

Vom Einfluss des Kontextes auf Kompetenzen im Rahmen von Experimentiertests

Masterarbeit an der Pädagogischen Hochschule
Zürich

Masterstudiengang Fachdidaktik
Naturwissenschaften

vorgelegt von:
David-Matthias Sichau

eingereicht bei:
Pitt Hild

05. Februar 2015, Zürich

Inhaltsverzeichnis

1 Einleitung	1
2 Theoretischer Rahmen	2
2.1 Transfer	2
2.1.1 Historischer Überblick	2
2.1.2 Kritik	4
2.1.3 Lobato und Siebert 2002	5
2.1.4 Elemente von Transfer	5
2.1.5 Konsequenzen für den Unterricht	6
2.1.6 Zusammenfassung zum Transfer	8
2.2 Kompetenz	8
2.2.1 Bildungsreformen	8
2.2.2 Definition von Kompetenz	9
2.2.3 Kompetenz und Transfer	12
2.3 Kompetenz des skalenbasierten Messens	12
2.4 Forschungsfrage	13
3 Methode	14
3.1 Anforderungen	14
3.2 Umsetzung	14
3.3 Untersuchung	16
3.3.1 Vorbereitung	16
3.3.2 Durchführung	17
3.3.3 Nachbereitung	17
4 Ergebnisse	18
4.1 Kodierung	18
4.1.1 Items	18
4.1.2 Qualitätsstandards	18
4.1.3 Niveau	20
4.2 Fragebogen	21
4.3 Unterschiede zwischen den Klassen	22
4.4 Korrelation der Niveaus des skalenbasierten Messens	23
4.5 Rasch-Analyse	26
4.5.1 Parameterschätzung	26
4.5.2 Modellkontrolle des Rasch-Modells	28
4.5.3 Unterschied in den Schwierigkeiten der Qualitätsstandards	30
4.5.4 Unterschiede in den latenten Personenfähigkeiten	33

4.5.5	Zusammenhang zwischen Rasch-Modell und Fragebogen	36
4.6	Videoanalyse	38
4.6.1	Qualitätsstandards	38
4.6.2	Korrelation zwischen Video-Merkmalen und Qualitätsstufen . .	39
4.6.3	Messzeitpunkte und Messdauer	39
5	Diskussion	43
5.1	Kodierung	43
5.1.1	Items	43
5.1.2	Qualitätsstandards	43
5.1.3	Niveaus	44
5.2	Fragebogen	44
5.3	Unterschiede zwischen den Klassen	45
5.4	Ist das Abschneiden in den Tests unterschiedlich	45
5.5	Rasch-Analyse	46
5.5.1	Parameterschätzung	46
5.5.2	Modellkontrolle	47
5.5.3	Unterschied in den Schwierigkeiten der Qualitätsstandards . .	48
5.5.4	Unterschied in den latenten Personenfähigkeiten	48
5.5.5	Zusammenhang zwischen Rasch-Modell und Fragebogen	49
5.6	Videoanalyse	49
5.7	Zusammenfassung	50
6	Ausblick	51
6.1	Datengrundlage	51
6.2	Videoanalyse	51
6.3	Methoden	51
Literaturverzeichnis		53
Anhang		59
1	Urheberschaftsbestätigung	59
2	Daten und Auswertungen	59
3	Fragebogen	59
3.1	Fragebogen am Anfang	60
3.2	Fragebogen am Ende	64
4	Aufgabenstellung und Kodierungen	68
4.1	Test 201: Aufgabenstellung	69
4.2	Test 201: Kodierung	75
4.3	Test 301: Aufgabenstellung	77
4.4	Test 301: Kodierung	83
4.5	Test 305: Aufgabenstellung	86
4.6	Test 305: Kodierung	92

5	Einverständnis Erklärung für Video Aufnahme	94
---	---	----

Zusammenfassung

In der vorliegenden Masterarbeit wurde untersucht, ob die Kompetenz des skalenbasierten Messens von Lernenden auf der Sekundarstufe I in unterschiedlichen Kontexten gleich verfügbar ist. Diese Fragestellung ist aktuell interessant, aufgrund des stattfindenden Paradigmenwechsels der Bildungspolitik hin zur Kompetenzorientierung. Dabei ist es notwendig Kompetenzen zu entwickeln, welche unabhängig des Kontextes sind.

Um diese Fragestellung zu lösen wurden hands-on Experimentiertests des ExKoNawi Projektes der PH Zürich verwendet, bei denen die gleiche Kompetenz in unterschiedlichen Kontexten gemessen wird. In dieser Arbeit wurde dafür die Kompetenz des skalenbasierten Messens im Kontext der Temperaturmessung und Kraftmessung in den Fächern Physik und Chemie verwendet.

Die Forschungsfrage konnte mit Hilfe von Signifikanztests und Korrelationstests und mit der probabilistischen Testtheorie, basierend auf den erhobenen Daten, beantwortet werden. Es konnte gezeigt werden, dass für die hier untersuchte Stichprobe (72 1. Sek A Schülerinnen und Schüler), die Kompetenz des skalenbasierten Messens unabhängig ist vom fachlichen oder inhaltlichen Kontext ist. Aufgrund der geringen Datengrundlage, ist dieses Ergebnis nicht generalisierbar.

1 Einleitung

In den letzten Jahrzehnten fand ein Wandel in den Bildungssystemen zu einer Output-Orientierung statt. Der Wandel wurde durch das schlechte Abschneiden einiger Länder in internationalen Studien wie PISA (PISA-Konsortium Deutschland 2004) initialisiert. Dadurch fand ein Perspektivwechsel statt, sodass die Resultate des Bildungssystems überprüft werden und die Bildungssysteme nicht mehr über den Input gesteuert werden.

Dieser Wandel spiegelt sich auch in den neu entwickelten Bildungsstandards, in welchen die Kompetenzen der Schülerinnen und Schüler im Vordergrund stehen, welche diese nach dem Besuch des Bildungssystems erreicht haben sollen (Oelkers et al. 2008).

Oft werden diese angestrebten Kompetenzen unabhängig von einem inhaltlichen oder fachlichen Kontext gestellt (so z.B. bei HarmoS Konsortium HarmoS Naturwissenschaften+ (2010)). Deshalb wird erwartet, dass die Kompetenzen generalisierbar sind und teilweise auf andere Situationen übertragen werden können (Hartig und Klieme 2006).

Diese Generalisierbarkeit und Transferbarkeit von Kontexten soll in dieser Masterarbeit genauer analysiert werden. Im Rahmen der Validierung von hands-on Testaufgaben im Projekt ExKoNawi der PH Zürich (Metzger et al. 2013), soll untersucht werden, inwiefern die Qualitätsstandards der Kompetenz des skalenbasierenden Messens vom Kontext abhängig sind.

Dabei soll herausgefunden werden, ob das Erreichen eines bestimmten Qualitätsstandards von fachlichen oder inhaltlichen Kontext abhängig ist. Diese Fragestellung ist hinsichtlich der Kompetenzorientierung des Bildungssystems elementar. Nicht transferierbare Kompetenzen wären von geringem Nutzen, da diese erworbenen Kompetenzen dann nur an einen genau definierten Kontext angewendet werden können. Dies ist jedoch nicht das Ziel des Bildungssystems, da die erworbene Kompetenzen auf ausserschulische Kontexte angewendet werden sollten.

Die Forschungsfrage, ob eine gewisse Kompetenz von Lernenden auf der Sekundarstufe I in unterschiedlichen Kontexten gleich verfügbar ist, soll mithilfe von mehreren hands-on Experimentiertests zu einer Kompetenz in unterschiedlichen Kontexten beantwortet werden. Bevor diese Frage jedoch beantwortet werden kann, soll ein Überblick über den theoretischen Hintergrund zum Kontext, zum einen basierend auf dem Begriff des Transfers und zum anderen basierend auf dem Kompetenzbegriff, gegeben werden.

2 Theoretischer Rahmen

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 Blickwinkel 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. Greeno, Collins und Resnick (1996) definierten Transfer als „the process of applying knowledge in new situations“. Aber auch innerhalb der schulischen Bildung gibt es einen Transfer zwischen den verschiedenen Fächern. So wird von Schülerinnen und Schülern erwartet, dass inhaltliches, prozedurales und epistemologisches 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 zu verorten, soll zuerst ein historischer Überblick über den Begriff des Transfers gegeben werden.

Woodworth 1901

Eines der ersten Experimente zu Transfer wurde von Woodworth und Thorndike (1901) durchgeführt. Dabei mussten Probanden die Grösse von Rechtecken schätzen. Nachdem die Personen sich durch Wiederholungen verbessert hatten, wurde ihnen zwei neue Testsets 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 Woodworth und Thorndike, dass kein Transfer stattgefunden haben kann.

Ein ähnliches Resultat auf universitärem Niveau konnte von Renkl et al. (1994) gezeigt werden. Sie konnten zeigen, dass Nichtökonomien 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. Laut Woodworth und Thorndike (1901) basiert Transfer auf der Idee von identischen Elementen (Pea 2013). In diesem Theorieverständnis entsteht Transfer, wenn Wissen auf zwei verschiedenen Aufgaben, welche jedoch identische Merkmale/Elemente besitzen, angewendet wird. Dieses Transferverständnis basiert und stützt das Reiz-Reaktions-Modell des Lernens (Detterman 1993; Mietzel 2007).

Ferguson 1956

Eine alternative Theorie zum Transfer wurde von Ferguson (1956) entwickelt. Fergusons Theorie basiert darauf, dass die Intelligenz einer Person sich auf deren Transferleistung auswirkt. So findet nach Ferguson (1956) bei dem Lernen permanent ein Transfer statt, da jede Lernaufgabe von der anderen unterschiedlich ist und daher Transfer stattfinden muss. Im Unterschied zu Woodworth und Thorndike (1901) betrachtet Ferguson Transfer als einen kontinuierlichen Prozess, welcher durch Lernen verbessert werden kann. Wichtig ist jedoch zu beachten, dass Fergusons Theorie nur Nah-Transfer beschreibt. Unter Nah-Transfer wird Transfer zwischen sehr ähnlichen Situationen definiert.

Judd 1908

Eine der grundlegenden Studien zu Ferntransfer wurde von Judd (1908) gemacht. Bei Ferntransfer wird erworbenes Wissen auf Kontexte angewendet, welche sich deutlich vom Kontext unter welchem das Wissen erworben wurde, unterscheiden. Im Vergleich zu Woodworth und Thorndike 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, 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. 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. Judd (1908) erklärte dieses damit, dass die einen wussten, welches Prinzip sie auch bei der zweiten Situation anwenden können. Die Schülerinnen und Schüler, welchen das Prinzip hingegen nicht erklärt wurde, haben nur gelernt ihren Wurf auf die erste Situation anzuwenden; dieses Wissen konnten sie jedoch nicht generalisieren, da dies spezifisch für die Situation erworben wurde.

Im Vergleich zu Woodworth und Thorndike beinhaltet die Theorie von Judd ein ko-

gnitivistisches 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. Detterman (1993) kritisiert an dieser Studie die Verwendung von Transfer. So erklärt Judd einem Teil der Personen das zugrunde liegende Prinzip. Dies ist laut Detterman jedoch äquivalent, als ob 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 Gick und Holyoak (1980) durchgeführt. Dabei wurde untersucht, unter welchen Bedingungen Lernende Analogien verwenden, um strukturell ähnliche Probleme zu lösen. Ein Beispielproblem, welches sie den Lernenden gaben, war folgendes: „wie kann ein Tumor mit Strahlung zerstört werden, ohne dass gesundes Gewebe geschädigt wird?“ Dieses Problem wurde erstmals von Duncker und Lees (1945) verwendet. Eine mögliche Lösung ist mehrere Strahlen zu verwendet, welche sich nur im Tumor überlagern. Bevor sie dieses Problem lösten, war den Lernenden eine Geschichte erzählt worden, bei welcher das gleiche Prinzip verwendet wurde. 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 analog zu lösen, solange die Lernenden nicht auf die Ähnlichkeit aufmerksam gemacht werden.

2.1.2 Kritik

Lave (1988) kritisiert die unter 2.1.1 vorgestellten Untersuchungen. Da bei allen angenommen wird, dass 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 die Teilnehmer der verschiedenen Studien überhaupt lernen. So greift sie insbesondere die Annahme an, dass die Teilnehmer der Studien kontextunabhängig lernen. Sie lernen immer kontextspezifisch. Ein anderer Punkt, welchen sie angreift, ist, wer die Ähnlichkeit der Probleme definiert. Ist die Ähnlichkeit der Probleme für die Teilnehmer der Studie auch greifbar? Auch Detterman (1993) kritisiert die Studien. So sollten seiner Meinung nach alle Studien zu Transfer als Doppel-Blindstudien durchgeführt werden, da der Studienleiter unbewusst die Leistung der Probanden ändern könnte. Detterman fordert

daher:

No transfer experiment should be carried out without using a double blind procedure, particularly experiments assessing general transfer (Detterman 1993, S. 10).

2.1.3 Lobato und Siebert 2002

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

Lobato und Siebert (2002) möchte einen Kritik-Punkt von Lave (1988) lösen. So kritisierte Lave, dass der Untersucher festlegt, was Transfer von Wissen ist. Daher legten Lobato und Siebert (2002) als Messung für Transferleistung fest, welche Ähnlichkeit der Proband selbst zwischen verschiedenen Situationen zieht. So untersuchten sie einen Schüler, welcher eine Rollstuhllampe erhöhen sollte, ohne die Steigung zu verändern. Der Schüler löste dieses Problem, indem er die Verhältnisse von Höhe zu Länge konstant hielt. Er verwendete dafür jedoch nicht die im Mathematikunterricht 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 hatte, indem er das Konzept von konstanter Geschwindigkeit als Verhältnis von zurückgelegter Strecke zur Zeit auf dieses Problem angewendet hatte.

Lobato und Siebert (2002) konnten damit zeigen, dass 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.

Marini, McKeough und Lupart (1995) definieren drei Elemente, welche zu einem Transfer führen. Das erste Element besteht aus Merkmalen 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 (siehe Seite 189ff). Dies führt dazu, dass lösungsrelevantes Wissen von dem vorhanden und dem verarbeitbaren Wissen abhängt. Zusätzlich kann in einem bestimmten Kontext jedoch nicht alles Wissen abgerufen werden, da man mit trägem Wissen rechnen muss und auch der aktuellen Motivation des Lernenden.

Als zweites Element von Transfer geben Marini, McKeough und Lupart 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 (Marini, McKeough und Lupart 1995, s. S. 279). Aufgrund dessen haben Novizen oft Probleme 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 Godden und Baddeley (1975). 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 Schoenfeld (1988); So hatten Lernende keine Schwierigkeiten mit einer Divisionsaufgabe. Wenn die Aufgabe jedoch in einen Kontext gestellt wurde, 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 Woodworth und Thorndike (1901), welche nur den Aspekt der Aufgabenmerkmale genauer untersucht hatten. Erst neuere Arbeiten berücksichtigen alle Elemente und insbesondere den Kontext (Lobato und Siebert 2002; Detterman 1993; Greeno, Collins und Resnick 1996).

2.1.5 Konsequenzen für den Unterricht

Wie Claxton (1990) zeigte, darf jedoch nicht davon ausgegangen werden, dass in der Schule erworbene Wissen ohne Weiteres auf andere Alltagsprobleme angewendet werden kann. So sprach Whitehead (1929) 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 Whitehead (1929) entsteht träges Wissen oft unter schulischen oder universitären Bedingungen.

Detterman 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 (Detterman 1993, S. 21).

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 (Bransford, Brown und Cocking 2000, Kapitel 3).

Überlernen von Fähigkeiten

Eine Möglichkeit, gute Transferleistung zu erreichen, ist das intensive Einüben von Grundfertigkeiten, wie zum Beispiel in der Grundschule. LaBerge und Samuels (1974) untersuchten dies bei der Fertigkeit des Lesens. Dabei wird das Üben nicht abgebrochen, wenn die Schülerinnen und Schüler die Fertigkeit subjektiv (aus der Sicht der Lehrperson) bereits können, sondern noch einige Zeit fortgesetzt. LaBerge und Samuels 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 Perkins und Salomon (1989) 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 (LaBerge und Samuels 1974; Mietzel 2007).

Diese Erkenntnis deckt sich mit den Forderungen vom Whitehead (1929), welcher bereits 1929 davor warnte, dass in der Schule trügerisches Wissen entsteht. Daher soll in der Schule darauf geachtet werden, nicht zu viel in zu kurzer Zeit zu erreichen. So fordert er auch wenige Themengebiete gründlich zu erarbeiten. Diese Forderung wurde auch von neueren Studien bestätigt (Porter 1989; Brophy 1992; Millar und Osborne 1999).

Bransford, Brown und Cocking (2000) beschreibt diesen Aspekt unter dem Begriff „initial learning“ und definiert ihn etwas breiter. Im Unterschied zu den anderen Autoren hält er auch die Motivation für einen wichtigen Aspekt beim Einüben von Grundfertigkeiten. Daher schreibt er auch der Problemstellung einen wichtigen Stellenwert zu, da Probleme nicht zu einfach und auch nicht zu schwer sein sollten, damit die Motivation der Schülerinnen und Schüler nicht zerstört wird.

Entkontextualisieren

Wie bereits vorher angesprochen, hängt Wissen sehr stark vom Kontext ab unter welchem es gelernt wird (Godden und Baddeley 1975; Schoenfeld 1988). Anderson, Reder und Simon (1996) 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 (Perkins und Salomon 1989). Entkontextualisieren von Wissen ist jedoch nicht ausreichend, zusätzlich müssen Lernende lernen, wann und wo welches Wissen angewendet werden muss (Wiggins 1993). Diese Erkenntnis deckt sich mit den Ergebnissen von Gick und Holyoak (1980), bei denen die Teilnehmer einen höheren Transfer aufwiesen, wenn auf die Ähnlichkeit der Situationen hingewiesen wurden.

Problemorientierter Unterricht

Williams (1992) untersuchte viele Lernsituationen im medizinischen Studium auf ihre Möglichkeiten zu Transfer. Sie stellte fest, dass das Wissen, welches in Vorlesungen gelernt wurde, im klinischen Teil der Ausbildung vergessen ist. Ein Grund dafür ist, dass in vielen Lehrbüchern und Vorlesungen theoretisches Wissen losgelöst von Anwendungen dargestellt wird. 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 Shuell (1996), 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 Barrows (1985), Michael et al. (1993), Shuell (1996), Corte (2003), Reusser (2005), Fässler (2007) und Pea (2013)). Das Ziel dabei ist das 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 zum Transfer

Es wurde im 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.

2.2 Kompetenz

Nachdem ein Überblick über den Transferbegriff erarbeitet wurde, soll in diesem Abschnitt versucht werden die Konsequenzen aus der Betrachtung zum 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, um die Ressourcen effektiver einzusetzen.

Dieser Perspektivwechsel wurde von verschiedenen grossen Bildungsstudien (PISA (PISA-Konsortium Deutschland 2004), TIMSS (Martin und Mullis 2003) und IGLU (Bos et al. 2003)) 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 hinsichtlich 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 sollen (Oelkers et al. 2008).

Diese Bemühungen führten in vielen Ländern zur Entwicklung von neuen Bildungsstandards (Berner und Stolz 2006). In der Schweiz wurde dies von der EDK Schweizer Konferenz der Kantonalen Erziehungsdirektoren (2004) unter dem Titel „Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat)“ angestossen. In Deutschland wurde neue Bildungsstandards von der Kultusministerkonferenz (2004) 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 (McGee 1996). Auch in Australien wurde ein ähnliches Bildungskonzept 2000 unter dem Name „outcome-based education (OBE)“ umgesetzt (Killen 2000). Auch in England wurde zu Beginn des Jahrtausends Bildungsreformen gefordert (Millar und Osborne 1999), welche dann um 2005 umgesetzt wurden (Huber et al. 2006).

