumentation der Messung bewertet. Dieser Qualitätsstandard wird nur erreicht wenn Item 1.2 (Werden alle Messungen und Messergebnisse vollständig dargestellt) mindestens den Wert von 2 erreicht hat.

Qualitätsstandard 3

Im dritten Qualitätsstandard wird das Begründen des richtigen Messinstrumentes bewertet. Dieser Standard wird nur erreicht wenn Item 3.1 (Ist das Korrekte Messinstrument gewählt worden) und Item 3.2 (Wird die Wahl des Messinstrumentes korrekt begründet) zusammen zwei ergeben.

Qualitätsstandard 4

Qualitätsstandard 4 beurteilt die Messwiederholung. Es wird aus Item 4.1 (mehrmaliges Messen), 4.2 (identische Messung), 4.3 (wurde Mittelwert gebildet) und 4.4 (korrekter Mittelwert) gebildet. Dieser Level wird erreicht wenn die Items addiert mindestens zwei ergeben.

Qualitätsstandard 5

Der letzte Qualitätsstandard 5 zeigt auf, inwiefern die Schülerinnen und Schüler Fehlerquellen der Messung begründen können. Dieser Standard besteht aus Item 5.1 (Fehlerkategorien nennen) und 5.2 (Verbesserungsvorschläge) welche zusammen mehr als eins ergeben müssen.

Erreichte Qualitätsstandards

In Tabelle 4.2 wird ein Überblick über die erreichten Qualitätsstandards aller Schülerinnen und Schüler gegeben. Zusätzlich werden auch die bedingten Qualitätsstandards angegeben, welche nur erreicht werden können, wenn der vorhergehende Qualitätsstandard erreicht wurde.

Test	p_{Q1}	p_{QS1}	p_{Q2}	p_{QS2}	p_{Q3}	p_{QS3}	p_{Q4}	p_{QS4}	p_{Q5}	p_{QS5}
201	0.51	0.51	0.34	0.27	0.05	0.04	0.08	0.03	0.16	0.03
301	0.62	0.62	0.31	0.31	0.09	0.04	0.09	0.01	0.39	0.01
305	0.72	0.72	0.30	0.29	0.35	0.14	0.11	0.01	0.50	0.01

Tabelle 4.2: Zusammenfassung der erreichten Qualitätsstandards, wobei $p_{Q1} - p_{Q5}$ den unbedingten Qualitätsstandards entsprechen. Die bedingten Qualitätsstandards werden mit $p_{QS1} - p_{QS5}$ bezeichnet.

4.1.3 Niveau

Basierend auf den Qualitätsstandards wurden zwei Niveaus gebildet, welche das erreichte Niveau der Schülerinnen und Schüler bei der Kompetenz des skalenbasierten Messens bezeichnen. Die Niveaus können einen Wert zwischen 0 und 5 annehmen. Eine Übersicht über die erreichten Niveaus wird in Tabelle 4.3 gegeben.

Code erhältlich auf:

GitHub

<http://git.io/bjn9qg>

Test	uLev						kLev					
	0	1	2	3	4	5	0	1	2	3	4	5
201	0.36	0.24	0.22	0.13	0.03	0.03	0.40	0.24	0.32	0.01	0	0.03
301	0.31	0.21	0.29	0.14	0.03	0.03	0.42	0.28	0.26	0.01	0	0.03
305	0.13	0.19	0.24	0.31	0.11	0.03	0.22	0.43	0.18	0.13	0.01	0.03

Tabelle 4.3: Prozedural erreichte Niveaus aller Schülerinnen und Schüler. Das unbedingte Niveau wird mit uLev und das bedingte Niveau mit kLev bezeichnet.

Unbedingtes Niveau

Dieses Niveau ist der Summenscore der einzelnen unbedingten Qualitätsstandards. In der Arbeit wird dieses Level mit *uLev* abgekürzt.

Bedingtes Niveau

Dieses Niveau ist der Summenscore der bedingten Qualitätsstandards. Dieses Niveau wird mit *kLev* abgekürzt.

4.2 Fragebogen

Im standardisierten Teil des Fragebogens wurden Fragen zum absoluten Selbstkonzept nach SESSKO gestellt (**Schone2002**). Die verwendeten Fragen sind in Tabelle 4.4 aufgeführt.

Skala	Frage	α_d
SESSKO 18(a)	Ich bin für die Schule sehr begabt.	0.71
SESSKO 19(a)	Neues zu lernen fällt mir schwer.	0.76
SESSKO 20(a)	Ich bin sehr intelligent.	0.71
SESSKO 21(a)	Ich kann in der Schule viel.	0.72
SESSKO 22(a)	In der Schule fallen mir viele Aufgaben schwer.	0.74

Tabelle 4.4: Fragen von SESSKO zur Skala „Schulisches Selbstkonzept - absolut“ (**Schone2002**). α_d bezeichnete das standardisierte Cronbach Alpha wenn dieses Item weggelassen würde.

