

Forschungsarbeit an der Pädagogischen Hochschule
Zürich

Masterstudiengang Fachdidaktik
Naturwissenschaften

vorgelegt von:
David-Matthias Sichau

eingereicht bei:
Pitt Hild

05. Februar 2015, Zürich

Inhaltsverzeichnis

1. Einleitung	4
2. Theoretischer Rahmen	4
2.1. Kompetenz des skalenbasierten Messens	4
2.2. ExKoNawi	5
3. Testentwicklung	5
4. Methode	6
5. Anforderungen	6
6. Umsetzung	6
7. Untersuchung	8
7.1. Vorbereitung	8
7.2. Durchführung	8
7.3. Nachbereitung	9
8. Ergebnisse	9
9. Kodierung	9
9.1. Items	10
9.2. Qualitätsstandards	11
9.2.1. Erreichte Qualitätsstandards	12
9.3. Niveau	12
9.3.1. Unbedingtes Niveau	13
9.3.2. Bedingtes Niveau	13
10. Fragebogen	13
11. Unterschiede zwischen den Klassen	14
12. Korrelation der Niveaus des skalenbasierten Messens	15
13. Rasch-Analyse	18
13.1. Parameterschätzung	18
13.2. Modellkontrolle des Rasch-Modells	20
13.3. Unterschied in den Schwierigkeiten der Qualitätsstandards	22
13.4. Unterschiede in den latenten Personenfähigkeiten	25
13.5. Zusammenhang zwischen Rasch-Modell und Fragebogen	28

14. Videoanalyse	30
14.1. Qualitätsstandards	30
14.1.1. Korrekt und präzise messen	30
14.1.2. Messung wiederholen	30
14.2. Korrelation zwischen Video-Merkmalen und Qualitätsstufen	30
14.3. Messzeitpunkte und Messdauer	31
15. Diskussion	33
16. Kodierung	33
16.1. Items	33
16.2. Qualitätsstandards	35
16.3. Niveaus	35
17. Fragebogen	35
18. Unterschiede zwischen den Klassen	36
19. Ist das Abschneiden in den Tests unterschiedlich	36
20. Rasch-Analyse	37
20.1. Parameterschätzung	37
20.2. Modellkontrolle	38
20.3. Unterschied in den Schwierigkeiten der Qualitätsstandards	39
20.4. Unterschied in den latenten Personenfähigkeiten	39
20.5. Zusammenhang zwischen Rasch-Modell und Fragebogen	40
21. Videoanalyse	40
22. Zusammenfassung	41
23. Ausblick	41
24. Datengrundlage	41
25. Videoanalyse	42
26. Methoden	42
Literaturverzeichnis	43
A. Urheberschaftsbestätigung	45
B. Daten und Auswertungen	45

C. Fragebogen	46
C.1. Fragebogen am Anfang	47
C.2. Fragebogen am Ende	51
D. Aufgabenstellung und Kodierungen	55
D.1. Test 201: Aufgabenstellung	56
D.2. Test 201: Kodierung	62
D.3. Test 301: Aufgabenstellung	64
D.4. Test 301: Kodierung	70
D.5. Test 305: Aufgabenstellung	74
D.6. Test 305: Kodierung	80
E. Einverständnis Erklärung für Video Aufnahme	82

1. Einleitung

Im Rahmen einer Masterarbeit an der PH Zürich (Sichau 2015b), bei der Untersucht wurde, inwiefern die Kompetenz des skalenbasierenden Messens vom Kontext abhängig ist, musste ein neuer hands-on Experimentiertest entwickelt werden. Der hands-on Experimentiertest sollte auf bereits existierenden hands-on Testaufgaben des Projekt ExKoNawi der PH Zürich (Metzger et al. 2013) basieren. Die Grundlage für die Entwicklung des Tests waren dabei die existierenden Testaufgaben zur Kompetenz des skalenbasierenden Messens (Gut et al. 2014; Metzger et al. 2013).

Im Rahmen dieser Forschungsarbeit wird die Entwicklung dieses Testes dokumentiert. Zusätzlich soll dieser Test auch mit den existierenden Tests verglichen werden und auf Unterschiede untersucht werden.