2.2.2 Definition von Kompetenz

Der Begriff der Kompetenz wird im Moment sowohl fachlich als auch politisch sehr stark diskutiert. So spricht Weinert (2001a) von einer Inflation des Kompetenzbegrif-

fes. Daher stellte Weinert (2001b, S.27) folgende Definition des Kompetenzbegriffes auf:

die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können (S.27).

Dieser Kompetenzbegriff wurde die Basis für den Kompetenzbegriff, welcher in den Bildungsstandards (sowohl der Schweiz als auch von Deutschland) verwendet wird, und wurde von Klieme (2004) weiter präzisiert. Klieme (2004) 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 als „Kenntnisse“, „Routinen“ oder „Fertigkeiten“ bezeichnet.
3. Kompetenz als motivationaler Orientierungen, welche notwendig ist, um eine Aufgabe zu bewältigen.
4. Handlungskompetenz, als Integration der vorgängigen 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 dieser schwierig von der Definition der allgemeinen Intelligenz zu unterscheiden ist, insbesondere im Bezug auf den oben genannten ersten Punkt. Weinert (2001a) empfiehlt daher eine Einschränkung des Kompetenzbegriffes. So sollen Kompetenzen auf einen eingeschränkten Raum von Kontexten und Situationen bezogen und allgemeine intellektuellen Fähigkeiten ausgeschlossen werden. Dies begründet Weinert (2001a) 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 definier-

te Ziele zu erreichen. Dies führt auch 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 (Hartig und Klieme 2006).

Ein weiterer Unterschied zwischen dem Kompetenz- und dem Intelligenzkonzept beruht auf der Lernbarkeit. So fordert Baumert, Stanat und Demmrich (2001, S. 22), 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 daher für das Bildungsmonitoring von Interesse. Intelligenz wird hingegen als relativ stabil betrachtet, da sie hauptsächlich von genetischen Faktoren abhängt (Shakeshaft et al. 2013). 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 angeeignet haben sollte.

Des Weiteren gibt es bei der Erstellung von Leistungsmessungen Unterschiede. Bei Intelligenztests werden bestimmte Primärfaktoren verwenden, z.B. dreidimensionales Denken, Gedächtnis, usw., welche Unterschiede zwischen einzelnen Personen aufzeigen sollen. Kompetenzen hingegen werden durch die Anforderungen definiert (Rychen und Salganik 2001). 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 (Konsortium HarmoS Naturwissenschaften+ 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 Intelligenzleistung und Kompetenzleistungen festgestellt. So fand Rindermann (2006) 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 (Hartig und Klieme 2006).

2.2.3 Kompetenz und Transfer

Der Kompetenzbegriff wird in der Literatur sehr unterschiedlich definiert (Klieme 2004; Weinert 2001a). Dennoch ist die Definition des Kompetenzbegriffes für internationale Studien wie PISA (PISA-Konsortium Deutschland 2004), TIMSS (Martin und Mullis 2003) und IGLU (Bos et al. 2003) 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 (Hartig und Klieme 2006).

Im Abschnitt 2.1.5 wurde herausgearbeitet, welche Eigenschaften von Unterricht zu besserer Transferleistung führen können. Auch Lersch (2007) 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 Mietzel (2007) für das Entkontextualisieren von Unterricht, um Transferleistung zu fördern. Auch die Problemorientierung von Lerngelegenheiten wird von Lersch (2007) für kompetenzfördernden Unterricht als wichtig gehalten, insbesondere fordert er, dass „systematische Wissensvermittlung [...] um variable Anwendungssituationen“ ergänzt werden sollten. 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 (Klieme 2004).

Diese Verknüpfung von Transfer und Kompetenz befindet sich auch im Ansatz von Gott und Duggan (2002), bei dem transferfähiges Strategiewissen und kontextspezifisches Fachwissen erforderlich ist (Gott und Duggan 1996).

2.3 Kompetenz des skalenbasierten Messens

Im Rahmen von ExKoNawi hands-on Experimentiertests wurde ein Modell basierend auf dem Kompetenzmodell von Gott und Duggan (1996) entwickelt, um verschiedene hands-on Kompetenzen von Schülerinnen und Schüler auf der Sekundarstufe I in der Schweiz zu messen werden (Metzger et al. 2013). Einer der Kompetenzen, welche mit ExKoNawi hands-on Experimentiertests gemessen werden soll, ist die Kompetenz des *skalenbasiertes Messen* (Gut et al. 2014). Die Definition dieser Kompetenz basiert auf der Arbeit von Munier, Merle und Brehelin (2013). In dieser Kompetenz geht nach (Gut et al. 2014) darum „quantitative Größen mit gegebenen Messinstrumenten genau [zu] messen“. Bei dieser Kompetenz gibt es drei Teilbereiche, welche eine wichtige

Rolle spielen. Zum einen müssen die Schülerinnen und Schüler entscheiden, welches Messinstrument besser für eine Messung geeignet ist. Ein weiterer Teilaспект 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 (Munier, Merle und Brehelin 2013; Gut et al. 2014).

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 daher zur Frage, ob das erreichbare Kompetenzniveau von Schülerinnen und Schüler tatsächlich unabhängig vom inhaltlichen und fachlichen Kontext ist? 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 können sollten. Um diese Frage zu beantworten, soll in der vorliegenden Arbeit untersucht werden, ob die erreichten Kompetenzniveaus des skalenbasierten Messens unabhängig vom fachlichen oder inhaltlichen Kontext sind.

3 Methode

3.1 Anforderungen

Um die vorliegende Fragestellung zu beantworten, ist es notwendig mehrere Test zu verwenden, welche die Kompetenz des skalenbasierten Messens erheben. Zusätzlich müssen die Tests die Kompetenz des skalenbasierten Messens unter verschiedenen Kontexten ermitteln. 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 bestimmt wurde. Zusätzlich wurde ein dritter Test neu entwickelt, bei welchem eine Temperaturmessung im Fach Physik durchgeführt wurde. Die Teststellung wird in Sichau (2015) detailliert beschrieben. Der dritte Test wurde so entworfen, dass 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 Klassenzahl durchgeführt werden. Dies führte zur Bildung von zwei Gruppen, wobei 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 Testergebnisse haben. Die Tabelle 3.1 gibt eine Übersicht über die Gruppeneinteilung der Schülerinnen und Schüler 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 stammten sie aus verschiedenen Gemeinden. Die Klassen in Glattbrugg hatten beide dieselbe Lehrperson, die anderen Klassen hatten unterschiedliche Lehrpersonen. Ein Überblick über die wichtigsten Daten zu den einzelnen Klassen lässt sich der Tabelle 3.2 entnehmen.

Gruppe FABC	Gruppe FACB	Gruppe ABCF	Gruppe ACBF
Fragebogen	Fragebogen	Temperatur Physik 305	Temperatur Physik 305
Temperatur Physik 305	Temperatur Physik 305	Kraft Physik	Temperatur Chemie
Kraft Physik 301	Temperatur Chemie 201	Temperatur Chemie 201	Kraft Physik 301
Temperatur Chemie 201	Kraft Physik 301	Fragebogen	Fragebogen

Tabelle 3.1: Aufteilung der Gruppen innerhalb einer Klasse

	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	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: Wichtigste Informationen zu den einzelnen Klassen

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 hatten, in ein Zimmer gesetzt werden, indem keine Videoaufnahme stattfand hat. Die Erlaubnis zur Videoaufnahme wurde bereits vor der Durchführung von den Klassenpersonen organisiert und eingesammelt.

3.3 Untersuchung

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 benötigten Materialien in dieser Box finden konnten.



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

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 Videoauswertung aufgestellt

wurde. Die Videoaufnahme wurde vor Eintreten der Schülerinnen und Schüler gestartet, um die Ablenkung durch die Kamera möglichst gering zu halten.

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 konnten. 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 konnten. Nach dem dritten Test wurde wieder eine kurze fünfminütige Pause durchgeführt. Während der Tests wurden den Schülerinnen und Schülern nur Fragen zu Unklarheiten beantwortet, inhaltliche Fragen oder Fragen zum korrekten Vorgehen wurden zurückgewiesen.

3.3.3 Nachbereitung

Nachdem die Tests durchgeführt wurden, wurden die Auswertungsbögen eingesammelt und von David Sichau erstkodiert. Es wurde eine Zweitkodierung vor 15 % der Auswertungsbögen von Pitt Hild durchgeführt. Die 11 Auswertungsbögen zur Zweitkodierung wurden zufällig (random generator) ausgewählt, um sicherzugehen, dass ein Bias ausgeschlossen 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 Testsituation hatten. 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 beschrieben wurde die Erstcodierung von David Sichau durchgeführt. Es wurde eine Zweitcodierung von 15 % zufällig ausgewählten (per Random Generator) Auswertungsbögen von Pitt Hild durchgeführt. Dabei wurden die identischen Kodierschemata verwendet, welche sich im Anhang der Arbeit befinden (siehe Abschnitt 4 im Anhang).

4.1.1 Items

Es gab insgesamt elf Items, welche mit den Kodierschemata kodiert wurden.

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

Code erhältlich auf:
GitHub
<http://git.io/mk9z-Q>

4.1.2 Qualitätsstandards

Aus den elf Items wurden fünf Qualitätsstandards entwickelt vgl. Hild, Metzger und Parchmann (2014) und Gut et al. (2014). Es gibt bedingte und unbedingte Qualitätsstandards. Bei den bedingten Qualitätsstandards ist für das Erreichen dieser 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.

Item	201		301		301	
	p_0	κ	p_0	κ	p_0	κ
1.1	1.00	1.00	0.91	0.74	0.91	0.79
1.2	0.91	0.81	1.00	na	1.00	1.00
2.1	0.81	0.67	0.81	0.74	1.00	1.00
3.1	1.00	1.00	0.91	0.81	1.00	1.00
3.2	1.00	na.	1.00	1.00	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	1.00	na.	1.00	na.	1.00	na.
4.4	1.00	na.	1.00	na.	1.00	na.
5.1	1.00	na.	1.00	na.	1.00	na.
5.2	0.91	na.	1.00	1.00	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)

Qualitätsstandard 2

Bei Qualitätsstandard 2 wird die Dokumentation der Messung bewertet . Dieser Qualitätsstandard wird nur erreicht, wenn Item 2.1 (Werden alle Messungen und Messergebnisse vollständig dargestellt?) mindestens den Wert von 2 erreicht hat.

Qualitätsstandard 3

Im dritten Qualitätsstandard wird das Begründen des richtigen Messinstruments bewertet. Dieser Standard wird nur erreicht, wenn Item 3.1 (Ist das Korrekte Messinstrument gewählt worden?) und Item 3.2 (Wird die Wahl des Messinstruments korrekt begründet?) zusammen 2 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. Diese Level wird erreicht, wenn die Items addiert mindestens 2 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 1 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 angeben, 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>

Unbedingtes Niveau

Dieses Niveau ist der Summenscore der einzelnen unbedingten Qualitätsstandards.

Test	unbedingtes Niveau						bedingtes Niveau					
	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: Erreichte Niveaus aller Schülerinnen und Schüler. Beim bedingten Niveau ist es jeweils erforderlich, dass alle vorhergehenden Qualitätsstandards erreicht worden sind.

Bedingtes Niveau

Dieses Niveau ist der Summenscore der bedingten Qualitätsstandards.

4.2 Fragebogen

Im standardisierten Teil des Fragebogens wurden Fragen zum absoluten Selbstkonzept nach SESSKO gestellt (Schöne et al. 2002). 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“ (Schöne et al. 2002). α_d bezeichnet das standardisierte Cronbach Alpha wenn dieses Item weggelassen würde.

Zusätzlich wurden nach Dierks, Höffler und Parchmann (2014) Fragen zum Selbstkonzept zu Schulversuchen entwickelt und angepasst, welche in Tabelle 4.5 aufgeführt sind.

Es wurde die innere Konsistenz beider Skala überprüft. Für die innere Konsistenz wurde Cronbachs Alpha verwendet, da dies nach Eisinga, Grotenhuis und Pelzer (2013) eher zu einer Unterschätzung der innere Konsistenz führt. Bei der Skala „Schulisches Selbstkonzept - absolut“ wurde ein standardisiertes Cronbach Alpha $\alpha = 0.77$ erreicht.

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 Schulversuchen weiss ich gleich: „Das verstehe ich nie.“	0.65
NatSK5	Für Schulversuche habe ich einfach keine Begabung.	0.63
NatSK6	Mit den Aufgaben bei Schulversuchen 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 Dierks, Höffler und Parchmann (2014). α_d bezeichnet das standardisierte Cronbach Alpha, wenn dieses Item weggelassen würde.

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 Alpha $\alpha = 0.69$ erreicht. Insgesamt konnten dabei 64 vollständige Fragebögen ausgewertet werden. Die Annahme des standardisierten Cronbach Alpha, dass alle Items die gleiche Varianz haben, kann aufgrund des geringen Unterschiedes zum absoluten Cronbach Alpha angenommen werden: $\alpha_{std} = 0.69$ versus $\alpha_{abs} = 0.68$.

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 bedeutend an Informationsgehalt ein-

gebüsst zu haben.

Aufgrund der geringen Anzahl an Beobachtungen für einzelne Qualitätsstandards wurde der exakte Test nach Fisher verwendet (Fisher 1922). Es wurden Kontingenztafeln für jeden Qualitätsstandard (Q1 bis Q5 und QS1 bis QS5) erstellt und in jeder Tafel die beiden Levels (0 und 1) unter 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 befinden sich in Tabelle 4.6. Bei keinem der 60 Tests konnte die Nullhypothese abgelehnt werden ($p < 0.05$). 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

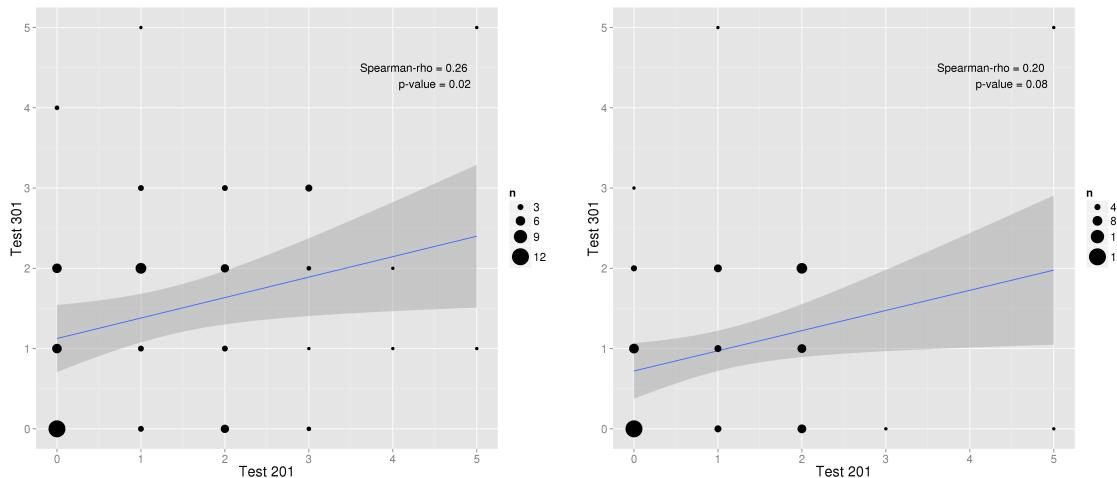
In einem nächsten Schritt wurde untersucht, inwiefern die bedingten und unbedingten Niveau-Stufen zwischen den einzelnen Tests korrelieren. Dazu wurde als Rangkorrelationskoeffizient Spearmans ρ berechnet. Der Vorteil dieser Methode ist, dass keine Annahmen über die zugrundliegenden Daten gemacht werden müssen. Des Weiteren bietet diese Methode den Vorteil, dass sie gegenüber Ausreisern robust ist (Kowalski 1972).

Da die Korrelation alleine keinen Aufschluss darüber gibt, ob diese signifikant ist, wurde sie 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 Darstellung 4.1). In die Streudiagramme wurde die Gerade der linearen Regression eingetragen mit dem zugehörigen 95% Vertrauensintervall. Zusätzlich wurde 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	0.08	0.20
201 vs. 305	1e-4	0.44	4e-3	0.33
301 vs. 305	2e-3	0.36	0.89	0.01

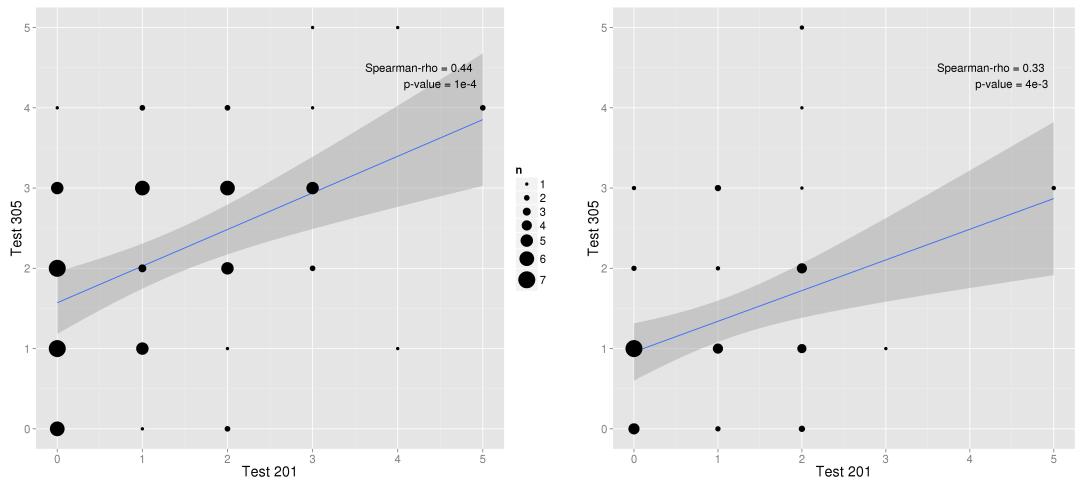
Tabelle 4.7: Spearmans ρ und p-Werte für die Korrelation zwischen den unbedingten Niveaus (uLev) und den bedingten Niveaus (kLev) zwischen den einzelnen Tests.



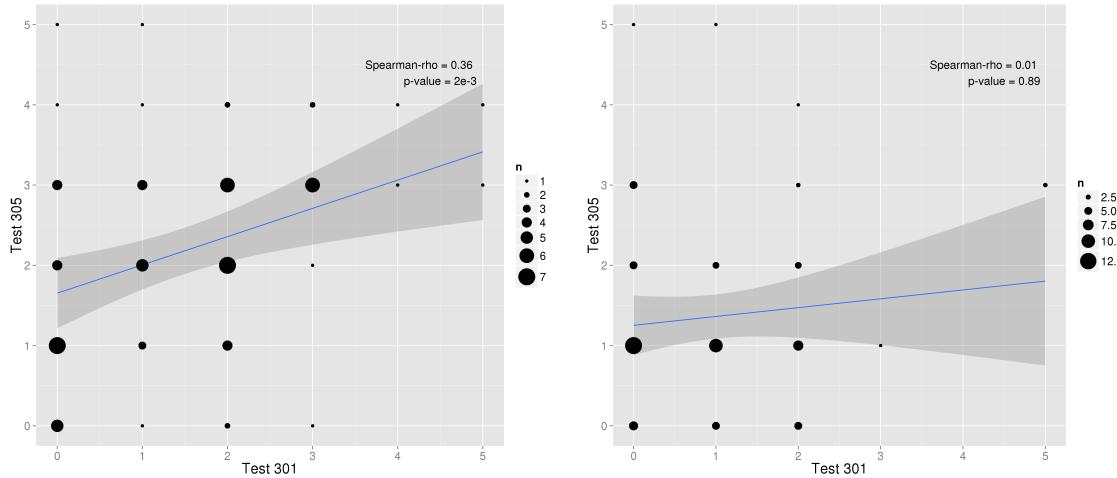
Code erhältlich auf:

GitHub

<http://git.io/FnbD>



(c) Correlation of unconditioned Niveau-Stufen between Test 305 and 201. (d) Correlation of conditioned Niveau-Stufen between Test 305 and 201.