Zusätzlich wurden nach **Dierks2014** Fragen zum Selbstkonzept zu Schulversuchen entwickelt und angepasst. Die entwickelten Fragen sind in Tabelle 4.5 aufgezeigt.

Kürzel	Frage	α_d
NatSK1	Schulversuche liegen mir nicht besonders.	0.65
NatSK2	Schulversuche würde ich viel lieber machen, wenn sie nicht so schwer wären.	0.69
NatSK3	Schulversuche fallen mir schwerer als vielen meiner Mitschüler/innen.	0.65
NatSK4	Bei manchen Schulversuche weiss ich gleich: „Das verstehe ich nie.“	0.65
NatSK5	Für Schulversuche habe ich einfach keine Begabung.	0.63
NatSK6	Mit den Aufgaben bei Schulversuche komme ich besser zurecht als viele meiner Mitschüler/innen	0.67
NatSK7	Ich denke, ich bin für Schulversuche begabter als viele meiner Mitschüler/innen.	0.66

Tabelle 4.5: Fragen zum Selbstkonzept bei Schulversuchen abgewandelt nach **Dierks2014**. α_d bezeichnete das standardisierte Cronbach Alpha wenn dieses Item weggelassen würde.

Es wurde die innere Konsistenz beider Skala überprüft. Bei den der Skala „Schulisches Selbstkonzept - absolut“ wurde ein standardisiertes Cronbach α von 0.77 erreicht. Die

Anzahl vollständig ausgefüllter Fragebögen betrug dabei 69. Alle unvollständigen Items wurden vor der Analyse entfernt. Bei der Skala zum Selbstkonzept bei Schulversuchen wurde ein standardisiertes Cronbach α von 0.69 erreicht. Insgesamt konnten dabei 64 vollständige Fragebögen ausgefüllt werden.

Code erhältlich auf:

GitHub

<http://git.io/WyJH6Q>

4.3 Unterschiede zwischen den Klassen

Um festzustellen, ob alle Datensätze der einzelnen Klassen kombiniert werden dürfen wurden zuerst alle Klassen einzeln gegeneinander auf folgende Nullhypothese überprüft:

Besteht kein Unterschied in den Qualitätsstandards zwischen den einzelnen Klassen?

Es wurden dabei die Qualitätsstandards verglichen, da diese im Vergleich zu den Items ein geringeres Rauschen aufweisen, ohne jedoch gross an Informationsgehalt eingebüsst zu haben.

Aufgrund der geringen Anzahl an Beobachtungen für einzelne Qualitätsstandards wurde der exakter Test nach Fisher verwendet. Es wurden Kontingenztafeln für jeden Qualitätsstandard (Q1 bis Q5 und QS1 bis QS5) erstellt und in jeder Tafel die beiden Levels (0 und 1) gegenüber den Klassen verglichen.

Klasse	Q1	Q2	Q3	Q4	Q5	QS1	QS2	QS3	QS4	QS5
1 vs. 2	0.68	1.00	1.00	0.60	1.00	0.51	0.59	1.00	1.00	1.00
1 vs. 3	1.00	0.72	1.00	1.00	1.00	1.00	0.72	1.00	1.00	1.00
1 vs. 4	0.43	0.72	0.22	0.32	0.65	0.42	0.72	0.48	1.00	1.00
2 vs. 3	0.68	0.72	1.00	0.22	1.00	0.68	1.00	1.00	1.00	1.00
2 vs. 4	1.00	0.72	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3 vs. 4	0.43	1.00	0.22	0.10	0.65	0.43	1.00	0.48	1.00	1.00

Tabelle 4.6: p-Werte für den exakten Test nach Fisher für die Vergleiche der einzelnen Klassen untereinander auf allen Qualitätsstandards. Kein p-Wert in dieser Tabelle liegt unter 0.05.

Die Resultate des exakten Tests nach Fisher befindet sich in Tabelle 4.6. Bei keinem der 60 Tests konnte die Nullhypothese abgelehnt werden ($p \leq 0.05$). Daher gibt es keinen signifikanten Unterschied zwischen den erreichten Qualitätsstandards in den einzelnen Klassen.