2. Theoretischer Rahmen

2.1. Kompetenz des skalenbasierten Messens

Die theoretische Grundlage des Testes stellt die Kompetenz des skalenbasierten Messens dar. Die Definition dieser Kompetenz basiert auf der Arbeit von Munier, Merle und Brehelin (2013). In dieser Kompetenz geht es darum „quantitative Grössen mit gegebenen Messinstrument genau [zu] messen“(Gut et al. 2014). Diese Kompetenz kann in drei Teilbereiche aufgegliedert werden. Zum einen ist die Wahl des Messinstrumentes, welches für eine Messung geeignet ist, eine wichtige Teilkompetenz. Nur mit einem geeigneten Messinstrument kann die Messung durchgeführt werden. Zusätzlich sind bei genauen Messungen die Wiederholung der Messung elementar. Nur mit Messwiederholungen lässt sich die Genauigkeit des Resultates abschätzen. Da-

her ist die Messwiederholung ein weiterer wichtiger Teilapekt der Kompetenz des Messens. Der dritte Teilapekt ist die korrekte Messung. Wenn das Messinstrument nicht korrekt verwendet wird, kann kein korrektes Messergebnis erhoben werden. Diese Teilkompetenzen ergeben zusammen die Kompetenz des skalenbasierten Messens (Munier, Merle und Brehelin 2013; Gut et al. 2014).

Die Kompetenz des skalenbasierten Messens beruht auf dem Kompetenzmodell von Gott und Duggan (1996), bei welchem Kompetenzen aus transferfähigem Strategiewissen und kontextspezifischem Fachwissen bestehen Gott und Duggan (2002).

2.2. ExKoNawi

Beim Projekt ExKoNawi der PH-Zürich (Metzger et al. 2013) geht es darum hands-on Experimentiertests zu entwickeln, welche verschiedene Kompetenzen von Schülern und Schülerinnen auf der Sekundarstufe I in Schweizer Schulen zu messen. Eine der gemessenen Kompetenzen ist die Kompetenz des skalenbasierten Messens. Für diese Kompetenz gibt es bereits existierende Tests (Metzger et al. 2013; Gut et al. 2014; Hild, Metzger und Parchmann 2014b, 2014a). Diese Tests sind in verschiedenen Kontexten angesiedelt, messen aber alle die Kompetenz des skalenbasierten Messens. Im Test 201 wird die Temperaturveränderung von Wasser bei auflösen eines Salzes gemessen. Im Test 301 wird gemessen, bei welcher Kraft ein Faden reisst.

3. Testentwicklung

Wenn wir nun den fachlichen Kontext untersuchen der Test analysieren, so ist der Test 201 im Bereich Chemie angesiedelt der Test 301 jedoch in der Physik. Beim inhaltlichen Kontext geht es um eine Temperaturmessung und einmal um eine Kraftmessung. Um nun messen zu können ob es einen Unterschied bei der Kompetenz des skalenbasierten Messens zum einen bei fachlichen und zum anderen beim inhaltlichen Kontext gibt, musste ein weiterer Test entwickelt werden. Dieser Test mit der Nummer 305 musste daher fachlich in der Physik angesiedelt sein. Inhaltlich sollte dieser Test eine Temperaturmessung enthalten.

Da dieser Test auf der Sek I in der Klasse 7 durchgeführt werden sollte, musste das Themengebiet und die Aufgabe leicht verständlich sein. Zusätzlich wird der Test extern durchgeführt, daher sollte der Materialaufwand gering sein und eine schnelle Einrichtung von Arbeitsplätzen möglich sein. Daher wurde entschieden als Aufgabe die möglichst genaue Bestimmung der Mischtemperatur von kaltem und warmen Wasser zu stellen. Der gesamte Test wurde sehr stark an der Aufgabenstellung von Test 201 angelehnt. Auch das Kodierschemata wurde vom existierenden abgeleitet und angepasst. Die Tests finden sie im Anhang unter Abschnitt D.

4. Methode

4.1. Umsetzung

Die Tests wurden im Rahmen

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, bei welcher eine zuerst den Fragebogen ausfüllte und die andere Gruppe den Fragebogen am Ende ausfüllte. Zusätzlich wurde noch der zweite und dritte Test in jeder Gruppe vertauscht um zu untersuchen ob Müdigkeit oder die Wiederholungen Einfluss auf die Test-Ergebnisse haben. Die Tabelle 1 gibt eine Übersicht über die Gruppeneinteilung der Schülerinnen und Schüler innerhalb einer Klasse an.