(e) Correlation of unconditioned Niveau-Stufen between Test 305 and 301. (f) Correlation of conditioned Niveau-Stufen between Test 305 and 301.

Abbildung 4.1: Korrelation zwischen den Niveau-Stufen der 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 zugrundeliegenden Daten, der dunkelgraue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

4.5 Rasch-Analyse

Als probabilistische Testmethode wurde das Rasch-Modell verwendet. Der Grund für diese Methodik ist, dass es sich bei der Kompetenz des skalenbasierten Messens um ein latentes Merkmal handelt. In anderen Worten: Die Kompetenz des skalenbasierten Messens ist nicht direkt beobachtbar.

Es wurde 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 die Personen ist und $j = 1, \dots, m$ die Zählvariable für die Aufgaben darstellt. Die Variable $u_{ij} \in \{0, 1\}$ bezeichnet die dichotome Antwort einer Person auf eine Aufgabe. 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. Damit ist gemeint, dass das Lösen einer Aufgabe immer zufällig ist, jedoch die Lösungswahrscheinlichkeit von der Fähigkeit einer Person und der Schwierigkeit der Aufgabe abhängig ist. 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 Parameter werden basierend auf den tatsächlichen Testergebnissen u_{ij} geschätzt.

4.5.1 Parameterschätzung

Für die Parameterschätzung des Rasch-Modells gibt es verschiedene Ansätze. Da die beste Methode von den Daten abhängig ist, wird in einem ersten Schritt das Rasch-Modell sowohl mit der bedingten Maximum-Likelihood-Schätzung, als auch mit der marginalen Maximum-Likelihood-Schätzung getestet und die Resultate verglichen.

Bei der bedingten Maximum-Likelihood-Schätzung wird ein zweistufiges Vorgehen gewählt. Zuerst werden die Aufgabenparameter geschätzt, ohne die Personenparameter zu beachten. Erst in einem zweiten Schritt werden die Personenparameter geschätzt. Ein Problem dieser Methodik ist, dass Personenfähigkeiten von Personen, welche keine oder alle Aufgaben gelöst haben, nicht geschätzt werden können (Mair und Hatzinger 2007).

In der marginalen Maximum-Likelihood-Schätzung wird angenommen, dass für die Personenfähigkeiten in der Stichprobe eine bestimmte Verteilung vorliegt. Meistens wird dabei eine Normalverteilung angenommen, da diese einfacher zu berechnen ist

(Strobl 2012). Diese Annahme ist insbesondere dann problematisch, wenn nur eine Stichprobe der Gesamtbevölkerung verwendet wird (Rizopoulos 2006).

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 der Schätzung der Schwierigkeitsgrade der einzelnen Qualitätsstandards.

Bei der Schätzung der Personenparametern θ konnte die bedingte Maximum-Likelihood-Schätzung alle 72 Personenfähigkeiten ohne Extrapolationen berechnen. Die marginalen Maximum-Likelihood-Schätzung konnte jedoch nur die Personen-Fähigkeiten von 64 Personen berechnen. Daher wird in der weiteren Arbeit für alle Rasch-Modelle jeweils der bedingten Maximum-Likelihood-Schätzer verwendet.

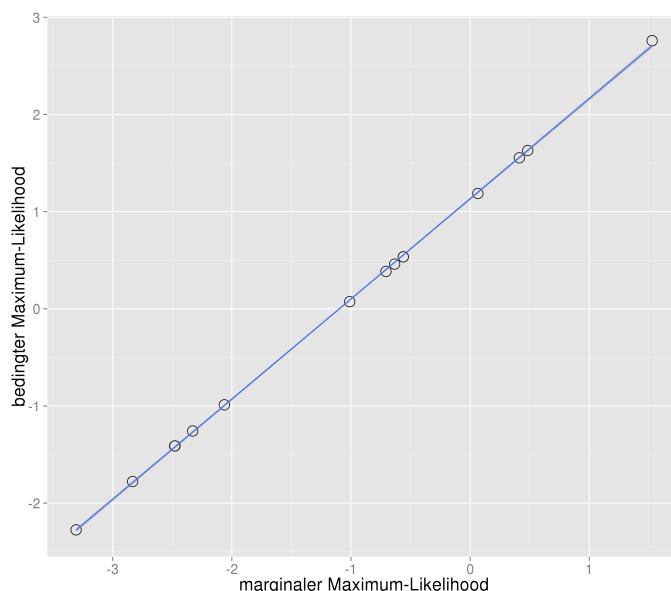


Abbildung 4.2: Vergleich des Rasch-Modells mit der bedingten Maximum-Likelihood-Schätzung und der marginalen Maximum-Likelihood-Schätzung. Da alle Punkte auf einer Geraden liegen, gibt es keinen Unterschied zwischen den unterschiedlichen Schätzmethoden für Schwierigkeitsgrad der Qualitätsstandards in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

Es gibt noch weitere Parameter-Schätzer wie den Bayesianischen Ansatz, welcher Markov-Chain-Monte-Carlo Methoden verwendet. Dieser trifft jedoch auch Annahmen über die Verteilung der Personenparameter (Fischer und Molenaar 1995, siehe Kapitel 3). Die Annahmen decken sich daher mit dem marginalen Maximum-

Likelihood Schätzer.

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 Personenrandsummen verwendet.

Der p-Wert des Andersens Likelihood-Quotienten Test beträgt $p = 0.14$. Daher liegt keine signifikante Modellverletzung vor, dass heisst, die Aufgabenparameter 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.

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 Personenrandsummen 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>

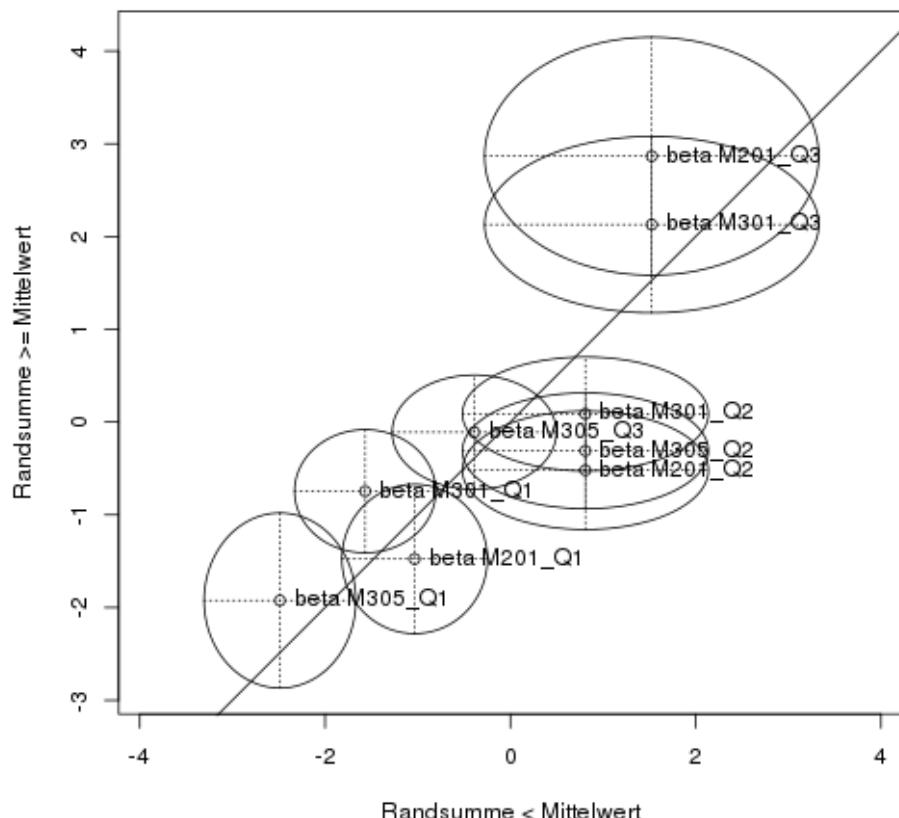
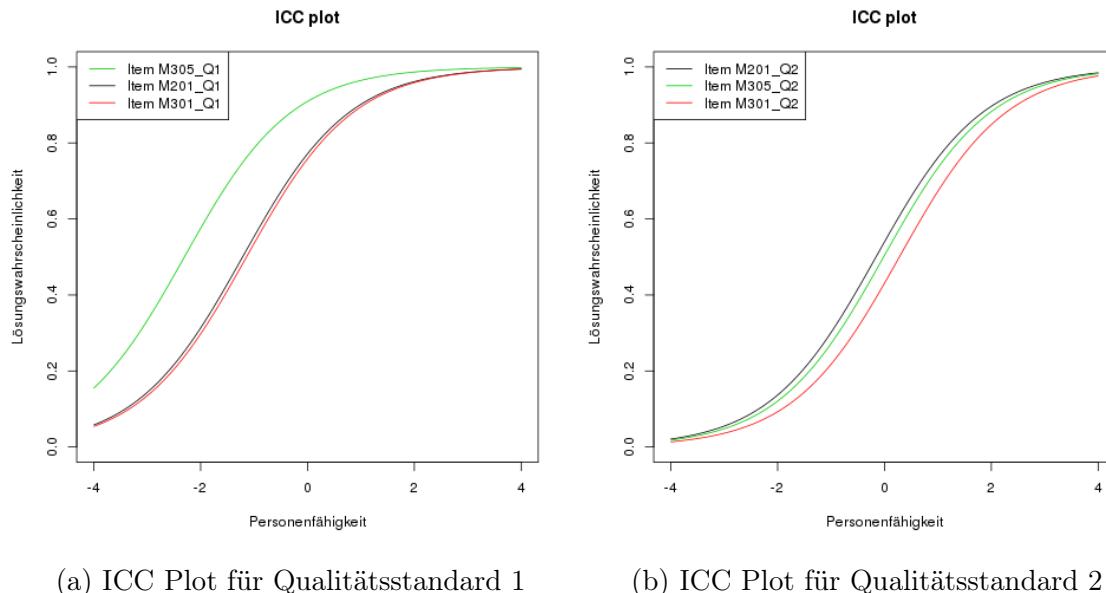


Abbildung 4.3: Modellkontrolle des Rasch-Modells: kein Qualitätsstandard hat eine signifikante Abweichung von der Diagonalen, daher gibt es keine signifikanten Unterschiede für Personen mit niedrigen und hohen Randsummen in den Qualitätsstandards.

4.5.3 Unterschied in den Schwierigkeiten der Qualitätsstandards

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



In Tabelle 4.9 finden sich die Aufgabenparameter β_j der einzelnen Qualitätsstandards.

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: Aufgabenparameter β_j für die einzelnen Qualitätsstandards.

Mit den so gewonnenen Aufgabenparametern β_j wurde nun die Korrelation zwischen den einzelnen Test berechnet. Da mit dem bisherigen Rasch-Modell der Personenparameter θ_i über alle drei Tests identisch ist, sollten sich die Schwierigkeitsgrade der einzelnen Qualitätsstufen in den Tests nicht unterscheiden. Die Ergebnisse dieser Analyse sind im der Darstellung 4.6 und in Tabelle 4.10 angegeben. Wichtig dabei ist, dass der Stichprobenumfang mit 3 sehr gering ist.

Code erhältlich auf:

GitHub

<http://git.io/FVZt>

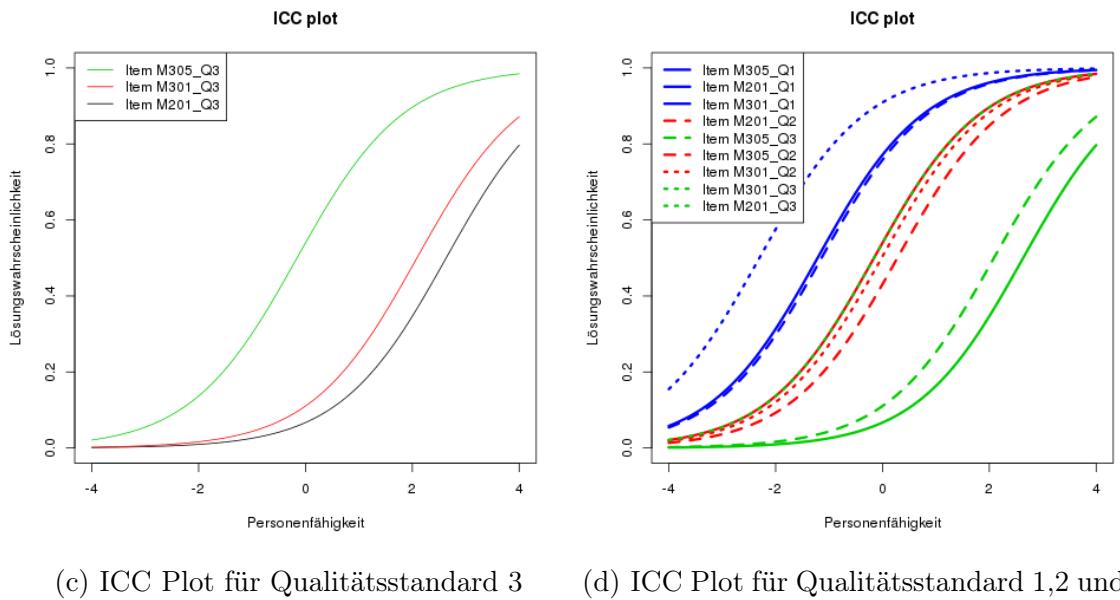


Abbildung 4.4: Aufgabencharakteristische Kurven für die Qualitätsstandards 1, 2 und 3 für alle drei Tests.

201 vs 301		201 vs 305		301 vs 305	
p-Wert	ks	p-Wert	ks	p-Wert	ks
1.00	0.33	1.00	0.33	0.60	0.67

Tabelle 4.10: Resultate des Kolmogorow-Smirnow-Test für die Übereinstimmung der Schwierigkeiten der Qualitätsstandards. Wobei ks die Test-Statistik des Kolmogorow-Smirnow-Test ist.

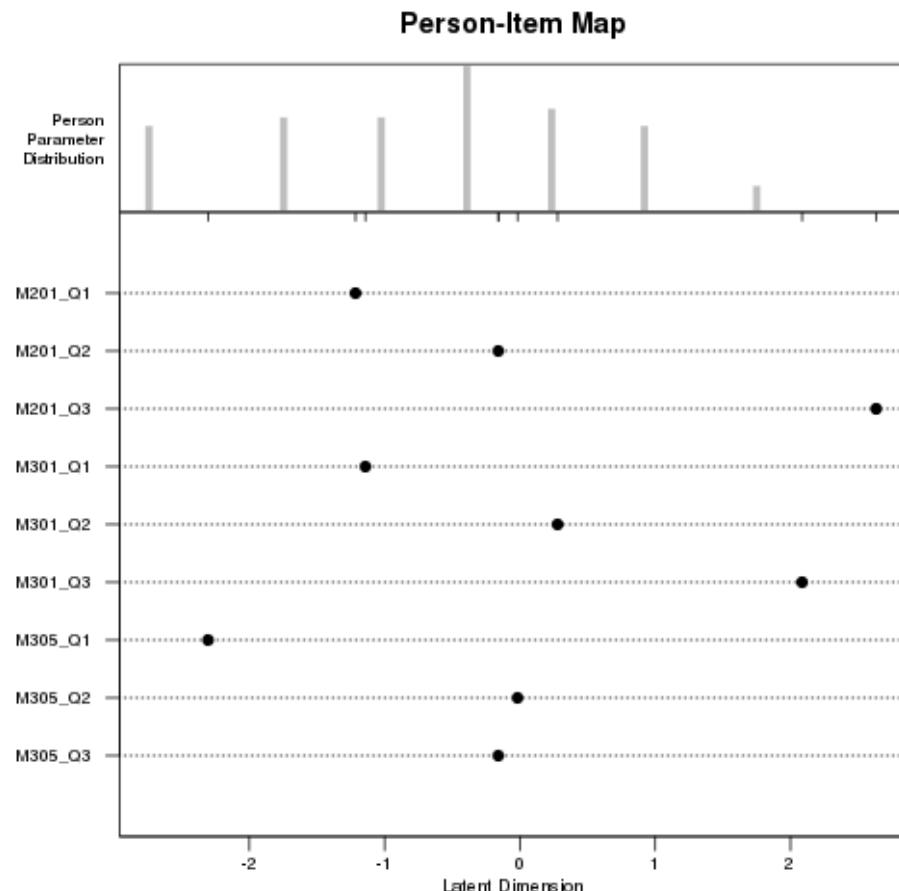
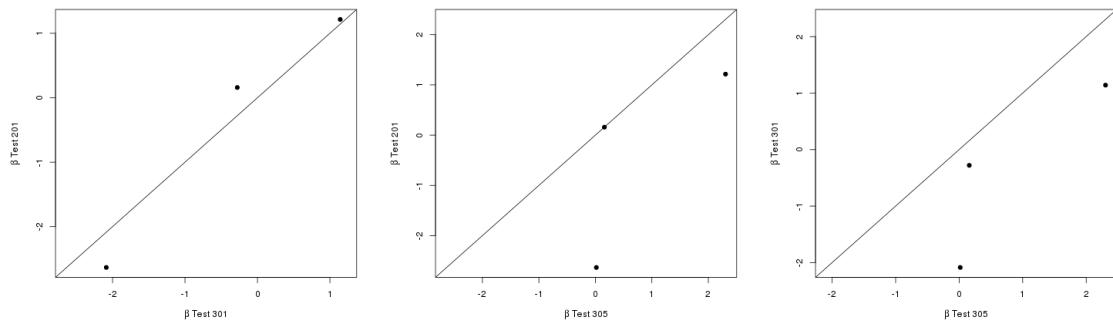


Abbildung 4.5: Person-Item-Map, auf welcher die Verteilung der Personen basierend auf der latenten Skala ersichtlich ist und die Lage der Aufgabenparameter auf der latenten Skala. Anhand dieser Darstellung kann man sehen, dass z.B. der Qualitätsstandard 1 im Test 201 und im Test 301 einen sehr ähnlichen Schwierigkeitsgrad besitzen.



- (a) Vergleich der Aufgaben-
parameter β_j zwischen
Test 201 und 301.
(b) Vergleich der Aufgaben-
parameter β_j zwischen
Test 201 und 305.
(c) Vergleich der Aufgaben-
parameter β_j zwischen
Test 301 und 305.

Abbildung 4.6: Vergleich der Aufgabenparameter zwischen den einzelnen Tests. Wenn die Schwierigkeiten der Qualitätsstandards übereinstimmen würden, müssten alle Punkte auf der Winkel-Halbierenden liegen.

4.5.4 Unterschiede in den latenten Personenfähigkeiten

Nachdem in einem ersten Schritt die Schwierigkeit der Qualitätsstandards untersucht und festgestellt wurde, dass keine signifikante Unterschiede in den Schwierigkeitsgraden zwischen den einzelnen Tests existieren, wurde nun ein neues Rasch-Modell entwickelt.