Code erhältlich auf:

GitHub

<http://git.io/0DOelQ>

4.4 Korrelation der Niveaus des skalenbasierten Messens

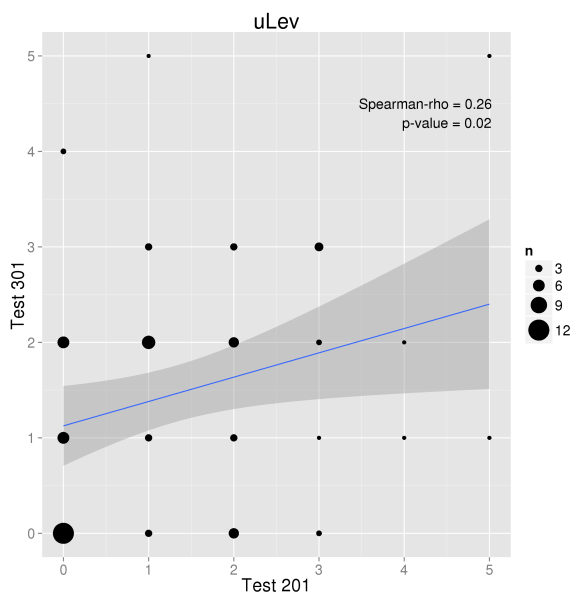
In einem nächsten Schritt wurde untersucht inwiefern die Niveau-Stufen (uLev und cLev) zwischen den einzelnen Tests korrelieren. Dazu wurde als Rangkorrelationskoeffizient Spearmans ρ berechnet. Der Vorteil dieser Methode ist, dass keine Annahmen über die zugrundeliegenden Daten gemacht werden muss. Des Weiteren bietet diese Methode den Vorteil, dass sie gegenüber Ausreißern robust ist.

Da die Korrelation alleine keinen Aufschluss darüber gibt, ob diese Korrelation signifikant ist, wurde die Korrelation zusätzlich auf Signifikanz getestet. Wichtig bei dieser Analyse ist, dass die Korrelation keine Aussage über die Kausalität zulässt.

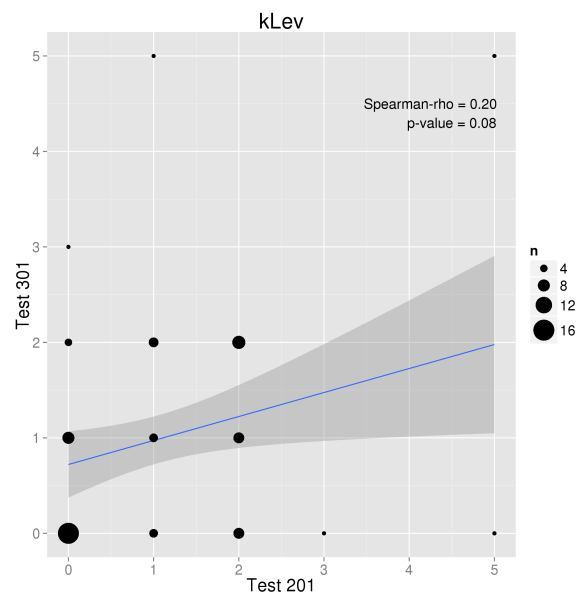
Die Ergebnisse wurden grafisch als Streudiagramme dargestellt (siehe 4.4). In die Streudiagramme wurde die Gerade der linearen Regression eingetragen mit dem zugehörigen 95% Vertrauensintervall. Zusätzlich wurde die noch Spearmans ρ und der p-Wert des Signifikanztests angegeben, diese Werte sind auch in Tabelle 4.7 zusammengefasst.

Test	uLev		kLev	
	p-Wert	ρ	p-Wert	ρ
201 vs. 301	0.02	0.26	0.20	0.08
201 vs. 305	0.44	1e-4	0.33	4e-3
301 vs. 305	0.36	2e-3	0.01	0.89

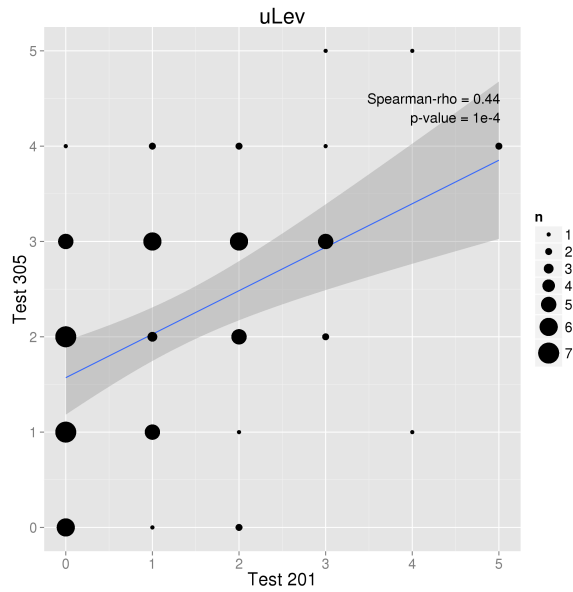
Tabelle 4.7: Spearmans ρ und p-Werte für die Korrelation zwischen den unbedingten Niveaus (uLev) und den bedingten Niveaus (kLev) zwischen den einzelnen Tests.