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 1: Aufteilung der Gruppen, innerhalb einer Klasse

Die Namen der Gruppen aus Tabelle 1 wurden auch für die Kodierung der Tests verwendet, sodass jeder Test einer Gruppe zuordenbar ist.

Die vier Klassen waren alle von derselben Schulstufe (7. Schuljahr) jedoch in verschiedenen Gemeinden. Die Klassen in Glattbrugg hatten beide dieselbe Lehrperson, die anderen Klassen hatten unterschiedliche Lehrpersonen. Ein Überblick über die wichtigsten Daten zu den einzelnen Klassen befindet sich in Tabelle 2.

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 haben, wurden in ein Zimmer gesetzt werden, wo keine Videoaufnahme stattgefunden hat. Die Erlaubnis zur Videoaufnah-

	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 2: Aufteilung der Gruppen, innerhalb einer Klasse

me wurde im Vorhinein zur Durchführung von den Klassenpersonen organisiert und eingesammelt.

5. Untersuchung

5.1. Vorbereitung

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

Zusätzlich wurden die Auswertungsbögen in der richtigen Reihenfolge und bereits mit einer Kodierung versehen, in einem Schnellhefter bereitgestellt. Ein für die Durchführung vorbereiteter Klassenraum ist im Bild 1 ersichtlich.

Im Bild 1 sieht man auch gut, wie die Kamera für die Videoauswertung aufgestellt wurde. Die Videoaufnahme wurde bevor die Schülerinnen und Schüler den Klassenraum betreten haben gestartet, um die Ablenkung durch die Kamera möglichst gering zu halten.

5.2. Durchführung

Nachdem die Schülerinnen und Schüler in die beiden Räume aufgeteilt wurden, wurden sie von den Versuchsleitern jeweils begrüßt. Die Begrüssung war stichwortartig



Abbildung 1: Klassenzimmer für die Durchführung des ersten Durchganges vorbereitet.

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

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

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

5.3. 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 das 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 Test Situation haben. Insgesamt ist Material zu 8 Schülerinnen und Schüler verwertbar, da die andern zu weit entfernt sind und daher ihre Aktionen nicht beobachtbar waren.

6. Ergebnisse

7. Kodierung

Wie bereits beschrieben wurde die Erstkodierung von David Sichau durchgeführt. Es wurde eine Zweitkodierung 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 D im Anhang).

7.1. 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 3 sind alle Ergebnisse zusammengefasst.

Code erhältlich auf:

GitHub

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

7.2. Qualitätsstandards

Aus den elf Items wurden fünf Qualitätsstandards entwickelt (Hild, Metzger und Parchmann 2014b). 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.

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 3: 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 1

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

Qualitätsstandard 2

Bei Qualitätsstandard 2 wird die Dokumentation der Messung bewertet . Dieser Qualitätsstandard wird nur erreicht wenn Item 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 Messinstrumentes bewertet. Dieser Standard wird nur erreicht wenn Item 3.1 (Ist das Korrekte Messinstrument gewählt worden?) und Item 3.2 (Wird die Wahl des Messinstrumentes korrekt begründet?) zusammen 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.

7.2.1. Erreichte Qualitätsstandards

In Tabelle 4 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: Zusammenfassung der erreichten Qualitätsstandards, wobei $p_{Q1} - p_{Q5}$ den unbedingten Qualitätsstandards entsprechen. Die bedingten Qualitätsstandards werden mit $p_{QS1} - p_{QS5}$ bezeichnet.

7.3. 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 5 gegeben.

Code erhältlich auf:

GitHub

<http://git.io/bjn9qg>

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 5: Prozedural erreichte Niveaus aller Schülerinnen und Schüler. Beim bedingten Niveau ist es jeweils erforderlich, dass alle vorhergehenden Qualitätsstandards erreicht worden sind.

7.3.1. Unbedingtes Niveau

Dieses Niveau ist der Summenscore der einzelnen unbedingten Qualitätsstandards.

7.3.2. Bedingtes Niveau

Dieses Niveau ist der Summenscore der bedingten Qualitätsstandards.

8. Fragebogen

Im standardisierten Teil des Fragebogens wurden Fragen zum absoluten Selbstkonzept nach SESSKO gestellt (Schöne et al. 2002). Die verwendeten Fragen sind in Tabelle 6 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 6: Fragen von SESSKO zur Skala „Schulisches Selbstkonzept - absolut“ (Schöne et al. 2002). α_d bezeichnete 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 7 aufgezeigt sind.