Es werden jetzt drei Rasch-Modelle gebildet, bei denen jeder Test und dessen Qualitätsstandards 1-3 in einem Modell kombiniert wurden. Aus den drei Modellen wurden die Personenfähigkeiten berechnet und dann mit dem Kolmogorow-Smirnow-Test auf den Goodness of fit überprüft. Dabei wurden Personenparameter, welche aufgrund des bedingten Maximum-Likelihood Schätzers nicht berechnet werden konnten, aus den Daten heraus gefiltert. Wichtig hierbei ist jedoch, dass diese drei Rasch-Modelle aufgrund der Probleme mit dem Parameter-Schätzer nicht validiert werden konnten, da die Grösse der Datensätze zu gering war. Um diesen Vergleich sinnvoll durchzuführen, bräuchte es einen neuen, besseren Schätzer. Der marginal Maximum-Likelihood Schätzer konnte deutlich weniger Personenparameter schätzen, als der bedingte Maximum-Likelihood Schätzer.

Die Ergebnisse der Test befinden sich in den Darstellungen 4.7 und die wichtigsten Test Parameter in Tabelle 4.11.

Code erhältlich auf:

GitHub

<http://git.io/FVjL>

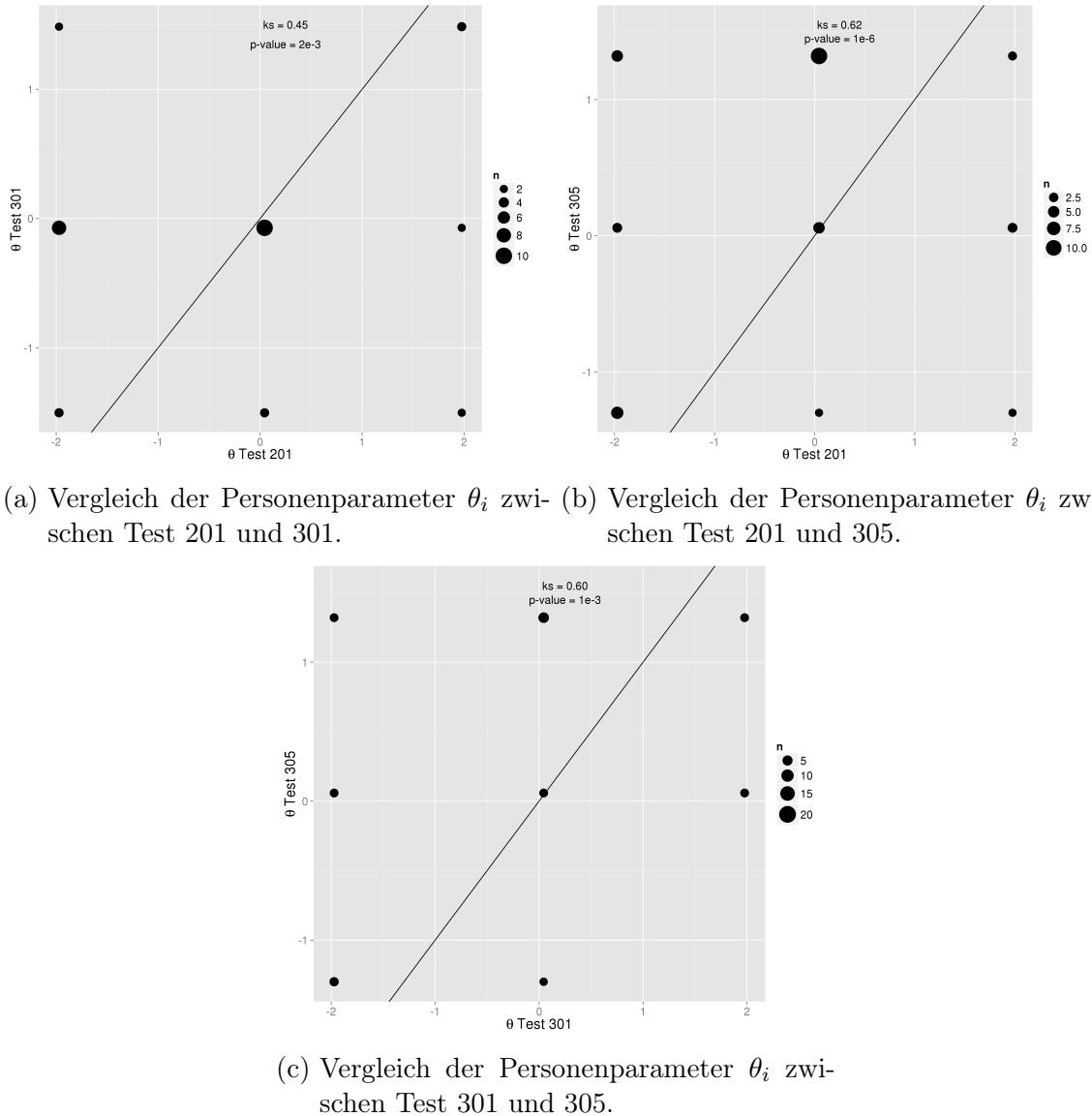


Abbildung 4.7: Vergleich der Personenparameter für die drei Tests. Zusätzlich sind der p-Wert des Kolmogorow-Smirnow-Test und die Test-Statistik ks angegeben.

201 vs 301			201 vs 305			301 vs 305		
p-Wert	ks	n	p-Wert	ks	n	p-Wert	ks	n
2e-3	0.45	33	1e-6	0.62	37	1e-3	0.60	20

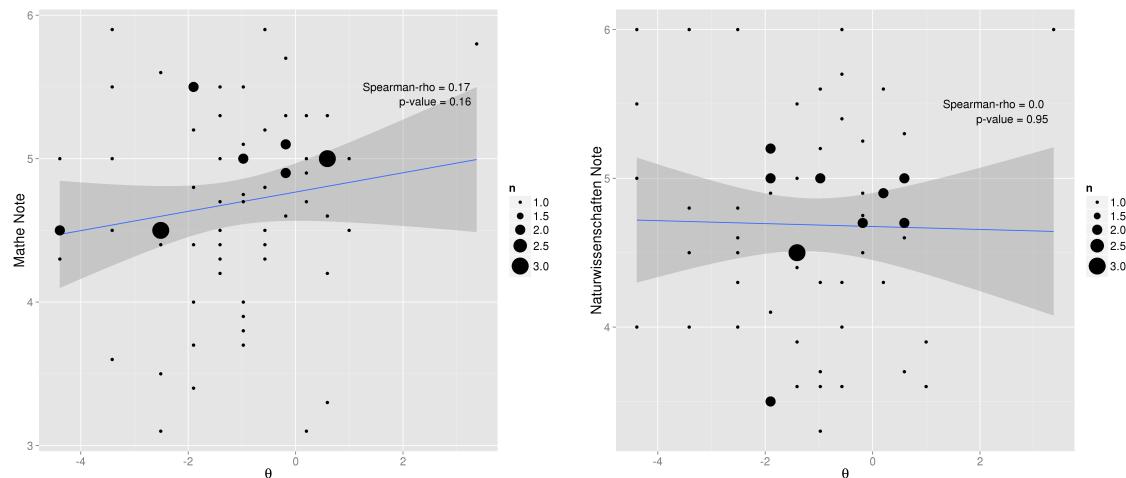
Tabelle 4.11: Resultate des Kolmogorow-Smirnow-Test für die Übereinstimmung der Personenparameter zwischen den drei Tests. Wobei ks die Test-Statistik des Kolmogorow-Smirnow-Test ist. Mit n wird die Anzahl an Personenparametern angegeben, welche für den Test verwendet werden konnten.

4.5.5 Zusammenhang zwischen Rasch-Modell und Fragebogen

Hierfür wurde wieder das erste Rasch-Modell verwendet, bei dem die Qualitätsstandards 1 bis 3 als Items verwendet wurden und pro Person nur eine Personenfähigkeit geschätzt wurde. Die so geschätzten Personenfähigkeiten wurden mit den Ergebnissen des Fragebogens korreliert. Die Resultate befinden sich in den Darstellungen und die Testergebnisse nochmals zusammengefasst in Tabelle 4.12.

Note Mathe		Note NatW.		SESSKO		Selbskonzept Schulversuche	
p-Wert	ρ	p-Wert	ρ	p-Wert	ρ	p-Wert	ρ
0.16	0.17	0.95	0.0	0.46	0.09	0.04	0.23

Tabelle 4.12: Spearmans ρ und p-Werte für die Korrelation zwischen der Personenfähigkeit θ und verschiedenen Skalen.



- (a) Korrelation der Personenfähigkeit θ mit der Note in Mathe. (b) Korrelation der Personenfähigkeit θ mit der Note in den Naturwissenschaften.

Code erhältlich auf:

GitHub

<http://git.io/FwCx>

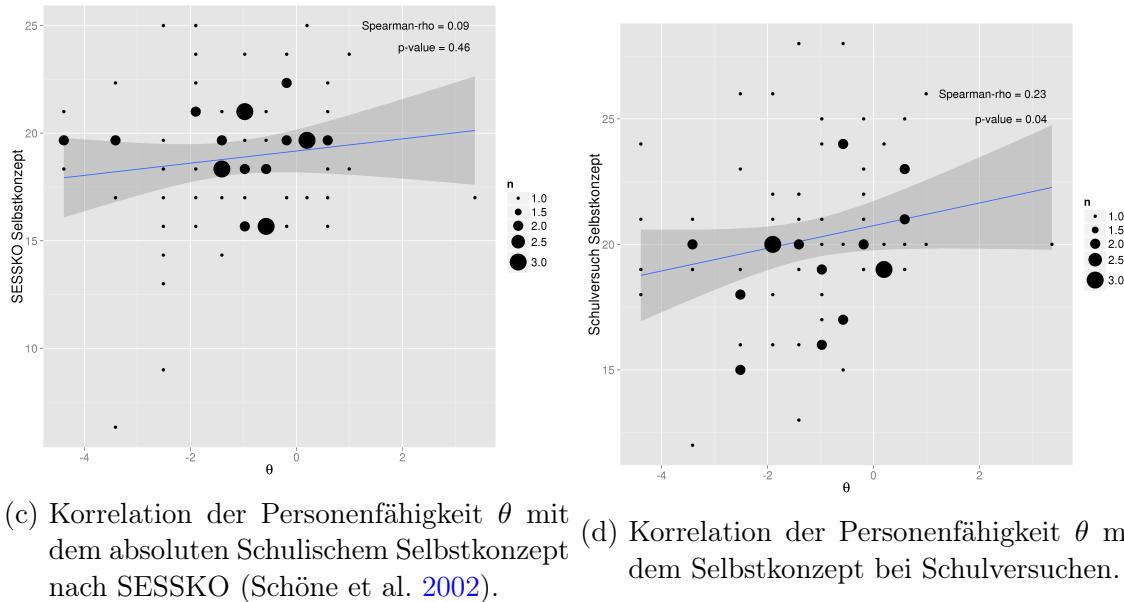


Abbildung 4.8: Korrelation zwischen der Personenfähigkeit θ und verschiedenen per Fragebogen erhobenen Daten. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die blaue Gerade ist die lineare Regression der zugrunde liegenden Daten, der dunkelgraue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

4.6 Videoanalyse

Insgesamt sind vier Stunden Videomaterial angefallen. Es wurde, wie bereits erwähnt, nur in einer Halbklasse eine Videoaufnahme durchgeführt. Aufgrund der Position der Videokamera konnten nur die Aktionen von je zwei Schülerinnen und Schülern analysiert werden. So konnten von 8 Schülerinnen und Schülern die Aktionen per Video analysiert werden.

4.6.1 Qualitätsstandards

Es wurden die existierende Qualitätsstandards auf Überprüfbarkeit per Video analysiert und die Qualitätsstandards 1 und 4 als analysierbar identifiziert. Für diese beiden Standards wurde jeweils eine Kodierung definiert.

Korrekt und präzise messen

Es wurde eine Kodierung, welche an Schreiber (2012) angelehnt war, verwendet. Bei der Kodierung des Merkmals korrekt und präzise messen, wurde von einer Gütestufe von 3 ausgegangen. Wenn ein Schüler oder Schülerin nicht korrekt abgemessen hat (z.B. schräg abgelesen), wurde Gütestufe 2 kodiert. Wenn die Schülerin oder der Schüler bei den einzelnen Messungen unterschiedlich gemessen hat, wurde die Gütestufe 1 vergeben.

Messung wiederholen

Test	Messung korrekt			Messwiederholung		
	1	2	3	1	2	3
201	0.25	0.63	0.13	0.63	0.38	0.00
301	0.13	0.75	0.13	0.63	0.38	0.00
305	0.25	0.63	0.13	0.38	0.38	0.25

Tabelle 4.13: Die erreichten Gütestufen für die Merkmale Messung wiederholen und korrekt und präzise messen. Die Anzahl kodierter Personen beträgt 8.

Bei diesem Merkmal wurde von einer Gütestufe von 1 ausgegangen. Wenn der Schüler oder die Schülerin die Messung wiederholt hat, wurde die Gütestufe 2 erreicht. Als Messwiederholung wurde eine Messung in einem neuen Experiment definiert. Es reichte also nicht, mehrmals den Messwert abzulesen um diese Gütestufe zu erreichen,

sondern es musste das Experiment erneut durchgeführt werden. Gütestufe 3 wurde erreicht, wenn das Experiment identisch durchgeführt wurde.

Die Resultate der Kodierungen befinden sich in Tabelle 4.13.

4.6.2 Korrelation zwischen Video-Merkmalen und Qualitätsstufen

Da die Videokodierung Merkmal basierend auf den Qualitätsstandards entwickelt hat, wurde untersucht, ob zwischen den Merkmalen und den Qualitätsstandards eine Korrelation existiert. Diese Resultate befinden sich in Darstellung 4.9 und in Tabelle 4.14. In keinem der Korrelationstests wird die Signifikanzschwelle überschritten, daher gibt es keine signifikante Korrelation zwischen den Qualitätsstandards und den Merkmalen der Videokodierung.

201 Q1		201 Q4		301 Q1		301 Q4	
p-Wert	ρ	p-Wert	ρ	p-Wert	ρ	p-Wert	ρ
0.76	-0.13	1.00	0.0	1.00	0.0	1.00	0.0

305 Q1		305 Q4	
p-Wert	ρ	p-Wert	ρ
0.53	0.26	0.87	0.07

Tabelle 4.14: Spearmans ρ und p-Werte für die Korrelation zwischen Qualitätsstandards und den Merkmalen aus der Videokodierung..

4.6.3 Messzeitpunkte und Messdauer

Zusätzlich zu den zwei Merkmalen wurde für jede Messung noch erhoben, wann die Messung begonnen und wann sie beendet wurde. Bei der Temperaturmessung war die Definition der Messung nicht trivial. Es wurde folgende Definition für eine Messung verwendet: Für eine Temperaturmessung muss dass Thermometer aus dem Medium entfernt und abgelesen werden. Ein Ablesen ohne dass das Thermometer aus dem Medium herausgenommen wird, gilt nicht als Messung. Der Hauptgrund für diese eingeschränkte Definition ist, dass der Ablesevorgang nur sehr schwierig eindeutig beobachtbar ist. Daher wurde dieser mit dem Entfernen des Thermometers verknüpft, sodass die Kodierung einfacher ist. Ein Problem dabei war der Test 201, da dort die

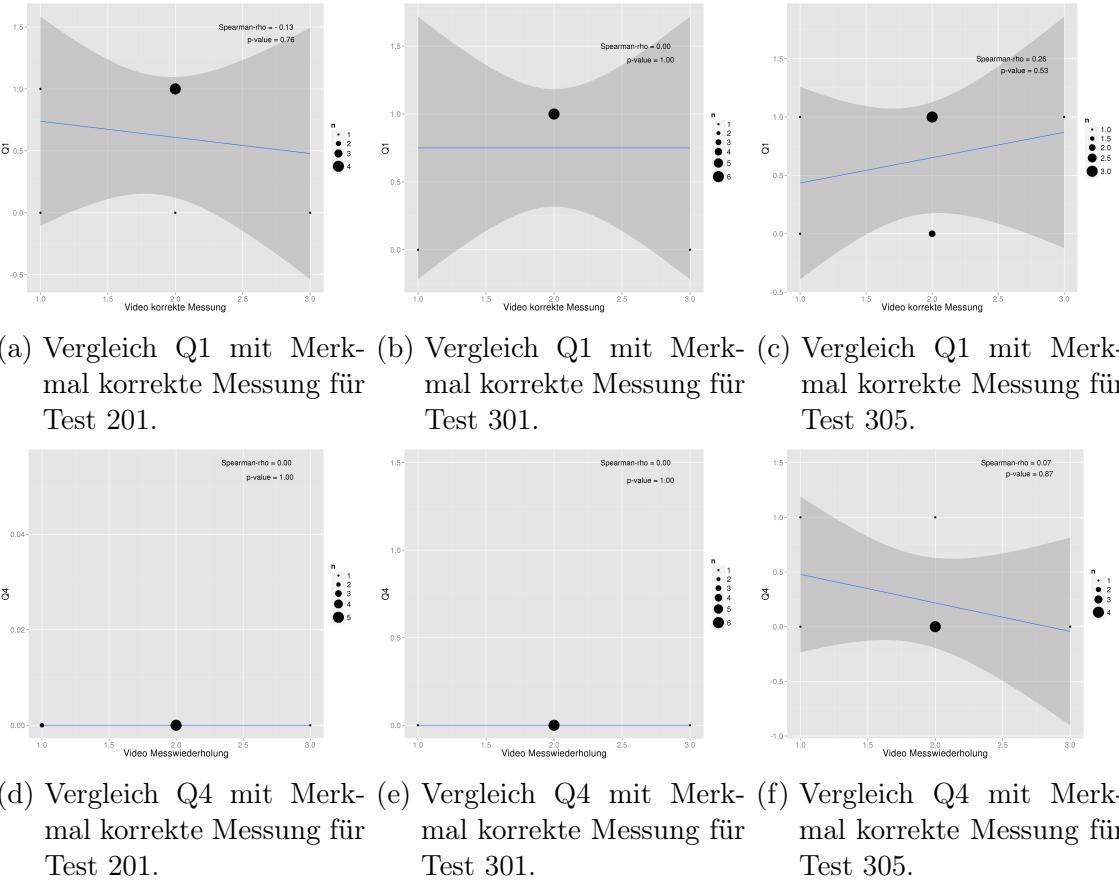


Abbildung 4.9: Vergleich der Merkmale der Videokodierung mit den Qualitätsstandards 1 und 4. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die blaue Gerade ist die lineare Regression der zugrundeliegenden Daten, der dunkelgraue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

Thermometer über das Video nicht unterscheidbar waren. Daher wurden dort die Messinstrumente mit 1 und 2 kodiert. Die Resultate finden sich in Darstellung 4.10.

Code erhältlich auf:

GitHub

<http://git.io/bvQS>



Abbildung 4.10: Übersicht über alle Messungen der videografierten Schülerinnen und Schüler. In der ersten Spalte ist der Identifizierungs-Code. Die Tests wurden hier unterschiedlich kodiert, es gilt 1 = 305, 2 = 201, 3 = 301. In Schwarz wird jeweils markiert, wann eine Messung durchgeführt wird. Die Linie in der Mitte entspricht der Halbzeit des Testes (10 min).

5 Diskussion

Nachdem im letzten Kapitel die Ergebnisse präsentiert wurden, soll in diesem Kapitel versucht werden mit Hilfe der Ergebnisse die Forschungsfrage zu beantworten.