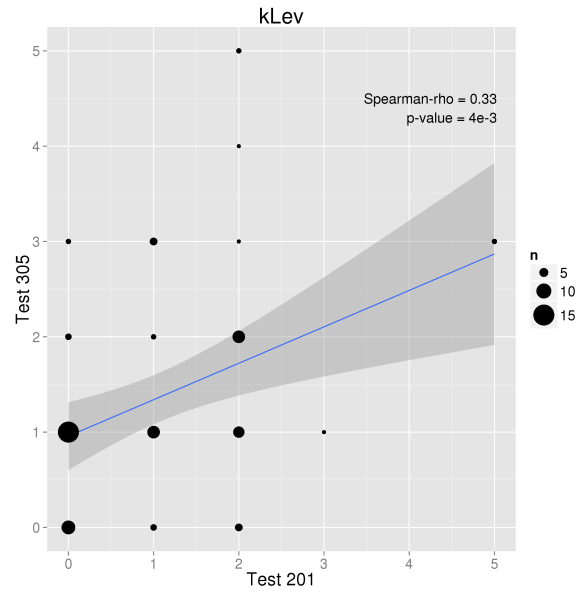
(a) unbedingte Niveau-Stufen



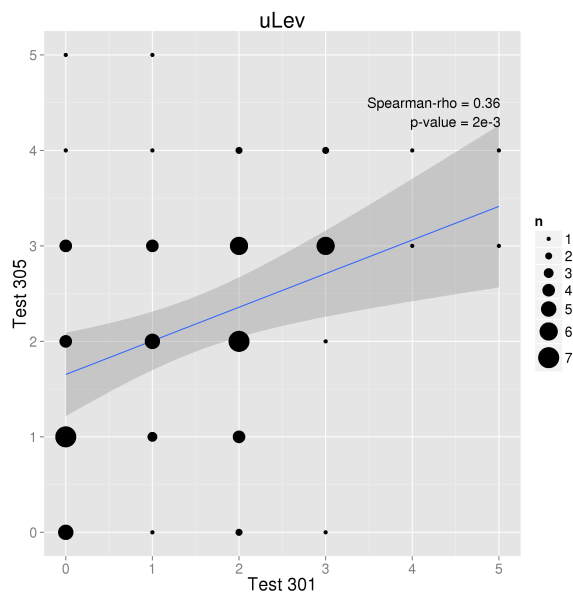
(b) bedingte Niveau-Stufen



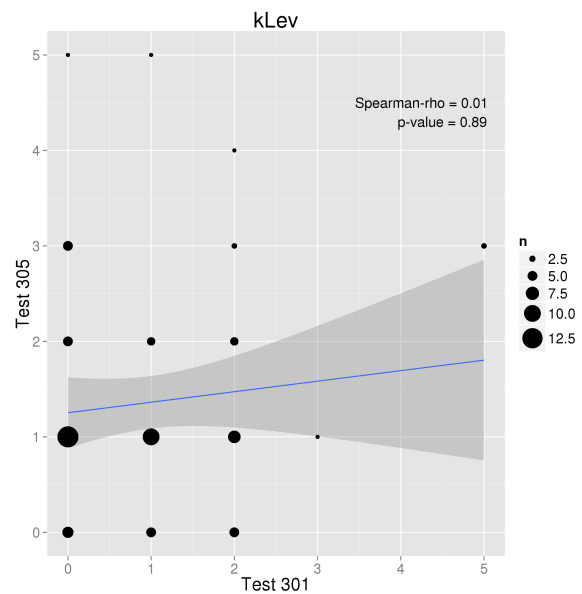
(c) unbedingte Niveau-Stufen



(d) bedingte Niveau-Stufen



(e) unbedingte Niveau-Stufen



(f) bedingte Niveau-Stufen

Abbildung 4.1: Korrelation zwischen den Niveau Stufen zwischen den einzelnen Tests. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die blaue Gerade ist die lineare Regression der zugrunde liegenden Daten, der dunkel graue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

Code erhältlich auf:

GitHub

<http://git.io/FnbD>

4.5 Rasch-Analyse

Als Probabilistische Test-Methode wurde das Rasch Modell verwendet. Der Grund für diese Methodik war, dass es sich bei der Kompetenz des skalenbasierten Messens um ein latentes Merkmal handelt. In anderen Worten die Kompetenz des skalenbaiserten Messens ist nicht direkt beobachtbar.