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 7: Fragen zum Selbstkonzept bei Schulversuchen abgewandelt nach Dierks, Höffler und Parchmann (2014). α_d bezeichnete das standardisierte Cronbach Alpha wenn dieses Item weggelassen würde.

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. 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.

Code erhältlich auf:

GitHub

<http://git.io/WyJH6Q>

9. Unterschiede zwischen den Klassen

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

Besteht kein Unterschied in den Qualitätsstandards zwischen den einzelnen Klassen?

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

Aufgrund der geringen Anzahl an Beobachtungen für einzelne Qualitätsstandards wurde der 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 8: 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 8. 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>

10. Korrelation der Niveaus des skalenbasierten Messens

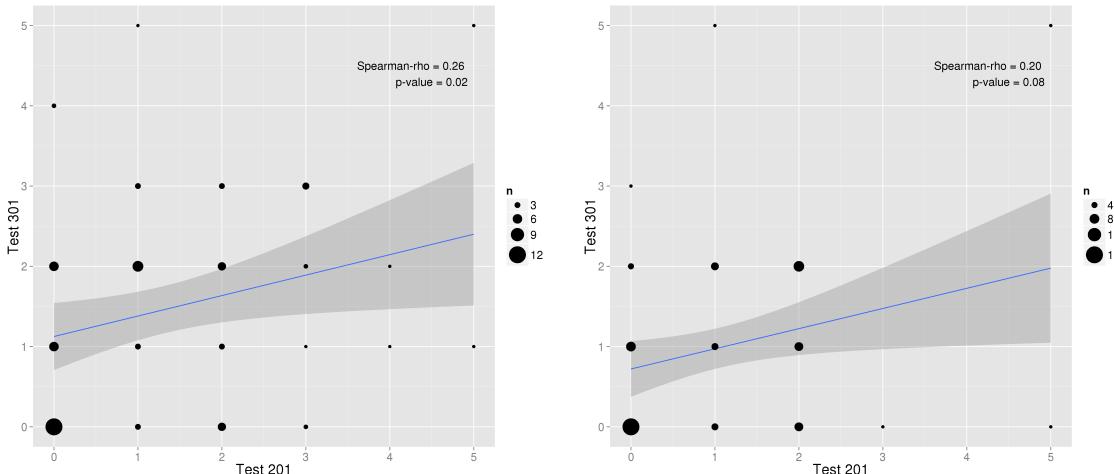
In einem nächsten Schritt wurde untersucht inwiefern die bedingten und unbedingten Niveau-Stufen zwischen den einzelnen Tests korrelieren. Dazu wurde als Rangkorrelationskoeffizient Spearmans ρ berechnet. Der Vorteil dieser Methode ist, dass keine Annahmen über die zugrundliegenden Daten gemacht werden 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 Korrelation signifikant ist, wurde die Korrelation zusätzlich auf Signifikanz getestet. Wichtig bei dieser Analyse ist, dass die Korrelation keine Aussage über die Kausalität zulässt.

Die Ergebnisse wurden grafisch als Streudiagramme dargestellt (siehe Darstellung 2). 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 9 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 9: Spearmans ρ und p-Werte für die Korrelation zwischen den unbedingten Niveaus (uLev) und den bedingten Niveaus (kLev) zwischen den einzelnen Tests.



(a) Korrelation der unbedingten Niveau-Stufen zwischen Test 301 und 201. (b) Korrelation der bedingten Niveau-Stufen zwischen Test 301 und 201.

Code erhältlich auf:

GitHub

<http://git.io/FnbD>

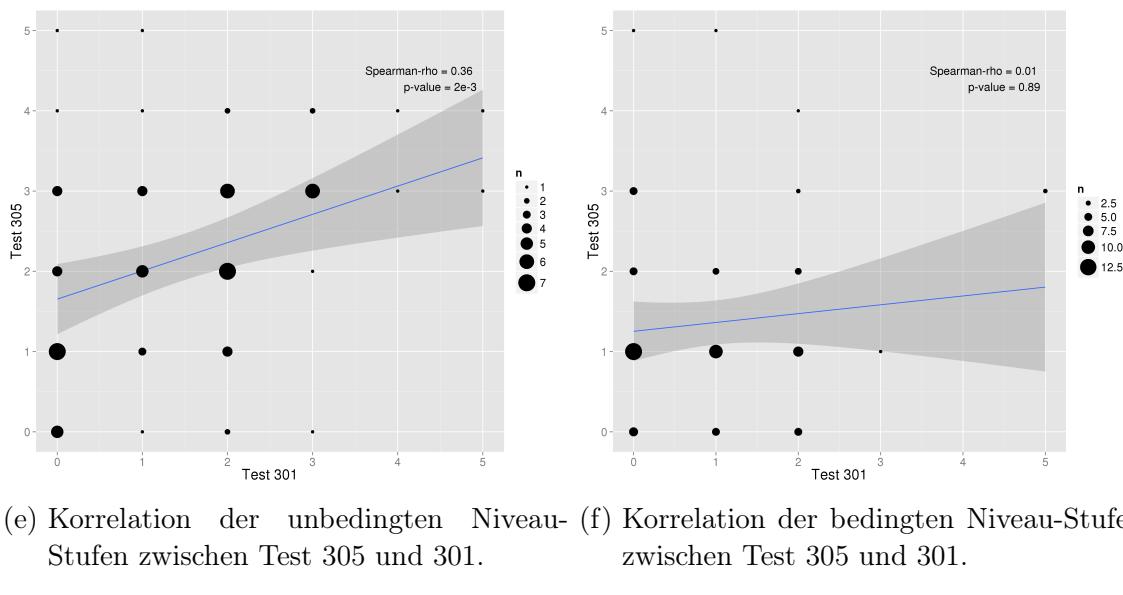
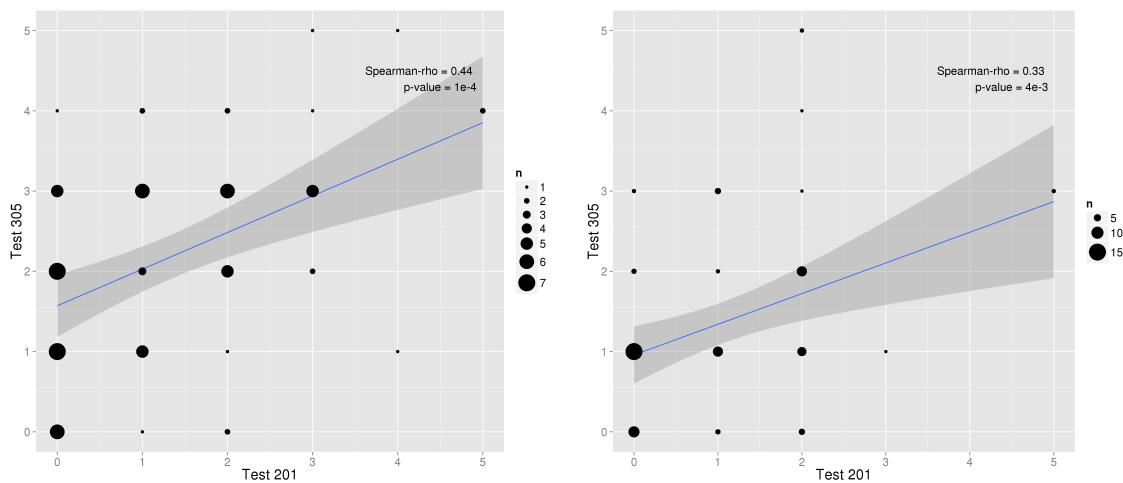


Abbildung 2: Korrelation zwischen den Niveau-Stufen der einzelnen Tests. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die (blaue) Gerade ist die lineare Regression der zugrundeliegenden Daten, der dunkel graue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

11. Rasch-Analyse

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

Es wurde folgendes Rasch-Modell verwendet.

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

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.

11.1. Parameterschätzung

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

Bei der bedingten Maximum-Likelihood-Schätzung wird ein zweistufiges Vorgehen gewählt. Zuerst werden die 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 3 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.