5.1 Kodierung

5.1.1 Items

Da sowohl die Qualitätsstandards als auch die Niveaus auf den Items basieren, ist eine gute Kodierung derselben elementar. Durch die Zweitkodierung der Items sollte sichergestellt werden, dass die Kodierung der Items verlässlich und wiederholbar ist. In Tabelle 4.1 sind die Ergebnisse für die Interrater-Reliabilität aufgeführt. Bis auf wenige Ausnahmen befinden sich alle Werte oberhalb von $\kappa > 0.75$, was nach Greve und Wentura (1997, S.111) sehr gut bis ausgezeichnet ist. Landis und Koch (1977) bezeichnet jedoch auch die niedrigen κ -Werte, bei denen $\kappa > 0.61$ ist, als „substantial strength of agreement“.

Ein Problem bei der Kodierung der Items und der Überprüfung war jedoch, dass viele Schülerinnen und Schüler bestimmte Items nicht erreichten. Daher konnte Cohens κ nicht für alle Items berechnet werden. Da die Übereinstimmung dort jedoch sehr hoch war, kann auch bei diesen Items von einer korrekten Kodierung ausgegangen werden. Dieses Problem kann auch eine Erklärung für die sehr gute Übereinstimmung bei bestimmten Items sein. So war es meistens sehr klar, wenn ein Schüler oder eine Schülerin ein Item nicht erreicht hatte. Daher war die Kodierung meistens sehr eindeutig.

Aufgrund dieser Ergebnisse kann davon ausgegangen werden, dass die Zweitkodierung aller Schülerinnen und Schüler keine deutlich abweichende Resultate liefert hätten und daher die Zweitkodierung von 15% der Schülerinnen und Schüler ausreichend war um die Qualität und Reliabilität der Kodierung festzustellen.

Daher kann davon ausgegangen werden, dass die Reliabilität der Kodierung gegeben und die Kodierung korrekt und nachvollziehbar ist.

5.1.2 Qualitätsstandards

Ein Problem bei der Definition der Qualitätsstandards ist die unterschiedliche Definition in der Literatur. So verwendete Gut et al. (2014) noch eine andere Reihenfolge der Qualitätsstandards. Die in dieser Arbeit verwendete Reihenfolge der Qualitäts-

standards basiert auf den Arbeiten von Metzger et al. (2013) und Gut et al. (2014). Ein Problem dabei ist jedoch, dass die Schwellenwerte für das Erreichen der Qualitätsstandards nicht publiziert sind. Die Schwellenwerte wurde daher von internen Dokumenten von Pitt Hild übernommen.

Die erreichten Qualitätsstandards in Tabelle 4.2 zeigen, dass insbesondere die Qualitätsstandards 3, 4 und 5 nur von einem geringen Prozentsatz der Schülerinnen und Schüler erreicht werden. Und es auch einen Unterschied in den erreichten Qualitätsstandards zwischen den einzelnen Test gibt. In dieser Arbeit wird nicht auf diese Unterschiede eingegangen. Dafür sei auf folgende Arbeit hingewiesen: Sichau (2015). Dieser Unterschied in den erreichten Qualitätsstandards deckt sich jedoch mit den Ergebnissen von Metzger et al. (2013).

5.1.3 Niveaus

Dieses schlechte Abschneiden der Klassen spiegelt sich auch in den erreichten Niveaus wieder. So sieht man in Tabelle 4.3, dass ein Grossteil der Schülerinnen und Schüler nicht über das Niveau 2 hinauskommt, sowohl beim unbedingten als auch beim bedingten Niveau. Im Vergleich zu Metzger et al. (2013) schneiden die Schülerinnen und Schüler in der 7. Klasse schlechter ab.

Da leider der Zeitpunkt der Datenerhebung in der Arbeit von Metzger et al. (2013) nicht aufgeführt ist, ist nicht klar ob der frühe Zeitpunkt des Testes (Beginn des ersten Halbjahres) einen eventuellen Einfluss auf das Abschneiden der Schülerinnen und Schüler hatte. So war dies bei allen Klassen bei denen diese Tests durchgeführt wurden, das erste Mal, dass sie in der Oberstufe experimentiert haben. Auch kannten die Schülerinnen und Schüler den Kraftmesser nicht und konnten nur durch ausprobieren herausfinden, wie dieser funktioniert. Daher die Vermutung, dass wenn der Test im zweiten Halbjahr der 7. Klasse durchgeführt worden wäre ein deutlich besseres Resultat hätte erzielt werden können.

5.2 Fragebogen

Die verwendeten Fragen im Fragebogen aus SESSKO (Schöne et al. 2002) und die abgewandelten Fragen nach Dierks, Höffler und Parchmann (2014) wurden auf innere Konsistenz überprüft. Beide Skalen erreichten wie in 4.2 beschrieben eine sehr gute innere Konsistenz, insbesondere da Cronbachs Alpha eher zu einer Unterschätzung der inneren Konsistenz führt (Eisinga, Grotenhuis und Pelzer 2013). Auch durch das Weglassen einzelner Fragen würde die innere Konsistenz nicht verbessert werden (siehe Tabelle 4.4 und Tabelle 4.5). Daher kann angenommen werden, dass beide Skalen das jeweilige Selbstkonzept konsistent widerspiegeln und ausreichend Fragen zu jeder Skala vorhanden sind.

Der Mittelwert aller Schülerinnen und Schüler beim „Schulisches Selbstkonzept - absolut“ kann mit den Werten aus der Literatur (Schöne et al. 2002) verglichen werden. Dabei hat die hier untersuchte Schülergruppe ein leicht überdurchschnittliches Selbstkonzept verglichen mit der Referenzgruppe (4. - 10. Klasse in verschiedenen Deutsch Schulformen und Bundesländern.). Der Grund dafür könnte der erst kürzlich erfolgte Übertritt auf die Oberstufe und dort die Einteilung in die Sek A sein.

5.3 Unterschiede zwischen den Klassen

Vor der weiteren Analyse der Daten musste erst festgestellt werden, ob die Datensätze der einzelnen Klassen kombiniert werden dürfen. Wichtig ist dabei, dass der exakte Test nach Fischer verwendet wird und nicht der Chi-Quadrat-Test, da bei kleinen Datensätzen (wie dem hier Vorliegenden) der Chi-Quadrat-Test sich nicht eignet (Mehta, Patel und Tsiatis 1984).

Für den exakten Fischer-Test wurden die erreichten Qualitätsstandards in den einzelnen Klassen verglichen. Die Qualitätsstandards wurden verwendet, da im Vergleich zu den Items das statistische Rauschen geringer ist und gleichzeitig nicht viel an Information verloren geht. Aus der Tabelle 4.6, kann geschlossen werden, dass kein signifikanter Unterschied zwischen den einzelnen Klassen existiert, da alle p-Werte über 0.05 liegen.

Es dürfen daher alle Datensätze kombiniert werden, da das Erreichen eines Qualitätsstandards nicht davon abhängt, in welcher Klasse ein Schüler oder eine Schülerin ist. Für alle weiteren Analysen wurden daher alle Datensätze kombiniert und nicht nach Klassen unterschieden.

5.4 Ist das Abschneiden in den Tests unterschiedlich

Nachdem gezeigt wurde, dass der Datensatz gesamthaft analysiert werden kann, wurde versucht die Forschungsfrage zu beantworten. Dafür ist es notwendig festzustellen, ob das Erreichen der Qualitätsstufen zwischen den unterschiedlichen Tests signifikant unterschiedlich ist.

Hierbei gibt es unterschiedliche Ergebnisse, wie in Tabelle 4.7 ersichtlich ist. So ist die Korrelation zwischen den unbedingten Niveaus zwischen allen Tests signifikant. Das Spearmans ρ liegt jeweils im leicht positiven Bereich, was auf eine leicht positive Korrelation hinweist. Bei dem bedingten Niveau ist nur der Test zwischen Test 201 und 305 signifikant.

Ein Grund für diese unterschiedlichen Resultate liegt vermutlich darin, dass beim bedingten Niveau nur sehr wenig hohe Werte erreicht wurden (siehe Tabelle 4.3). Daher

kommt es zu einer schlechten Datenmenge bei Niveaus über 2; dies kann man auch sehr gut in den Darstellungen 4.1 sehen. Dies führt zu Problemen bei der Berechnung des Korrelationstestes für bedingte Niveaus, da nur sehr wenige Datenpunkte im Bereich über 2 verfügbar sind, an denen eine Verankerung stattfinden könnte. Bei besseren Schülerinnen und Schülern, bei denen häufiger ein höheres Niveau erreicht würde, wären diese Probleme nicht so fatal und man würde vermutlich bei beiden Niveaus eine Korrelation feststellen können.

Aufgrund der geringen Datenmenge bei den bedingten Niveaus, wird der Fokus in der weiteren Arbeit auf die unbedingten Niveaus gesetzt. Aufgrund der Korrelationen zwischen diesen kann davon ausgegangen werden, dass das Erreichen eines unbedingten Niveaus in einem Test mit dem unbedingten Niveau in einem anderen Test signifikant leicht positiv korreliert. Dies ist ein erster Hinweis darauf, dass das Erreichen eines Niveaus nicht davon abhängig ist, in welchem Test dies erreicht wurde, sondern ausschliesslich von der Kompetenz des skalenbasierten Messens abhängt.

5.5 Rasch-Analyse

Nachdem sich in der klassischen Testtheorie erste Hinweise auf die Beantwortung der Forschungsfrage gezeigt haben, wurde zusätzlich die probabilistische Testtheorie verwendet. Ein Grund, diese Theorie zu verwenden, ist, dass das Abgeben einer korrekten Antwort ein Zufallsprozess ist und nicht deterministisch. Dieser Zufallsprozess hängt jedoch von der Aufgabenschwierigkeit und der latenten Personenfähigkeit ab. Aufgrund der zugrundeliegenden Daten wurde das dichotome Rasch-Modell verwendet. Für das Modell wurden nur die unbedingten Qualitätsstandards verwendet, da die bedingten Qualitätsstandards die Annahme des Rasch-Modells, dass alle Items unabhängig voneinander sind, verletzen.

5.5.1 Parameterschätzung

Ein grosses Problem bei der Rasch-Analyse ist die Parameterschätzung. Die grösste Schwierigkeit dabei ist, dass es im Moment in der Literatur nur zwei gängige Parameterschätzer gibt, welche im Detail analysiert wurden (Fischer und Molenaar 1995; Rost 2004; Strobl 2012). Wie bereits geschrieben machen diese beiden Parameterschätzer Annahmen über die zugrundeliegenden Daten. Bei den vorliegenden Daten kann insbesondere die Annahme über eine bestimmte Verteilung (der Einfachheit halber wird meistens eine Normalverteilung angenommen (Rost 2004)) der Personenfähigkeiten aufgrund der zugrundeliegenden Daten nicht angenommen werden.

Mit beiden Parameterschätzern können die Aufgabenschwierigkeiten β übereinstimmen geschätzt werden (siehe Darstellung 4.2). Nach Rost (2004) ist diese Schätzung jedoch deutlich unkritischer als die der Personenparameter. Bei den Personenpara-

metern θ gibt es jedoch Unterschiede zwischen beiden Schätzern. Bei der bedingen Maximum-Likelihood-Schätzung können alle Personenparameter ohne Extrapolation berechnet werden. Dies ist bei der marginalen Maximum-Likelihood-Schätzung nicht der Fall. Der Grund dafür liegt in der Annahme einer Normalverteilung der Personenparameter, die der marginalen Maximum-Likelihood-Schätzung zugrunde liegt. Bei grösseren Datensätzen mag diese Annahme gerechtfertigt sein, bei dem hier vorliegenden Datensatz ist dieser Schätzer jedoch nicht geeignet. Es wäre zwar prinzipiell möglich eine andere Verteilung als die Normalverteilung für die Personenparameter zu verwenden. Dafür müsste aber eine eigene Implementierung des Rasch-Modells vorgenommen werden, was den Rahmen dieser Arbeit sprengen würde.

Aufgrund dieses Vergleichs der Parameterschätzungen wurde für alle weiteren Rasch-Modelle in dieser Arbeit der bedingte Maximum-Likelihood-Schätzer verwendet. Dessen Annahmen, dass jeder Schüler oder Schülerin mindestens ein Item richtig oder falsch beantwortet haben muss, war jedoch bei der Aufteilung in kleinere Rasch-Modelle ein Problem. Daher sollten insbesondere für kleine Datensätze bessere Schätzer entwickelt werden, welche weniger Annahmen über die zugrundliegenden Daten machen. Eine Möglichkeit wäre ein Bootstrapping Algorithmus, welcher die Verteilung der Personenparameter aus den vorliegenden Daten selbst abschätzt und die Verteilung dann in den marginalen Maximum-Likelihood Schätzer einsetzt.

5.5.2 Modellkontrolle

Nachdem der beste Parameterschätzer identifiziert wurde, musste das Rasch-Modell jedoch noch verifiziert werden. Dafür wurde das Rasch-Modell basierend auf dem Mittelwert der Personenrandsummen gesplittet. Aufgrund der Annahmen für das Rasch-Modell sollten dann keine signifikanten Unterschiede zwischen den beiden neuen Modellen existieren. Dies wurde vom Andersens Likelihood-Quotienten Test bestätigt, nach dem die Qualitätsstufen 4 und 5 entfernt wurden. Das Problem mit diesen beiden Qualitätsstufen ist, dass für die untersuchte Personengruppe diese Standards sehr schwierig waren und sie daher kaum beantwortet wurden (siehe Tabelle 4.2). Aufgrund der Testergebnisse kann das Ausschliessen dieser Qualitätsstufen bestätigt werden, da dann ein valides Rasch-Modell vorliegt.

Zusätzlich wurden alle Qualitätsstandards noch überprüft, sowohl grafisch (siehe Darstellung 4.3), als auch mit dem Wald-Test (siehe Tabelle 4.8). Es gab dabei kein Qualitätsstandard, welcher als ungeeignet aus dem Modell ausgeschlossen werden musste, da er sich signifikant in den beiden Modellen unterschiedet.

Diese Resultate zeigen, dass das verwendete Rasch-Modell mit den Qualitätsstandards 1-3 valide ist. Dieses Resultat ist wichtig, da ansonsten die mit diesem Modell gewonnenen Parameter auf einer falschen Modellannahme beruhen würden.

5.5.3 Unterschied in den Schwierigkeiten der Qualitätsstandards

Die Schwierigkeit eines Qualitätsstandards sollte nicht davon abhängig sein, in welchem Test dieser Qualitätsstandard erreicht wurde. Dies wurde versucht mit Hilfe des Rasch-Modells zu verifizieren. Dazu wurden die *item characteristic curves* (ICC) gezeichnet, siehe Darstellung 4.4. Diese Darstellung lässt eine qualitative Überprüfung der Schwierigkeiten zu. Man sieht das bei Qualitätsstandard 1 und 3 die beiden Test 201 und 301 sehr ähnlich sind. Bei Test 305 sind die Qualitätsstandards meistens deutlich leichter im Bezug auf den Schwierigkeitsgrad. Dies liegt höchstwahrscheinlich daran, dass dieser Test im Vergleich zu den anderen beiden Test leichter ist (Sichau 2015). Dies sieht man auch in der Darstellung 4.5.

Zusätzlich zu der qualitativen Überprüfung wurde noch ein Kolmogorow-Smirnow-Test durchgeführt, um festzustellen, ob die Unterschiede in den Aufgabenparametern (siehe Tabelle 4.9) signifikant sind. Die Testergebnisse in Tabelle 4.10 zeigen, dass es keine signifikanten Unterschiede zwischen diesen Werten gibt. Wichtig ist dabei jedoch, dass diese Tests eine sehr geringe Power haben, da der Datensatz nur die Grösse von 3 hatte. Diese geringe Power zeigt sich auch in der Darstellung 4.6.

Durch die Kombination der qualitativen und quantitativen Resultaten kann jedoch die Aussage gestützt werden, dass es keine signifikanten Unterschiede in der Schwierigkeiten der Qualitätsstandards gibt. Dies ist ein weiterer Hinweis darauf, dass das Erreichen der Qualitätsstandards 1-3 nicht davon abhängig ist, welcher Test durchgeführt wurde.

5.5.4 Unterschied in den latenten Personenfähigkeiten

Nachdem es klar ist, dass die Aufgabenparameter sehr ähnlich sind, wurden die Personenparameter analysiert. Hierfür wurde das Rasch-Modell aufgeteilt und für jeden Test ein eigenes Rasch-Modell erstellt. Hierbei gibt es nun massive Probleme mit der Parameterschätzung, da nun die Wahrscheinlichkeit, dass ein Schüler keinen der drei Qualitätsstandards oder alle erreicht hat, signifikant höher ist. Daher konnten viele Personenparameter nicht geschätzt werden.

Diese Probleme mit der Parameterschätzung führten auch dazu, dass das Modell nicht validiert werden konnte. Die gewonnenen Personenparameter basieren daher auf einem nicht validierten Rasch-Modell und müssten daher mit Vorsicht interpretiert werden. Diesmal wurde untersucht, ob sich die Personen-Fähigkeiten zwischen den drei Rasch-Modellen unterscheiden. In Tabelle 4.11 und Darstellung 4.7 sind die Resultate dieses Testes dargestellt. Es kann daher davon ausgegangen werden, dass die Personenfähigkeiten zwischen den drei Tests nicht signifikant korrelieren.

Diese Resultate sind ein Gegenindiz zu den bisher vorliegenden Resultaten, da die Personenfähigkeit nicht von den durchgeführten Tests abhängen sollten. Aufgrund der

Datengrundlage und dem darauf basierenden Rasch-Modell sollten diese Ergebnisse jedoch nicht überbewertet werden, insbesondere da das Rasch-Modell nicht validiert werden konnte. Auch sieht man in Tabelle 4.11, dass meistens nur ein kleiner Teil der Personenparameter verglichen wurde, da der Schätzer nur für einen kleinen Teil der Personen fähig war den Personenparameter θ zu berechnen. Diese Ergebnisse beruhen daher Grossteils auf Problemen mit dem Parameterschätzer. Auch der marginale Maximum-Likelihood Schätzer hatte massive Probleme mit dem Datensatz und war noch ungeeigneter, daher wurden dessen Ergebnisse nicht präsentiert.

Aufgrund dieser Probleme sollten diese Gegenindizien nicht überinterpretiert werden, da sie auf einer sehr unsicheren Datengrundlage basieren. Dies zeigt jedoch, dass bessere Parameterschätzer notwendig sind, welche auch mit solchen Datensätzen umgehen können.

5.5.5 Zusammenhang zwischen Rasch-Modell und Fragebogen

Das erste Rasch-Modell, bei dem alle drei Test kombiniert wurden, wurde verwendet, um die latente Personenfähigkeit mit Resultaten des Fragebogens zu vergleichen. In Tabelle 4.12 sind die Ergebnisse der Korrelations-Test dargestellt. Es gibt nur einen signifikanten Zusammenhang zwischen dem Schulversuch-Selbstkonzept.

Dieses Ergebnis ist nicht überraschend, da in der Notengebung experimentellen hand-ons Test eher eine untergeordnete Rolle spielen. Auch das SESSKO Selbstkonzept (Schöne et al. 2002) ist vermutlich zu generell und korreliert daher nicht mit den Personenfähigkeiten des Rasch-Modells. Das letzte Selbstkonzept hingegen zielt sehr genau auf das Selbstkonzept bei Schulversuchen ab, welche nahezu identisch zu experimentellen hand-ons Test sind. Daher ist diese Korrelation zu erwarten. Um diese Skala jedoch zu verbessern, müsste diese noch im grösseren Rahmen validiert werden. Vor allem ist im Moment noch keine Normalverteilung der Daten gewährleistet.