Es wurde zuerst folgendes Rasch Modell verwendet.

$$P(U_{ij} = u_{ij} | \theta_i, \beta_j) = \frac{e^{u_{ij}(\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}} \quad (4.1)$$

Wobei $i=1, \dots, n$ die Zählvariable für eine Person ist und $j=1, \dots, m$ die Zählvariable für eine Aufgabe darstellt. Die Variable $u_{ij} \in \{0, 1\}$ die dichotome Antwort einer Person auf eine Frage ist. Die Variable β_j beschreibt den Schwierigkeitsgrad einer Aufgabe und θ_j die latente Fähigkeit einer Person.

Bei der Item-Response-Theorie (Probabilistische Test-Methoden) wird angenommen, dass das Ergebnis einer Person nicht deterministisch ist, sondern zufällig sein kann. Daher soll mit dem Rasch Modell die Lösungswahrscheinlichkeit jeder Aufgabe U_{ij} berechnet werden. Diese Lösungswahrscheinlichkeit hängt sowohl von der Fähigkeit der Person θ_j als auch von der Schwierigkeit der Aufgabe β_i ab. Diese Lösungswahrscheinlichkeiten werden basierend auf den Testergebnissen u_{ij} geschätzt.

4.5.1 Parameterschätzung

Für die Parameterschätzung gibt es verschieden Ansätze. Da die beste Methode von den Daten abhängig ist wurde in einem ersten Schritt das Rasch-Modell sowohl mit der bedingte Maximum-Likelihood-Schätzung, als auch mit der marginal Maximum-Likelihood-Schätzung getestet und die Resultate wurden verglichen.

Bei der bedingten Maximum-Likelihood-Schätzung wird ein zweistufiges Vorgehen gewählt. Zuerst werden die Aufgaben-Parameter geschätzt ohne die Personen Parameter zu beachten. Erst in einem zweiten Schritt werden die Personen-Parameter geschätzt. Ein Problem dieser Methodik ist, dass Personenfertigkeiten von Personen, welche keine oder alle Aufgaben gelöst haben nicht geschätzt werden können (**Mair2007**).

In der marginalen Maximum-Likelihood-Schätzung wird angenommen, dass für die Personenfähigkeiten in der Stichprobe eine Normalverteilung vorliegt. Diese Annahme ist insbesondere dann problematisch, wenn nur eine Stichprobe der Gesamtbevölkerung verwendet wird (**Rizopoulos2006**).

Da beide Schätzungen für den vorliegenden Datensatz problematisch sein könnten, wurde das Rasch-Modell mit beiden Ansätzen durchgeführt und die Resultate verglichen. Das Ziel war dabei, den besseren Ansatz für den vorliegenden Datensatz zu finden, um mit diesem Ansatz die weiteren Analysen durchzuführen. Als Datensatz für diesen Vergleich wurden die 15 unbedingten Qualitätsstandards verwendet. Die Resultate sind in Abbildung 4.2 ersichtlich. Es gibt für diesen Datensatz keinerlei Unterschied in beiden Parameterschätzern.