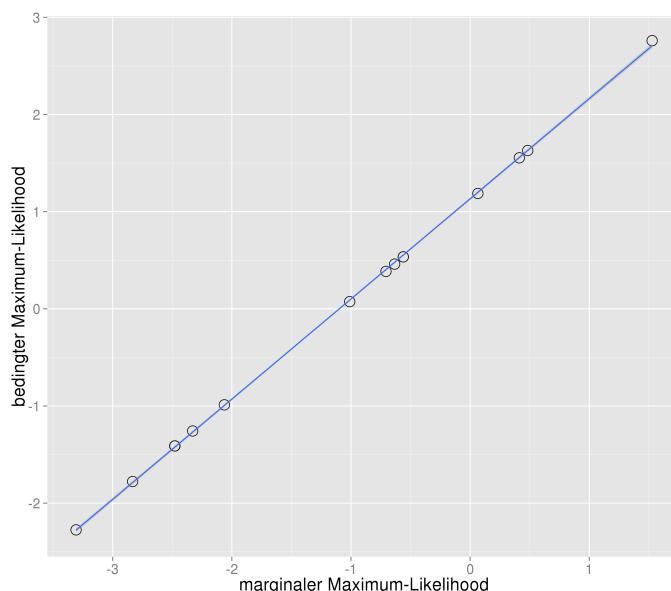


Abbildung 3: Vergleich des Rasch-Modells mit der bedingten Maximum-Likelihood-Schätzung und der marginalen Maximum-Likelihood-Schätzung. Da alle Punkte auf einer Geraden liegen, gibt es keinen Unterschied zwischen den unterschiedlichen Schätzmethoden für Schwierigkeitsgrad der Qualitätsstandards in dem 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>

11.2. 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 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 10 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 10: 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>

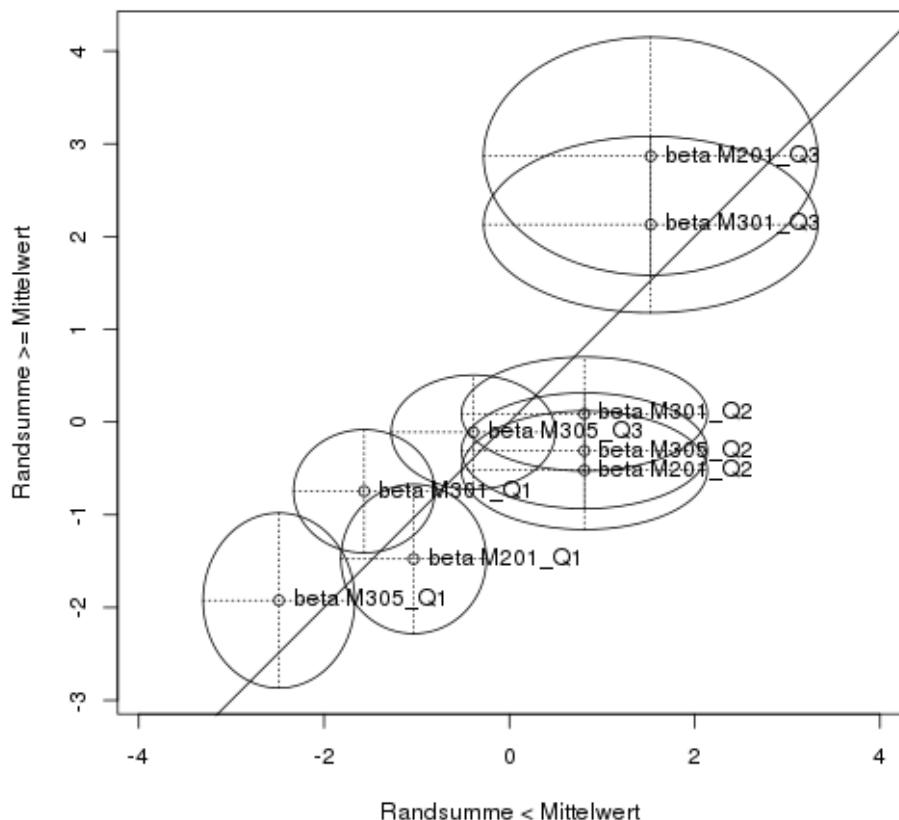
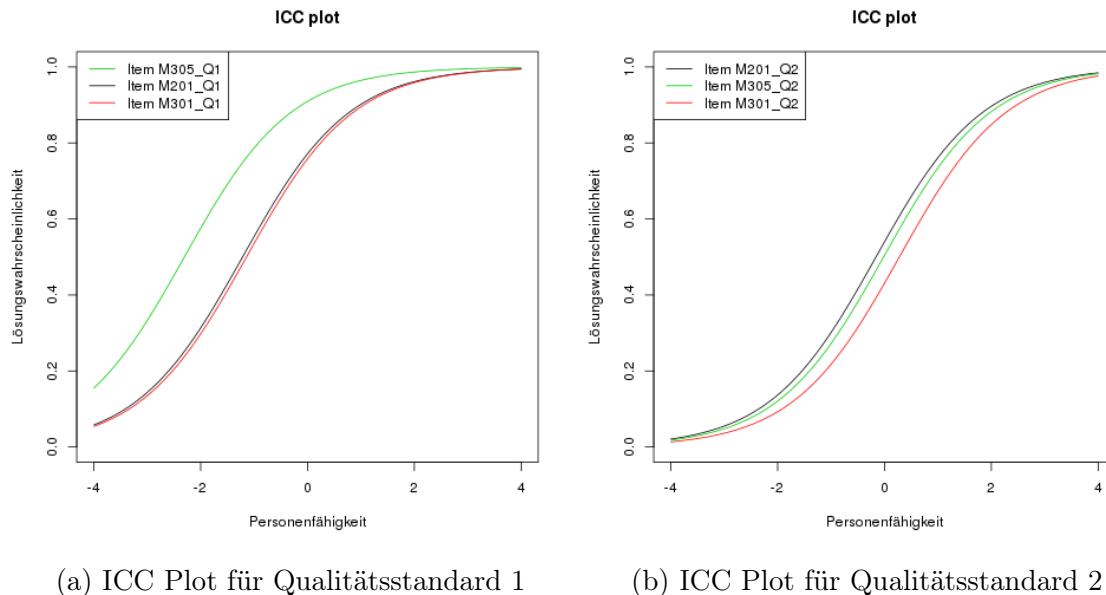