5.6 Videoanalyse

In einem letzten Schritt wurden die Videos analysiert. Die dabei entwickelten Merkmale wurden mit den Qualitätsstandards korreliert. Wie in Tabelle 4.14 und Darstellung 4.9 ersichtlich gibt es keinen Zusammenhang zwischen den im Video kodierten Merkmalen und den Qualitätsstandards, auf denen die Merkmale beruhen. Diese Ergebnisse sind zuerst enttäuschend, da die Merkmale eigentlich die Qualitätsstandards widerspiegeln sollten. Mit Beobachtungen, welche jedoch während der Testdurchführung gemacht wurden, lassen sich diese Ergebnisse jedoch erklären. Viele Schülerinnen und Schüler waren während der Testdurchführung stark auf die experimentelle Seite fokussiert und haben insgesamt sehr wenig der Datenbögen ausgefüllt. Dies zeigt sich auch im insgesamt eher schlechten Abschneiden der Schülerinnen und Schüler (siehe

Tabelle 4.3). Daher widerspiegeln die Qualitätsstandards nur den Teil des Experiments wieder, welche die Schülerinnen und Schüler dokumentiert haben.

Diese Resultate zeigen jedoch klar, dass für die Kompetenz des skalenbasierenden Messens auch der Aspekt der Dokumentation eine entscheidende Rolle spielt. Dies widerspricht sich jedoch nicht, da zu einer experimentellen Kompetenz die Fähigkeit zu Dokumentieren sehr wichtig ist. Für Schülerinnen und Schüler jedoch, welche sprachliche Schwächen haben, könnte der Einsatz von Videoanalysen hilfreich sein. Auch bei niedrigeren Schulstufen, wäre der Einsatz von Videoanalysen angebracht. Ein Nachteil ist jedoch der hohe Aufwand, welcher für die Kodierung der Videos anfällt.

Ein weiteres Problem ist die Interpretierbarkeit der Daten. So ist es sehr schwierig aus der Darstellung 4.10 sinnvolle Schlüsse zu ziehen. Diese Daten sind nur qualitativ analysierbar. Solange aber dieser Datensatz nicht grösser ist, sollten aus diesen Daten auch keine qualitativen Schlüsse gezogen werden.

5.7 Zusammenfassung

Abschliessend lässt sich sagen, dass sowohl mit der klassischen als auch mit der probabilistische Testtheorie die Forschungsfrage beantwortbar ist. Mit beiden Theorien konnten starke Hinweise darauf gefunden werden, dass bei dem vorliegenden Datensatz die Kompetenz des skalenbasierten Messens unabhängig vom fachlichen oder inhaltlichen Kontext ist. Es gibt zwar auch Gegenanzeigen gegen dieses Resultat, bei diesen ist aber oft die Datengrundlage sehr schlecht, im Vergleich zu den unterstützenden Hinweisen. Daher kann die Forschungsfrage mit der durchgeföhrten Methode beantwortet werden. Bevor aber generelle Schlüsse gezogen werden, müsste die Untersuchungsgruppe massiv vergrössert werden.

Das Resultat dieser Arbeit ist daher, folgendes:

Die Kompetenz des skalenbasierten Messens in der vorliegenden Untersuchungsgruppe ist unabhängig vom fachlichen oder inhaltlichen Kontexten.

6 Ausblick

6.1 Datengrundlage

Mit dieser Arbeit wurde ein erster Versuch durchgeführt zu zeigen, dass bestimmte Kompetenzen kontextunabhängig sind. Für eine Generalisierbarkeit der Resultate sind jedoch grössere Untersuchungsgruppen notwendig. Daher sollte ein erster Schritt dahin gehen, die Datengrundlage dieser Arbeit zu vergrössern. Damit sollten die bisher vorliegenden Hinweise stärker hervortreten und die Korrelationen besser abschätzbar sein. Hierbei denke ich jedoch, dass die bisherigen Ergebnisse bestätigt werden und keine gegensätzlichen Resultate gefunden werden. Die bisherigen Ergebnisse sind jedoch aufgrund der zu geringen Stichprobe nicht generalisierbar.

6.2 Videoanalyse

In dieser Arbeit wurde versucht zusätzliche Informationen mit Videoanalyse zu generieren. Dies ist nur sehr beschränkt gelungen, insbesondere da der Hauptfokus auf quantitativen Forschungsmethoden gelegt wurde und nicht auf qualitative. Dennoch hat sich gezeigt, dass die Ergebnisse der Videoanalyse nicht mit den über Pen- und Paper-Tests erhobenen Daten übereinstimmen. Für genauere Analyse dieser Ergebnisse sollte der Fokus gezielt auf den Vergleich zwischen Videoanalyse und Pen- und Paper-Tests gelenkt werden.

Interessant könnten die Videoanalysen insbesondere bei sprachlich schwächeren Schülern und Schülerinnen sein, welche aufgrund sprachlicher Schwierigkeiten in Pen- und Paper-Tests nur schwache Leistungen zeigen. Insbesondere hier könnten Videoanalysen helfen, festzustellen, ob es wirklich ein sprachliches Problem ist oder ob diese Schüler und Schülerinnen in diesen Tests auch tatsächlich schlechtere Leistungen erbringen.

6.3 Methoden

Auch methodisch wirft diese Arbeit weitere Fragen auf: Wie gezeigt ist insbesondere die Parameterschätzung des Rasch-Modells bei kleinen Datensätzen problematisch. Da die bisherigen Ansätze meistens Annahmen treffen, welche von kleinen Datensätzen nicht erfüllt werden können. Daher bräuchte es dringend neue Ansätze für die Schätzung der Personenparameter. Eine Möglichkeit wäre der marginale Maximum-Likelihood Schätzer. Dieser erfordert eine Annahme über die Verteilung

der zugrundeliegenden Personenparametern. In den meisten existierenden Software-packages wird eine Normalverteilung angenommen (Rost 2004; Rizopoulos 2006). Diese Annahme einer Normalverteilung ist nicht festgelegt für den marginale Maximum-Likelihood Schätzer. Dieses Problem könnte vielleicht mit einem neuen Bootstrapping-Algorithmus, welcher mit den vorliegenden Daten eine Abschätzung über die Verteilung der Personenparametern macht, gelöst werden. Diese Abschätzung könnte dann als Initialisierung für den marginal Maximum-Likelihood Schätzer verwendet werden. Durch weitere Iterationen könnte dann das Ergebnis eventuell noch verbessert werden. Dies würde insbesondere bei kleinen Datensätzen das Rasch-Modell verbessern und nützlicher machen.

Ein weiteres grosses Problem der sozialwissenschaftlichen Forschung ist die geringe Auseinandersetzung mit den verwendeten Methodiken. Auch das Verwenden von closed-source Programmen ist sehr fragwürdig, da oft die Dokumentation nicht ausreichend ist, um die Ergebnisse nachvollziehen zu können (z.B. SPSS). Meiner Meinung nach hat hier die Literatur in der sozialwissenschaftlichen Forschung grossen Nachholbedarf. So sollten die verwendeten Source-Codes für Analysen frei verfügbar gemacht werden (insbesondere bei Publikationen), damit andere Personen die Resultate nachvollziehen können, wie dies z.B. in der Bio-Informatik Standard ist. Es wurde versucht diese Forderung in dieser Arbeit umzusetzen, daher hier nochmals der Link zum vollständigen Source-Code der Auswertung.

Code erhältlich auf:

GitHub

<http://git.io/buGR>

Literaturverzeichnis

- Anderson, John R., Lynne Reder und Herbert A. Simon Mai 1996. "Situated Learning and Education". 1996. *Educational Researcher* 25, Nr. 4 (Mai): 5–11.
- Barrows, Howard S. 1985. *How to design a problem-based curriculum for the preclinical years*. 1985.
- Baumert, Jürgen, Petra Stanat und Anke Demmrich 2001. "PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie". 2001. In *PISA 2000*, herausgegeben von Manfred Weiß. Opladen.
- Berner, Esther, und Stefanie Stolz 2006. *Literaturanalyse zu Entwicklung, Anwendung und insbesondere Implementation von Standards in Schulsystemen: Nordamerika*. 2006. Technischer Bericht. Zürich: Pädagogisches Institut der Universität Zürich.
- Bos, Wilfried, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Renate Valtin und Gerd Walther 2003. *Erste Ergebnisse aus IGLU*. 2003. April 2003.
- Bransford, John D., Ann L. Brown und Rodney R. Cocking 2000. *How people learn*. 2000. Technischer Bericht. Washington, D.C.: Commision on Behavioral, Social Sciences und Education.
- Brophy, Jere 1992. "Probing the subtleties of subject-matter teaching." 1992. *Educational Leadership*.
- Claxton, Guy 1990. *Teaching to Learn: A Direction for Education*. 1990. Cassell education. Cassell.
- Corte, Erik De 2003. "Designing learning environments that foster the productive use of acquired knowledge and skills". 2003. Kap. 2 in *Powerful Learning Environments: Unravelling Basic Components and Dimensions*, 21–33.
- Detterman, Douglas K 1993. "The case for the prosecution: Transfer as an epiphenomenon". 1993. Kap. 1 in *Transfer on trial: Intelligence, cognition, and instruction*, herausgegeben von Douglas K (ED) Detterman und Robert J (ED) Sternberg, 1–24. Ablex Publishing.
- Dierks, Pay Ove, Tim Höffler und Ilka Parchmann 2014. "Interesse von Jugendlichen an Naturwissenschaften Ist es wirklich so schlecht wie sein Ruf?" 2014. *CHEMKON* 21 (3): 111–116.
- Duncker, Karl, und Lynne S. Lees 1945. "On problem-solving." 1945. *Psychological monographs* 58 (5): i.

- EDK Schweizer Konferenz der Kantonalen Erziehungsdirektoren 2004. *HARMOS Zielsetzungen und Konzeption Juni 2004*. 2004. Technischer Bericht. Bern: EDK Schweizer Konferenz der Kantonalen Erziehungsdirektoren.
- Eisinga, Rob, Manfred Te Grotenhuis und Ben Pelzer 2013. "The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?" 2013. *International Journal of Public Health* 58:637–642.
- Fässler, Lukas 2007. "Das 4-Schritte-Modell". 2007. Diss.
- Ferguson, George A. September 1956. "On transfer and the abilities of man." 1956. *Canadian journal of psychology* 10, Nr. 3 (September): 121–31.
- Fischer, Gerhard H Ed, und Ivo W Ed Molenaar 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. 1995, 436.
- Fisher, Ronald A. 1922. "On the interpretation of χ^2 from contingency tables, and the calculation of P". 1922. *Journal of the Royal Statistical Society* 85 (1): 87–94.
- Gick, Mary L., und Keith J. Holyoak 1980. "Analogical problem solving". 1980. *Cognitive psychology* 355:306–355.
- Godden, D.R., und A.D. Baddeley 1975. "Context-dependent memory in two natural environments: On land and underwater". 1975. *British Journal of psychology*.
- Gott, Richard, und Sandra Duggan 1996. "Practical work: its role in the understanding of evidence in science". 1996. *International Journal of Science Education* 18 (7): 791–806.
- Gott, Richard, und Sandra Duggan 2002. "Problems with the Assessment of Performance in Practical Science: which way now?" 2002. *Cambridge Journal of Education* 32 (2): 183–201.
- Greeno, James G., Allan M. Collins und Lauren B. Resnick 1996. "Cognition and learning". 1996. In *Handbook of Educational Psychology*, herausgegeben von D. Berliner und R. Calfee, 14–46. New York.
- Greve, Werner, und Dirk Wentura 1997. *Wissenschaftliche Beobachtung: eine Einführung*. 1997. Weinheim: PVU/Beltz.
- Gut, Christoph, Pitt Hild, Susanne Metzger und Josiane Tardent 2014. "Projekt Ex-KoNawi: Modell für hands-on Assessments experimenteller Kompetenzen". 2014. In *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht*, herausgegeben von Sascha Bernholt, 171–173.
- Hartig, Johannes, und Eckhard Klieme 2006. "Kompetenz und Kompetenzdiagnostik". 2006. Kap. 3 in *Leistung und Leistungsdiagnostik*, herausgegeben von Karl Schweizer, 127–143. Springer Berlin Heidelberg.

- Hild, Pitt, Susanne Metzger und Ilka Parchmann 2014. "Using feedback and feed forward to foster experimental competence in student-centered learning environments". 2014.
- Huber, Christina, Martina Späni, Claudia Schmellentin und Lucien Criblez 2006. *Bildungsstandards in Deutschland, Österreich, England, Australien, Neuseeland und Südostasien*. 2006. Technischer Bericht. Aarau: Fachhochschule Nordwestschweiz Pädagogische Hochschule.
- Judd, Charles H. 1908. "The relation of special training to general intelligence". 1908. *Educational review* 36 (28-42).
- Killen, Roy 2000. *Outcomes-based education: Principles and possibilities*. 2000. Technischer Bericht. Faculty of Education, University of Newcastle, Australia.
- Klieme, Eckhard 2004. "Was sind Kompetenzen und wie lassen sie sich messen?" 2004. *Pädagogik* 6:10–13.
- Konsortium HarmoS Naturwissenschaften+ 2010. *Wissenschaftlicher Kurzbericht und Kompetenzmodell*. 2010. Technischer Bericht.
- Kowalski, Charles J. 1972. "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient". 1972. *Applied Statistics* 21 (1): 1.
- Kultusministerkonferenz 2004. *Bildungsstandards der Kultusministerkonferenz*. 2004.
- LaBerge, David, und S.Jay Samuels April 1974. "Toward a theory of automatic information processing in reading". 1974. *Cognitive Psychology* 6, Nr. 2 (April): 293–323.
- Landis, J R, und G G Koch 1977. "The measurement of observer agreement for categorical data." 1977. *Biometrics* 33 (1): 159–174.
- Lave, Jean 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. 1988.
- Lersch, Rainer 2007. "Kompetenzfördernd unterrichten. 22 Schritte von der Theorie zur Praxis". 2007. *Pädagogik* 59 (12): 36–43.
- Lobato, Joanne, und Daniel Siebert Januar 2002. "Quantitative reasoning in a reconceived view of transfer". 2002. *The Journal of Mathematical Behavior* 21, Nr. 1 (Januar): 87–116.
- Mair, Patrick, und Reinhold Hatzinger 2007. "Extended Rasch Modeling : The eRm Package for the Application of IRT Models in R". 2007. *Journal Of Statistical Software* 20 (9): 1–20.
- Marini, Anthony, Anne McKeough und Judy Lupart 1995. *Teaching for Transfer: Fostering Generalization in Learning*. 1995. Mallory International.

- Martin, Michael O., und Ina V.S. Mullis 2003. "Overview of TIMSS 2003". 2003. *TIMSS:2–21*.
- McGee, Clive März 1996. "The Development of a New National Curriculum in New Zealand". 1996. *The Educational Forum* 60, Nr. 1 (März): 56–63.
- Mehta, C R, N R Patel und a a Tsiatis 1984. "Exact significance testing to establish treatment equivalence with ordered categorical data." 1984. *Biometrics* 40 (3): 819–825.
- Metzger, Susanne, Pitt Hild, Christoph Gut und Josiane Tardent 2013. "Projekt Ex-KoNawi: Aufgaben und erste Ergebnisse der hands-on Assessments". 2013. In *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht*, herausgegeben von Sascha Bernholt, 174–176.
- Michael, Ann L., Thomas Klee, John D. Brandsford und Steven F. Warren 1993. "The transition from theory to therapy: Test of two instructional methods". 1993. *Applied Cognitive ...* 7:139–153.
- Mietzel, Gerd 2007. *Pädagogische Psychologie des Lernens und Lehrens*. 2007. 567. Hogrefe Verlag.
- Millar, Robin, und Jonathan Osborne 1999. "Beyond 2000: Science education for the future". 1999. *Science And Technology*.
- Munier, Valérie, Hélène Merle und Danie Brehelin 2013. "Teaching Scientific Measurement and Uncertainty in Elementary School". 2013. *International Journal of Science Education* 35 (16): 2752–2783.
- Oelkers, Jürgen, Kurt Reusser, Esther Berner, Uli Halbheer und Stefanie Stolz 2008. *Qualität entwickeln- Standards sichern- mit Differenz umgehen*. 2008. Technischer Bericht. Bonn: Pädagogisches Institut der Universität Zürich.
- Pea, Roy D. 2013. "Putting Knowledge to Use". 2013. In *Technology in Education: Looking Toward 2020*, herausgegeben von Nickerson Raymond S. und Philop P. Zodhiates, 169–212. Routledge.
- Perkins, D.N., und Gavriel Salomon 1989. "Are cognitive skills context-bound?" 1989. *Educational researcher* 18 (1): 16–25.
- PISA-Konsortium Deutschland 2004. "PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland—Ergebnisse des zweiten internationalen Vergleichs". 2004.
- Porter, a. Juni 1989. "A Curriculum out of Balance: The Case of Elementary School Mathematics". 1989. *Educational Researcher* 18, Nr. 5 (Juni): 9–15.
- Renkl, Alexander, Hans Gruber, Heinz Mandl und Ludwig Hinkhofer 1994. "Hilft Wissen bei der Identifikation und Kontrolle eines komplexen ökonomischen Systems?" 1994. *Unterrichtswissenschaft* 22:195–202.

- Reusser, Kurt 2005. "Problemorientiertes Lernen - Tiefenstruktur, Gestaltungsformen, Wirkung". 2005. *Beitraege zur Lehrerbildung* 23 (2): 159–182.
- Rindermann, Heiner April 2006. "Was messen internationale Schulleistungsstudien?" 2006. *Psychologische Rundschau* 57, Nr. 2 (April): 69–86.
- Rizopoulos, D 2006. "ltm: An R package for latent variable modeling and item response theory analyses". 2006. *Journal of Statistical Software* 17 (5): 1–25.
- Rost, Jürgen 2004. *Lehrbuch Testtheorie - Testkonstruktion*. 2004. Verlag Hans Huber.
- Rychen, Dominique Simone, und Laura Hersh Salganik 2001. *Defining and selecting key competencies*. 2001. Herausgegeben von Dominique Simone Rychen und Laura Hersh Salganik. 251. Hogrefe & Huber.
- Schoenfeld, Alan H. 1988. "When good teaching leads to bad results: The disasters of 'well-taught' mathematics courses". 1988. *Educational psychologist* 23 (2): 145–166.
- Schöne, Claudia, Oliver Dickhäuser, Birgit Spinath und Joachim Stiensmeier-Pelster 2002. *SESSKO Skalen zur Erfassung des schulischen Selbstkonzeptes*. 2002. Göttingen: Hoegrefe Verlag.
- Schreiber, Nico 2012. *Diagnostik Experimenteller Kompetenz: Validierung Technologiegestützter Testverfahren Im Rahmen Eines Kompetenzstrukturmodells*. 2012. 273. Logos Verlag Berlin GmbH.
- Shakeshaft, Nicholas G., Maciej Trzaskowski, Andrew McMillan, Kaili Rimfeld, Eva Krapohl, Claire M a Haworth, Philip S Dale und Robert Plomin Januar 2013. "Strong genetic influence on a UK nationwide test of educational achievement at the end of compulsory education at age 16." 2013. *PloS one* 8, Nr. 12 (Januar): e80341.
- Shuell, Thomas J. 1996. "The role of educational psychology in the preparation of teachers". 1996. *Educational Psychologist* 31 (1): 37–41.
- Sichau, David 2015. "Entwicklung eines ExKoNawi hands-on Test zur skalenbasierten Messung". 2015. Forschungsarbeit, PH Zürich.
- Strobl, Carolin 2012. *Das Rasch-Modell*. 2012. 132. Rainer Hampp Verlag.
- Weinert, Franz E. 2001. "Concept of competence: A conceptual clarification." 2001a. In *Defining and selecting key competencies*, herausgegeben von D. S. Rychen und L. H. Salganik. Seattle: Hogrefe & Huber.
- Weinert, Franz E. 2001. *Leistungsmessungen in Schulen*. 2001b, 398. Beltz.
- Whitehead, Alfred North 1929. *The Aims of Education and Other Essays*. 1929, 13–26. New York: Free Press.