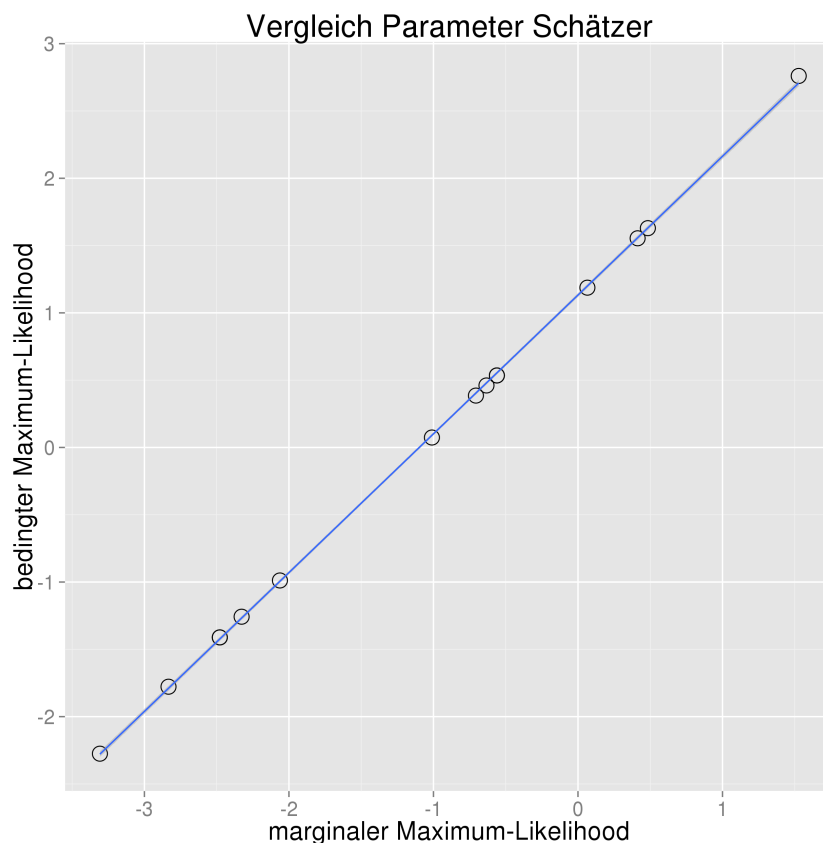


Abbildung 4.2: Vergleich des Rasch Modells mit der bedingten Maximum-Likelihood-Schätzung und der marginalen Maximum-Likelihood-Schätzung. Da alle Punkte auf einer Geraden liegen, gibt es keinen Unterschied zwischen den unterschiedlichen Schätzmethoden für den vorliegenden Datensatz der 15 unbedingten Qualitätsstandards.

Code erhältlich auf:

GitHub

<http://git.io/FRxZ>

4.5.2 Modellkontrolle des Rasch-Modells

Um das Rasch Modell zu Validieren wurde das Modell mit Hilfe des Andersens Likelihood-Quotienten Test validiert. Für alle 15 Qualitätsstufen führte dies zu Problemen und der

Test konnte nicht durchgeführt werden. Nachdem die Qualitätsstufen vier und fünf entfernt wurden, konnte das reduzierte Modell validiert werden. Als Splitkriterium wurde der Mittelwert der Personen-Randsummen verwendet.

Der p-Wert des Andersens Likelihood-Quotienten Test beträgt $p = 0.14$. Daher liegt jetzt keine signifikante Modellverletzung vor, die Aufgaben Parameter unterscheiden sich nicht signifikant für Personen mit niedrigen und hohen Randsummen. In der Grafik 4.3 sind die Resultate des Tests grafisch dargestellt. Es ist ersichtlich, dass keine Aufgabe das Modell verletzt, da die 95%-Konfidenz-Regionen alle die Diagonale berühren.

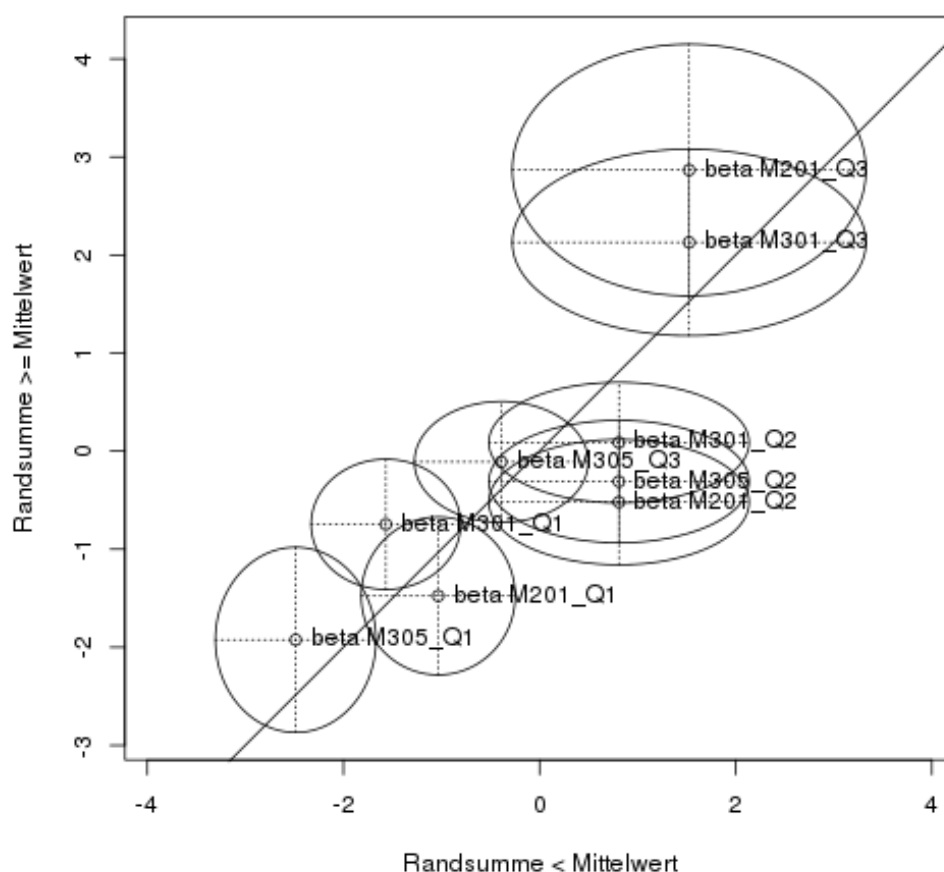


Abbildung 4.3: Modellkontrolle des Rasch-Modells: kein Qualitätsstandard hat eine signifikante Abweichung von der Diagonalen, daher gibt es keine signifikanten Unterschiede für Personen mit niedrigen und hohen Randsummen in den Qualitätsstandard.