- Wiggins, Grant 1993. "Assessment: Authenticity, context, and validity." 1993. *Phi Delta Kappan* 75 (3): 200–208.
- Williams, Susan M. 1992. "Putting case-based instruction into context: Examples from legal and medical education". 1992. *The Journal of the Learning Sciences* 2 (4): 367–427.
- Woodworth, Robert S., und Edward L. Thorndike 1901. "The influence of improvement in one mental function upon the efficiency of other functions. (I)." 1901. *Psychological Review* 8:247–261.

Anhang

1 Urheberschaftsbestätigung

Hiermit erkläre ich, dass die vorliegende Arbeit von mir eigenständig verfasst wurde und keine anderen als die von mir angegebenen Hilfsmittel verwendet wurden. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind mit Angaben der Quellen als solche gekennzeichnet.

Ich nehme zur Kenntnis, dass Arbeiten, die unter Bezug unerlaubter Hilfsmittel verfasst wurden und die fremde Textteile ohne entsprechenden Herkunftsnnachweis enthalten, verfolgt und geahndet werden.

Zürich den 05. Februar 2015



David Sichau

2 Daten und Auswertungen

Für die Auswertungen wurde offene Programmiersprache R¹ verwendet. Diese ist für alle Systeme kostenfrei verfügbar. Aller Code und die Daten dieser Masterarbeit befinden sich auf GitHub und sind frei verfügbar.

Code erhältlich auf:

GitHub

<http://git.io/buGR>

3 Fragebogen

Hier folgen die Fragebögen, welche die Schülerinnen und Schüler ausgefüllt haben. Da sie sich leicht unterschieden, wurden beide Fragebögen angehängt.

1. <http://www.r-project.org/>

3.1 Fragebogen am Anfang

Fragebogen

Code:

A. Allgemeine Fragen

Meine letzte Note im Fach Mathematik war:

Meine letzte Note im Fach Natur und Technik war:

Geschlecht: männlich
 weiblich

B. Fragebogen

Bitte kreuze in der Tabelle jeweils nur eine Spalte an.

	Ich stimme voll zu	Ich stimme eher zu	Ich stimme eher nicht zu	Ich stimme überhaupt nicht zu
1 <i>Ich bin für die Schule sehr begabt.</i> 18(a)				
2 Schulversuche würde ich viel lieber machen, wenn sie nicht so schwer wären. NAT_SEK_2				
3 Bei manchen Schulversuchen weiß ich gleich: „Das verstehe ich nie.“ NAT_SK_4				
4 <i>Ich kann in der Schule viel.</i> 21(a)				
5 Mit den Aufgaben bei Schulversuchen komme ich besser zurecht als viele meiner Mitschüler/innen Nat_SK_6				
6 Ich denke, ich bin für Schulversuche begabter als viele meiner Mitschüler/innen. Nat_SK_7				
7 <i>Ich bin sehr intelligent.</i> 20(a)				
8 Schulversuche liegen mir nicht besonders. Nat_SK_1				
9 Schulversuche fallen mir schwerer als vielen meiner Mitschüler/innen. Nat_SK_3				
10 <i>In der Schule fallen mir viele Aufgaben schwer.</i> 22(a)				
11 Für Schulversuche habe ich einfach keine Begabung. Nat_SK_5				
12 <i>Neues zu lernen fällt mir schwer.</i> 19(a)				

C. Offene Fragen (Schreiben Sie in kurzen Sätzen eine Antwort)

1. Das Ziel von Naturwissenschaftlichen Experimenten ist?

2. Was war dein letzter Schulversuch? Was hast du dort gemacht?

3. Hat dir dieser Schulversuch gefallen. Schreibe bitte eine kurze Begründung.

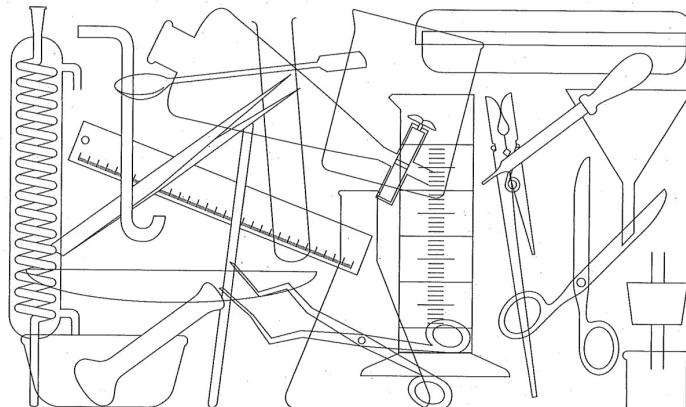
4. Welchen Versuch sollte man unbedingt im Unterricht durchführen und warum?

4. Sollte in der Schule mehr oder weniger experimentiert werden. Bitte schreibe auch eine Begründung.

D. Labor-Quiz

Finde die kurz beschriebenen Laborgeräte im Suchbild oder Buchstabensalat wieder. Markiere sie farbig.

1. Reagenzglas ist ein Glasrohr, das an einer Seite geschlossen und abgerundet ist.
2. In einen Trichter legt man ein Filterpapier ein
3. Eine Thermoskanne besteht aus einem Metallgehäuse und einem Deckel.
4. Eine Federwaage besteht aus einer Feder und einem Hacken und dient zum Kraft messen.
5. Die Pipette besitzt gegenüber der Spitze ein Gummihütchen
6. Eine Waage dient zum Abwiegen.
7. Der Messzylinder hat einen festen Stand und eine aufgedruckte Messskala
8. Ein Taschenrechner hilft beim Rechen.
9. Ein Glas – oben eng und unten weit – heisst Erlenmeyerkolben
10. Ein Bunsenbrenner hat einen Gasanschluss und erzeugt eine heisse Flamme.
11. Eine Schutzbrille trägt man um die Augen zu schützen.
12. Dieser spezielle Löffel heisst Spatel



Q	J	I	P	T	E	L	L	I	R	B	Z	T	U	H	C	S	D
Q	U	E	Q	D	F	Q	L	N	E	S	R	F	E	J	Y	T	B
P	J	G	W	G	I	H	D	J	N	K	I	F	X	S	R	R	D
V	G	A	I	P	K	W	V	X	N	E	X	Z	Z	L	B	G	H
M	U	A	V	Y	Q	N	J	L	E	G	C	F	I	Y	G	I	M
S	Z	W	T	X	N	K	P	T	R	V	J	X	S	G	Q	S	M
V	S	R	J	K	L	E	T	X	B	F	T	T	U	O	D	K	Y
T	H	E	R	M	O	S	K	A	N	N	E	P	I	G	P	T	R
B	G	D	O	T	A	S	C	H	E	N	R	E	C	H	N	E	R
N	R	E	T	C	O	Q	Y	S	S	B	N	K	N	T	I	Q	D
L	P	F	E	I	S	O	K	P	N	Y	E	E	V	Q	J	R	Z
Q	O	G	F	V	I	I	U	U	U	I	C	F	C	E	K	I	M
T	M	W	K	T	Q	S	N	Q	B	E	L	F	M	S	U	Y	S
T	Q	F	U	W	A	A	G	E	S	R	Q	X	S	C	E	C	T

3.2 Fragebogen am Ende

Fragebogen

Code:

A. Allgemeine Fragen

Meine letzte Note im Fach Mathematik war:

Meine letzte Note im Fach Natur und Technik war:

Geschlecht: männlich
 weiblich

B. Fragebogen

Bitte kreuze in der Tabelle jeweils nur eine Spalte an.

	Ich stimme voll zu	Ich stimme eher zu	Ich stimme eher nicht zu	Ich stimme überhaupt nicht zu
1 Ich bin für die Schule sehr begabt.				
2 Schulversuche würde ich viel lieber machen, wenn sie nicht so schwer wären.				
3 Bei manchen Schulversuchen weiß ich gleich: „Das verstehe ich nie.“				
4 Ich kann in der Schule viel.				
5 Mit den Aufgaben bei Schulversuchen komme ich besser zurecht als viele meiner Mitschüler/innen				
6 Ich denke, ich bin für Schulversuche begabter als viele meiner Mitschüler/innen.				
7 Ich bin sehr intelligent.				
8 Schulversuche liegen mir nicht besonders.				
9 Schulversuche fallen mir schwerer als vielen meiner Mitschüler/innen.				
10 In der Schule fallen mir viele Aufgaben schwer.				
11 Für Schulversuche habe ich einfach keine Begabung.				
12 Neues zu lernen fällt mir schwer.				

C. Offene Fragen (Schreiben Sie in kurzen Sätzen eine Antwort)

1. Das Ziel von Naturwissenschaftlichen Experimenten ist?

2. Welcher Versuch hat dir am besten gefallen? Schreibe bitte eine kurze Begründung.

3. Welcher Versuch hat dir am schlechtesten gefallen? Schreibe bitte eine kurze Begründung.

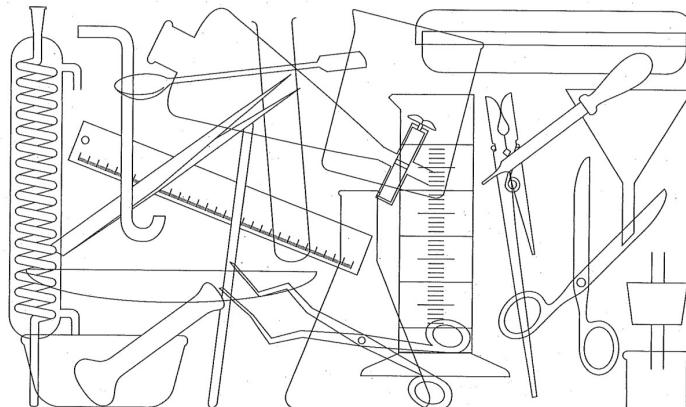
4. Was könnte man an den Versuchen verbessern?

4. Sollte in der Schule mehr oder weniger experimentiert werden. Bitte schreibe auch eine Begründung.

D. Labor-Quiz

Finde die kurz beschriebenen Laborgeräte im Suchbild oder Buchstabensalat wieder. Markiere sie farbig.

1. Reagenzglas ist ein Glasrohr, das an einer Seite geschlossen und abgerundet ist.
2. In einen Trichter legt man ein Filterpapier ein
3. Eine Thermoskanne besteht aus einem Metallgehäuse und einem Deckel.
4. Eine Federwaage besteht aus einer Feder und einem Hacken und dient zum Kraft messen.
5. Die Pipette besitzt gegenüber der Spitze ein Gummihütchen
6. Eine Waage dient zum Abwiegen.
7. Der Messzylinder hat einen festen Stand und eine aufgedruckte Messskala
8. Ein Taschenrechner hilft beim Rechen.
9. Ein Glas – oben eng und unten weit – heisst Erlenmeyerkolben
10. Ein Bunsenbrenner hat einen Gasanschluss und erzeugt eine heisse Flamme.
11. Eine Schutzbrille trägt man um die Augen zu schützen.
12. Dieser spezielle Löffel heisst Spatel



Q	J	I	P	T	E	L	L	I	R	B	Z	T	U	H	C	S	D
Q	U	E	Q	D	F	Q	L	N	E	S	R	F	E	J	Y	T	B
P	J	G	W	G	I	H	D	J	N	K	I	F	X	S	R	R	D
V	G	A	I	P	K	W	V	X	N	E	X	Z	Z	L	B	G	H
M	U	A	V	Y	Q	N	J	L	E	G	C	F	I	Y	G	I	M
S	Z	W	T	X	N	K	P	T	R	V	J	X	S	G	Q	S	M
V	S	R	J	K	L	E	T	X	B	F	T	T	U	O	D	K	Y
T	H	E	R	M	O	S	K	A	N	N	E	P	I	G	P	T	R
B	G	D	O	T	A	S	C	H	E	N	R	E	C	H	N	E	R
N	R	E	T	C	O	Q	Y	S	S	B	N	K	N	T	I	Q	D
L	P	F	E	I	S	O	K	P	N	Y	E	E	V	Q	J	R	Z
Q	O	G	F	V	I	I	U	U	U	I	C	F	C	E	K	I	M
T	M	W	K	T	Q	S	N	Q	B	E	L	F	M	S	U	Y	S
T	Q	F	U	W	A	A	G	E	S	R	Q	X	S	C	E	C	T

4 Aufgabenstellung und Kodierungen

Im folgenden Abschnitt befinden sich die Aufgabenstellungen der drei Tests und die Kodiermanuals.

4.1 Test 201: Aufgabenstellung

Salz lösen

Problem

Bei dieser Aufgabe sollst du herausfinden, wie sich die Temperatur des Wassers verändert, wenn du Pulver hinzugibst.

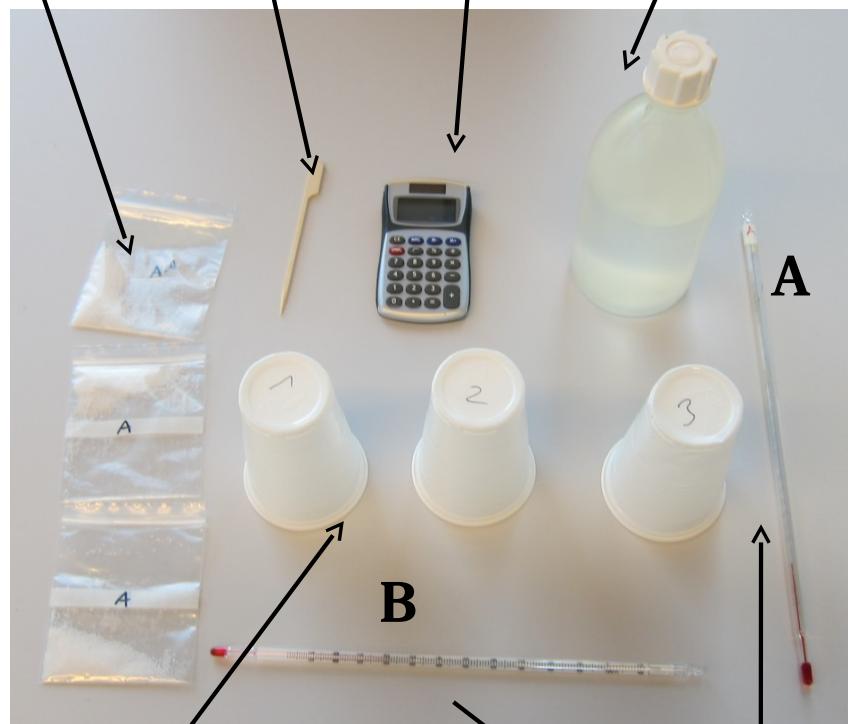
Material

3 Plastiktüten mit je 3 g Pulver

Holzstab

Taschenrechner

Wasserflasche



3 Plastikbecher (1,2 und 3)

2 Thermometer (A und B)

Achtung: Thermometer sind zerbrechlich und kosten viel Geld.

Messung

Aufgabe

Bestimme möglichst genau, wie sich die Temperatur verändert, wenn du 3 g eines Pulvers in 50 ml Wasser löst.

Überlege dir:

- Wie gehst du vor, damit du ein möglichst genaues Resultat erhältst?
- Mit welchem Thermometer misst du am genauesten?
- Wie viele Messungen sind notwendig?

Messprotokoll

- Schreibe zu jeder Messung das Resultat und das benutzte Thermometer (A oder B) auf.

EKN_12_M2_01_i01

Resultat

Die Temperatur des Wassers verändert sich um _____

EKN_12_M2_01_i02

Ist dein Resultat genau? Mache eine Einschätzung.

EKN_12_M2_01_i03

Wie kannst du noch genauer messen?
Begründe deine Antwort.

EKN_12_M2_01_i04

**Lege diese Seiten in dein Mäppchen.
Dann mach weiter mit Seite 5.**

Fragen

- Welches Thermometer hast du für dein Resultat benutzt? Kreuze an.
- Thermometer A
- Thermometer B

EKN_12_M2_01_i05

Kannst du mit beiden Thermometern gleich genau messen?
Begründe deine Antwort.

EKN_12_M2_01_i06

- Wie viel Mal hast du gemessen?

EKN_12_M2_01_i07

- Wie viele Messungen hast du für dein Endresultat gebraucht?

EKN_12_M2_01_i08

- Hast du für dein Endresultat einen Mittelwert berechnet?

Ja, weil ...

Nein, weil ...

EKN_12_M2_01_i09

**Lege das Blatt in dein Mäppchen.
Räume deinen Arbeitsplatz so auf, wie du ihn vorgefunden
hast.**

**Fahre mit dem nächsten Versuch erst nach der Pause
weiter.**

4.2 Test 201: Kodierung

Kodierschema			Temperatur	EKN_12_M2_01
	<i>häufig bei</i>			
QS 1			korrekt und präzise messen	
1.1	i01/i02	1P	Zeigt das Resultat eine richtige Tendenz?	Sinkt die Temperatur bei der Zugabe von Pulver A Erklärung: Das Lösen von Pulver A (Ammoniumchlorid) ist endotherm, d.h. die aufgenommene Hydratationsenthalpie ist grösser als die abgegebene Gitterenthalpie, die Umgebung wird kälter.
1.2	i01/i02	1P	Ist das Resultat vollständig/korrekt (korrekte Einheit)?	<ul style="list-style-type: none"> Wurde richtig vom Thermometer abgelesen und befinden sich, falls angegeben, die Anfangs- und Endtemperaturen bei mind. einer Messung in einem Bereich zwischen 17°C und 28°C (als richtig werden die folgenden Einheiten akzeptiert: °C, °, C) Liegt die entstandene Temperaturdifferenz in einem Bereich zwischen 1-8 °C. Erklärung: Die Raumtemperatur, sowie die Temperatur des Leitungswassers wurden im Vorfeld gemessen und liegen alle in einem Bereich zwischen 19°C und 25°C.
QS 2			Messung darstellen	
2.1	i01	3P	Werden alle Messungen und Messergebnisse vollständig dargestellt?	Je 1P pro Item Vollständigkeit: Bei jeder Messung wird klar 1. welcher Wert (Masszahl) gemessen wurde, 2. welches Messinstrument verwendet wurde 3. wie gemessen wurde (Skizze, muss nur 1mal vorhanden sein)
QS 3			Messinstrument begründen	
3.1	i05	1P	Ist die Wahl des Messinstruments korrekt?	Wahl des Messinstruments mit feinerer Skala: B
3.2	i06	1P	Wird die Wahl des Messinstruments korrekt begründet?	Korrekte Begründung: Feinere Skala
QS 4			Messung wiederholen	
4.1	i02/i07	1P	Entstand das Resultat durch mehrmaliges Messen?	

	i08		
4.2.	i02/i07 i08	1P	Falls ja, wurde mehrmals identisch gemessen?
			Identisch: Pulvermenge und Wassermenge. Die Wahl des Thermometers spielt hier keine Rolle.
4.3.	i02/i07 i08	1P	Falls ja, ist das Resultat durch korrekte Mittelwertbildung entstanden? (Methode)
			akzeptierte „Mittelwertbildung“ : 1. arithmetisches Mittel von mindestens 2 Messungen (identisches Messinstrumente) 2. Median/Extremwertausscheidung: Selektion des Zentralwertes bei einer ungeraden Anzahl (identischer) Messungen 3. Modalwert: Selektion des häufigsten Wertes (bei identischen Messungen)
4.4.	i02/i07 i08	1P	Ist das Resultat ein korrekter Mittelwert? (Ausführung)
			Korrekter Mittelwert wenn die „Mittelwertbildung“ bzw. Messwertselektion korrekt durchgeführt wurde.
QS 5			
5.1	i03/i04	3P	Wie viele Fehlerkategorien werden genannt?
		Je 1P	Messung ist genau und fehlerhaft, weil ... 1. Menge Wasser oder Menge Pulver ist nicht immer konstant, oder 2. Das Messinstrument misst zu ungenau, oder 3. Andere systematische oder zufällige Fehlerquellen werden erwähnt. <i>Fehlerkategorie: Mensch, Natur, Messinstrument (pro genannte Fehlerkategorie 1 Pkt)</i>
5.2	i03/i04	3P	Wie viele richtige Lösungsvorschläge zur Steigerung der Messgenauigkeit werden gemacht?
		Je 1P	Lösungsvorschläge 1. Verbesserungen bei der Messtechnik 2. Messwiederholung und „Mittelwertbildung“ Messwert-Selektion 3. Wahl Messinstrument (Messinstrument mit feinerer Skala)

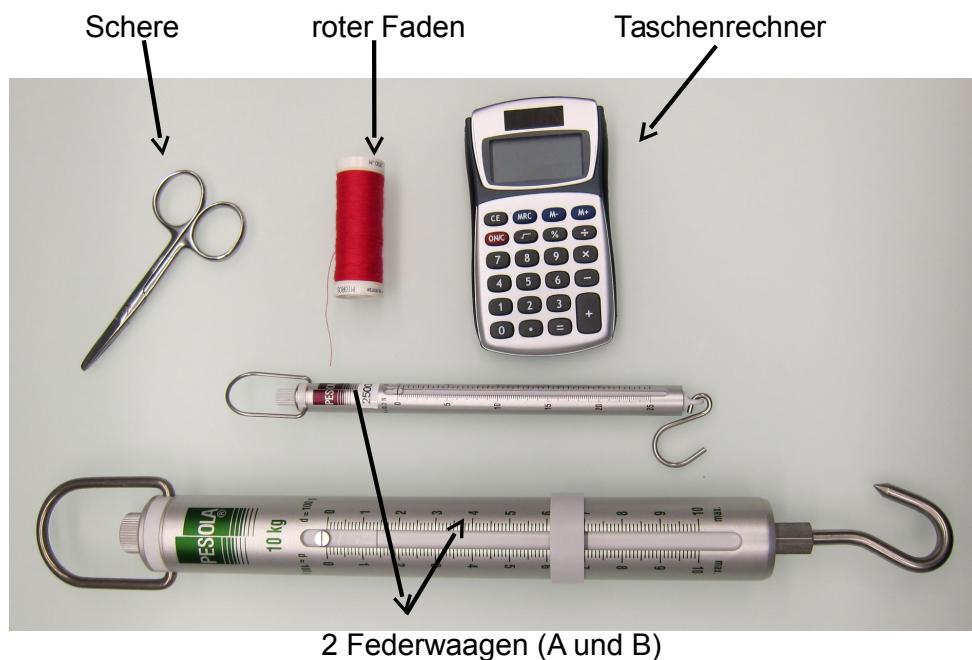
4.3 Test 301: Aufgabenstellung

Faden reissen

Problem

Bei dieser Aufgabe sollst du herausfinden, bei welcher Belastung ein Faden reisst.

Material



Messung

Aufgabe

Bestimme möglichst genau die Belastung, bei welcher der Faden reisst.

Überlege dir:

- Wie gehst du vor, damit du ein möglichst genaues Resultat erhältst?
- Mit welcher Federwaage misst du am genauesten?
- Wie viele Messungen sind notwendig?

Messprotokoll

- Zeichne auf, wie du die Belastung mit der Federwaage gemessen hast.
- Schreibe zu jeder Messung das Resultat und die benutzte Federwaage (A oder B) auf.

EKN_12_M3_01_i01

Resultat

- Der Faden reisst bei einer Belastung von _____.

EKN_12_M3_01_i02

- Ist dein Resultat genau? Mache eine Einschätzung.

EKN_12_M3_01_i03

- Wie kannst du noch genauer messen?
Begründe deine Antwort.

EKN_12_M3_01_i04

**Lege diese Seiten in dein Mäppchen.
Dann mach weiter mit Seite 5.**

Fragen

- Welche Federwaagen hast du für dein Resultat benutzt? Kreuze an.
- Federwaage A
- Federwaage B

EKN_12_M3_01_i05

- Kannst du mit beiden Federwaagen gleich genau messen?
Begründe deine Antwort.

EKN_12_M3_01_i06

- Wie viel Mal hast du gemessen?

EKN_12_M3_01_i07

- Wie viele Messungen hast du für dein Endresultat gebraucht?

EKN_12_M3_01_i08

- Hast du für dein Endresultat einen Mittelwert berechnet?

Ja, weil ...

Nein, weil ...

EKN_12_M3_01_i09

**Lege das Blatt in dein Mäppchen.
Bitte räume deinen Arbeitsplatz so auf, wie du ihn
vorgefunden hast!
Fahre mit dem nächsten Versuch erst nach der Pause
weiter.**

4.4 Test 301: Kodierung

Kodierschema			Faden	EKN_12_M3_01
	<i>häufig bei</i>			
QS 1			korrekt und präzise messen	
1.1	i01/10 2	1P	Ist das Resultat präzise (Masszahl innerhalb Toleranzbreite)? liegt das Resultat im Bereich 700-1400	
1.2	i01/10 2	1P	Ist das Resultat vollständig/korrekt (korrekte Einheit)? Präzision und Korrektheit der Lösung: Belastungsgrenze = 700g-1400g (2-Schlaufen-Ansatz) = 1400g-2800g (1-Schlaufen-Ansatz) Falls keine Skizze vorhanden ist, muss mindestens 1 Wert innerhalb der gesamten Toleranzbreite liegen. Werte aus dem Messprotokoll werden als Resultate interpretiert. Erklärung: Je nach Messvariante, wird die doppelte Belastungsgrenze gemessen (1-Schlaufen-Ansatz)!	
QS 2			Messung darstellen	
2.1	i01	3P	Werden alle Messungen und Messergebnisse vollständig dargestellt? Je 1P Die Antworten müssen im Messprotokoll ersichtlich sein. Vollständigkeit: Bei jeder Messung wird klar 1. welcher Wert (Masszahl, Einheit) gemessen wurde, 2. welches Messinstrument verwendet wurde, 3. wie gemessen wurde (Skizze, muss nur 1mal vorhanden sein)	
QS 3			Messinstrument begründen	
3.1	i05	1P	Ist die Wahl des Messinstruments korrekt? Wahl des Messinstruments mit feinerer Skala: A - „A ist genauer“ gilt nicht als Begründung -> mit 0 codiert	
3.2	i06	1P	Wird die Wahl des Messinstruments korrekt begründet? Korrekte Begründung: Feinere Skala	
QS 4			Messung wiederholen	
4.1	i02/10	1P	Entstand das Resultat durch mehrmaliges Messen?	

	7 i08		
4.2.	i02/i0 7 i08	1P	Falls ja, wurde mehrmals identisch gemessen?
			Identisch: gleiche Federwaage
4.3.	i02/i0 7 i08	1P	Falls ja, ist das Resultat durch korrekte Mittelwertbildung entstanden? (Methode)
			akzeptierte „Mittelwertbildung“ : 1. arithmetisches Mittel von mindestens 2 Messungen (identisches Messinstrumente) 2. Median/Extremwertausscheidung: Selektion des Zentralwertes bei einer ungeraden Anzahl (identischer) Messungen 3. Modalwert: Selektion des häufigsten Wertes (bei identischen Messungen)
4.4.	i02/i0 7 i08	1P	Ist das Resultat ein korrekter Mittelwert? (Ausführung)
			Korrekt er Mittelwert wenn die „Mittelwertbildung“ bzw. Messwertselektion korrekt durchgeführt wurde.
QS 5			
5.1	i03/i0 4	3P	Wie viele Fehlerkategorien werden genannt?
	Je 1P		Messung ist genau und fehlerhaft, weil ... 1. die Belastung an der Skala der Federwaage sehr rasch abgelesen werden muss (Beobachtungsschwierigkeiten) -> Mensch 2. der Faden nicht homogen ist (materialimmanente Variation) -> Natur 3. technische Schwierigkeit, Belastung kontinuierlich und langsam zu erhöhen (messtechnische Schwierigkeiten) -> Mensch 4. Reibung in der Federwaage (Mängel des Messinstruments) -> Messinstrument 5. ... <i>Fehlerkategorie: Mensch, Natur, Messinstrument (pro genannte Fehlerkategorie 1 Pkt)</i>
5.2	i03/i0	3P	Wie viele richtige Lösungsvorschläge zur Steigerung der Messgenauigkeit werden

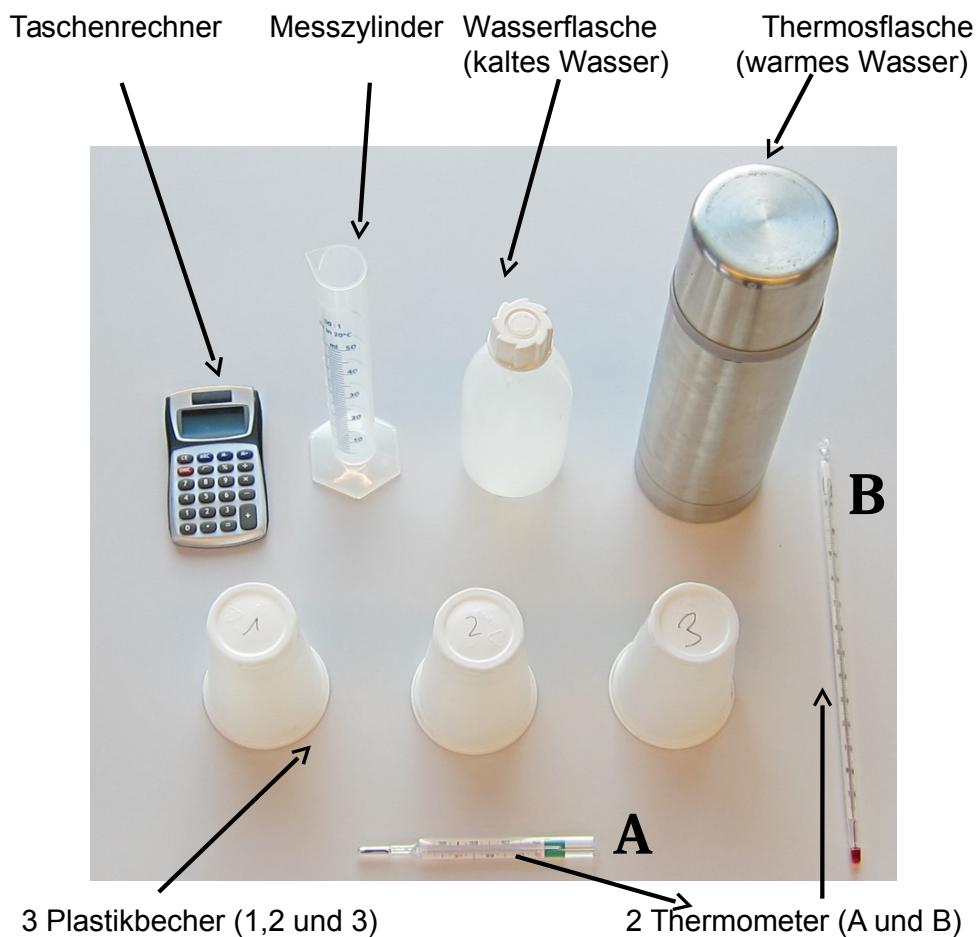
	4	gemacht?
	Je 1P	<u>Lösungsvorschläge</u> 1. Verbesserungen bei der Messtechnik (Mehr-Schlaufen-Ansatz, Technik, Kamera...) 2. Messwiederholung und „Mittelwertbildung“ Messwert-Selektion 3. Wahl Messinstrument (Messinstrument mit feinerer Skala, digitaler Kraftmesser)

4.5 Test 305: Aufgabenstellung

Wasser mischen

Problem

Bei dieser Aufgabe sollst du herausfinden, wie sich die Temperatur von warmen Wassers verändert, wenn du es mit Wasser mischt.



Achtung: Thermometer sind zerbrechlich und kosten viel Geld.

Messung

Aufgabe

Mische 50mL warmes Wasser mit 50mL Wasser aus der Flasche.
Bestimme die Endtemperatur.

Überlege dir:

- Wie gehst du vor, damit du ein möglichst genaues Resultat erhältst?
- Mit welchem Thermometer misst du am genauesten?
- Wie viele Messungen sind notwendig?

Messprotokoll

- Schreibe zu jeder Messung das Resultat und das benutzte Thermometer (A oder B) auf.

EKN_14_M3_05_i01

Resultat

- Die Temperatur des Wassers verändert sich um _____

EKN_14_M3_05_i02

- Ist dein Resultat genau? Mache eine Einschätzung.

EKN_14_M3_05_i03

- Wie könntest du noch genauer messen?
Begründe deine Antwort.

EKN_14_M3_05_i04

**Lege diese Seiten in dein Mäppchen.
Dann mach weiter mit Seite 5.**

Fragen

➤ Welches Thermometer hast du für dein Resultat benutzt? Kreuze an.

- Thermometer A
- Thermometer B

EKN_14_M3_05_i05

- Kannst du mit beiden Thermometern gleich genau messen?
Begründe deine Antwort.

EKN_14_M3_05_i06

- Wie viel Mal hast du gemessen?

EKN_14_M3_05_i07

- Wie viele Messungen hast du für dein Endresultat gebraucht?

EKN_14_M3_05_i08

- Hast du für dein Endresultat einen Mittelwert berechnet?

Ja, weil ...

Nein, weil ...

EKN_14_M3_05_i09

**Lege das Blatt in dein Mäppchen.
Räume deinen Arbeitsplatz so auf, wie du ihn vorgefunden
hast.**

**Fahre mit dem nächsten Versuch erst nach der Pause
weiter.**

4.6 Test 305: Kodierung

Kodierschema		Temperatur	EKN_14_M3_05
	<i>häufig bei</i>		
QS 1		korrekt und präzise messen	
1.1	i01/i02	1P Zeigt das Resultat eine richtige Tendenz?	Liegt die Endtemperatur zwischen der Temperatur des warmen und kalten Wassers (ca. In der Mitte)
1.2	i01/i02	1P Ist das Resultat vollständig/korrekt (korrekte Einheit)?	<ul style="list-style-type: none"> Wurde richtig vom Thermometer abgelesen und befinden sich, falls angegeben, die Anfangs- und Endtemperaturen bei mind. einer Messung in einem Bereich zwischen 25°C und 45°C (als richtig werden die folgenden Einheiten akzeptiert: °C, °, C und Celsius) Liegt die entstandene Temperaturdifferenz in einem Bereich zwischen 8-15 °C. <p><i>Erklärung: Die Temperatur des Warmwassers beträgt ca. 45-50 °C. Das kalte Wasser hat ungefähr 18-25 °C.</i></p>
QS 2		Messung darstellen	
2.1	i01	3P Werden alle Messungen und Messergebnisse vollständig dargestellt?	<p>Je Vollständigkeit: 1P Bei jeder Messung wird klar (je ite m) 1. welcher Wert (Masszahl) gemessen wurde, 2. welches Messinstrument verwendet wurde 3. wie gemessen wurde (Skizze, muss nur 1mal vorhanden sein)</p>
QS 3		Messinstrument begründen	
3.1	i05	1P Ist die Wahl des Messinstrumentes korrekt?	Wahl des Messinstruments mit korrekter Skala: B
3.2	i06	1P Wird die Wahl des Messinstrumentes korrekt begründet?	Korrekte Begründung: Skala liegt im korrekten Bereich oder die Temperatur sinkt bei Thermometer A nicht. Thermometer ist defekt wird auch als korrekt bewertet.
QS 4		Messung wiederholen	
4.1	i02/i07 i08	1P Entstand das Resultat durch mehrmaliges Messen?	

4.2.	i02/i07 i08	1P	Falls ja, wurde mehrmals identisch gemessen?
			Identisch: Wassermenge die gemischt wird. Die Wahl des Thermometers spielt hier keine Rolle.
4.3.	i02/i07 i08	1P	Falls ja, ist das Resultat durch korrekte Mittelwertbildung entstanden? (Methode)
			akzeptierte „Mittelwertbildung“ : 1. arithmetisches Mittel von mindestens 2 Messungen (identisches Messinstrumente) 2. Median/Extremwertausscheidung: Selektion des Zentralwertes bei einer ungeraden Anzahl (identischer) Messungen 3. Modalwert: Selektion des häufigsten Wertes (bei identischen Messungen)
4.4.	i02/i07 i08	1P	Ist das Resultat ein korrekter Mittelwert? (Ausführung)
			Korrechter Mittelwert wenn die „Mittelwertbildung“ bzw. Messwertselektion korrekt durchgeführt wurde.
QS 5			
5.1	i03/i04	3P	Wie viele Fehlerkategorien werden genannt?
		Je 1P	Messung ist genau und fehlerhaft, weil ... 1. Menge Wasser ist nicht immer konstant oder der Zeitpunkt der Messung ist verschieden, oder 2. Das Messinstrument misst zu ungenau, oder 3. Andere systematische oder zufällige Fehlerquellen werden erwähnt. <i>Fehlerkategorie: Mensch, Natur, Messinstrument (pro genannte Fehlerkategorie 1 Pkt)</i>
5.2	i03/i04	3P	Wie viele richtige Lösungsvorschläge zur Steigerung der Messgenauigkeit werden gemacht?
		Je 1P	<u>Lösungsvorschläge</u> 1. Verbesserungen bei der Messtechnik 2. Messwiederholung und „Mittelwertbildung“ Messwert-Selektion 3. Wahl Messinstrument (Messinstrument mit feinerer Skala)

5 Einverständnis Erklärung für Video Aufnahme



Schüler und Schülerinnen der Klasse von xy

Pädagogische Hochschule Zürich
David Sichau
c/o Pitt Hild
Zentrum für Didaktik der Naturwissenschaften
Lagerstrasse 2
CH-8090 Zürich
T +41 (0)44 63 88 12
E-Mail: sichau@inf.ethz.ch

Zürich, 03.11.2014

Erlaubnis für Videoaufnahmen

Sehr geehrte Eltern

Experimentieren kann am besten mit Hilfe von Videoaufnahmen veranschaulicht werden. Zu diesem Zweck möchten wir im Rahmen einer Doppellection Ihre Tochter bzw. Ihren Sohn filmen. Aus Gründen des Persönlichkeits- und Datenschutzes benötigen wir dafür Ihre Zustimmung.

Die PH Zürich verpflichtet sich, das Videomaterial ausschliesslich in Zusammenhang mit Master- und Forschungsarbeiten zu verwenden. Das Videomaterial wird nicht der Öffentlichkeit zugänglich gemacht.

Wir bitten Sie deshalb, uns durch Ihre Unterschrift zu bestätigen, dass man ihre Tochter bzw. ihren Sohn während der Doppellection filmen darf.

Merci vielmals für Ihre Mitarbeit und Ihr Verständnis.

David Sichau
Masterstudent Fachdidaktik Naturwissenschaften

Doppellection, am 6.11.2014,

Ja, ich bin mit den Videoaufnahmen einverstanden:

Name, Vorname

Unterschrift

--	--