Zusätzlich wurden die Qualitätsstandards mit dem Wald-Test überprüft. Damit können Qualitätsstandard, welche einen signifikanten Unterschied habe, identifiziert werden. In Tabelle 4.8 befinden sich die p-Werte des Wald-Test für die einzelnen Qualitätsstandards.

Test 201			Test 301			Test 305		
Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
0.44	0.08	0.24	0.11	0.33	0.56	0.38	0.14	0.61

Tabelle 4.8: p-Werte des Wald-Tests für die Qualitätsstandards, mit dem Mittelwert der Personen-Randsummen als Splitkriterium. Keine dieser p-Werte liegt unterhalb von 0.05 daher gibt es keine signifikanten Unterschiede in den Qualitätsstandards

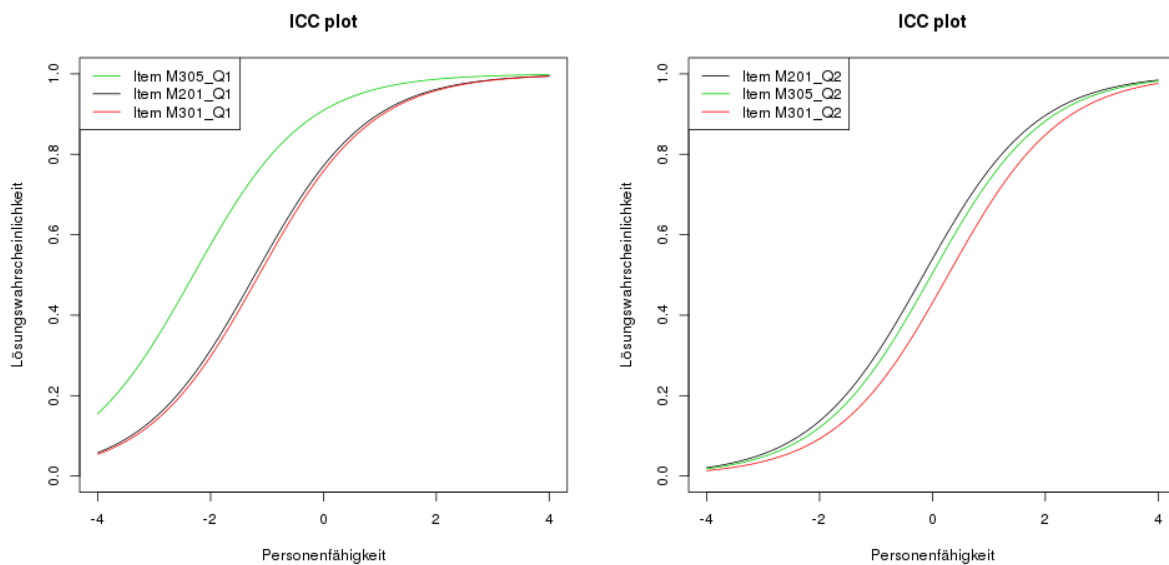
Code erhältlich auf:

GitHub

<http://git.io/FE3m>

4.5.3 Unterschied in den Qualitätsstandards

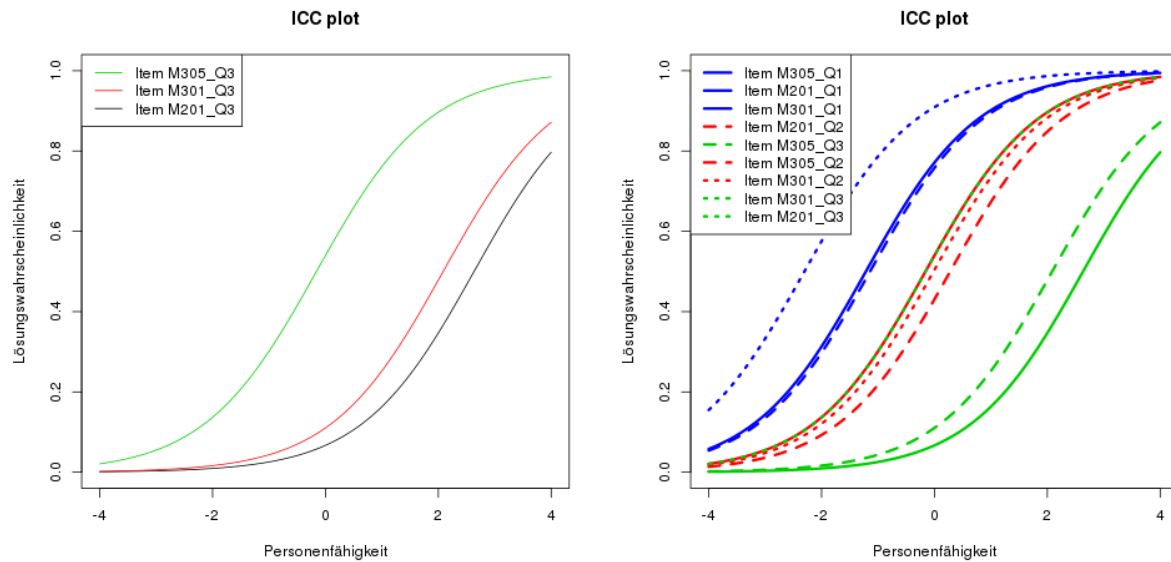
Nachdem das Modell kontrolliert wurde soll nun überprüft werden ob es einen Unterschied in den Qualitätsstandards zwischen den einzelnen Test gibt.



(a) ICC Plot für Qualitätsstandard 1

(b) ICC Plot für Qualitätsstandard 2

In Tabelle 4.9 finden sich die Aufgaben-Parameter β_j der einzelnen Qualitätsstandards.



(c) ICC Plot für Qualitätsstandard 3

(d) ICC Plot für Qualitätsstandard 1,2 und 3

Abbildung 4.4: Aufgabencharakteristische Kurven für die Qualitätsstandards 1,2 und 3 für alle drei Tests.

Test 201			Test 301			Test 305		
Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
1.215	0.159	-2.633	1.142	-0.278	-2.086	2.305	0.017	0.159

Tabelle 4.9: Aufgaben-Parameter β_j für die einzelnen Qualitätsstandards.

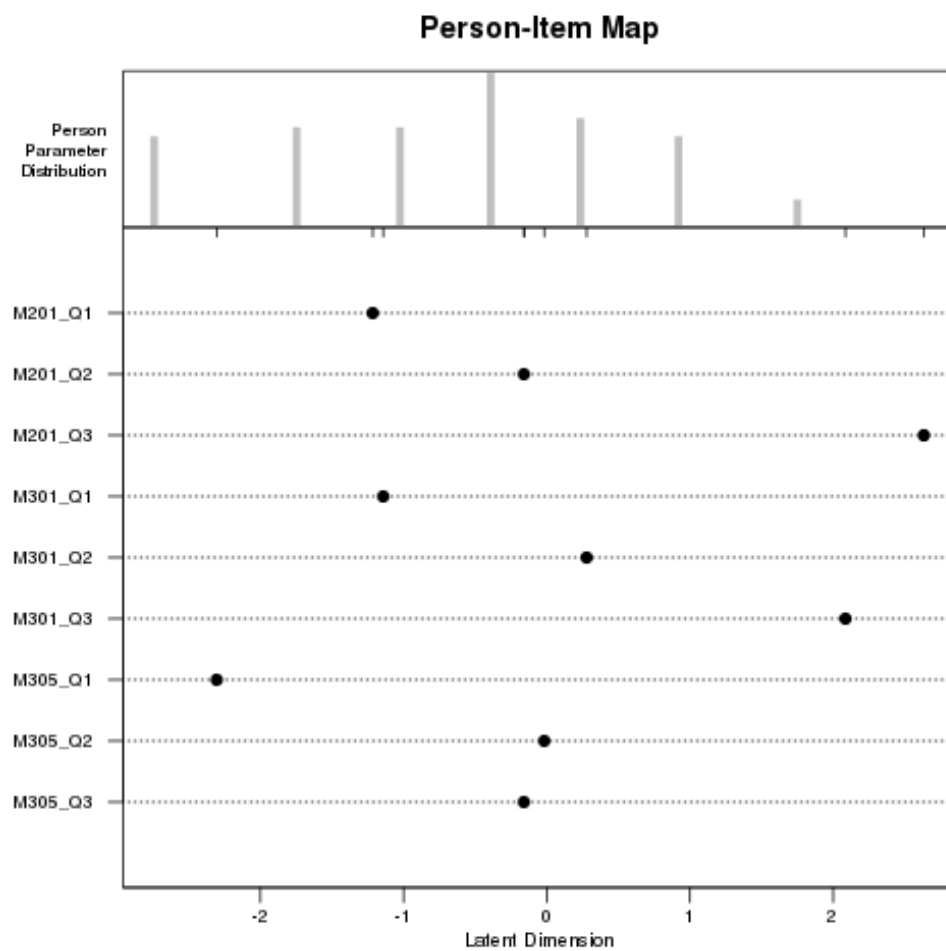


Abbildung 4.5: Person-Item-Map auf welcher die Verteilung der Personen basierend auf der latenten Skala ersichtlich ist und die Lage der Aufgaben-Parameter auf der latenten Skala.

5 Diskussion

6 Ausblick

Anhang