Abbildung 4: Modellkontrolle des Rasch-Modells: kein Qualitätsstandard hat eine signifikante Abweichung von der Diagonalen, daher gibt es keine signifikanten Unterschiede für Personen mit niedrigen und hohen Randsummen in den Qualitätsstandards.

11.3. 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 11 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 11: 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 7 und in Tabelle 12 angegeben. Wichtig dabei ist, dass der Stichprobenumfang mit 3 sehr gering ist.

Code erhältlich auf:

GitHub

<http://git.io/FVZt>

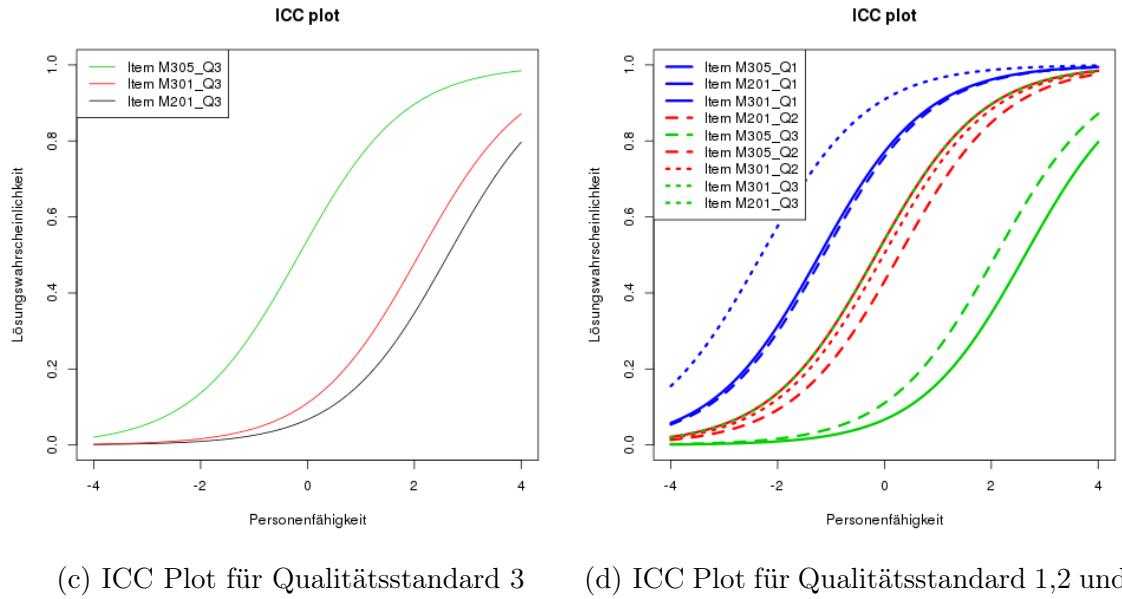


Abbildung 5: Aufgabencharakteristische Kurven für die Qualitätsstandards 1, 2 und 3 für alle drei Tests.

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

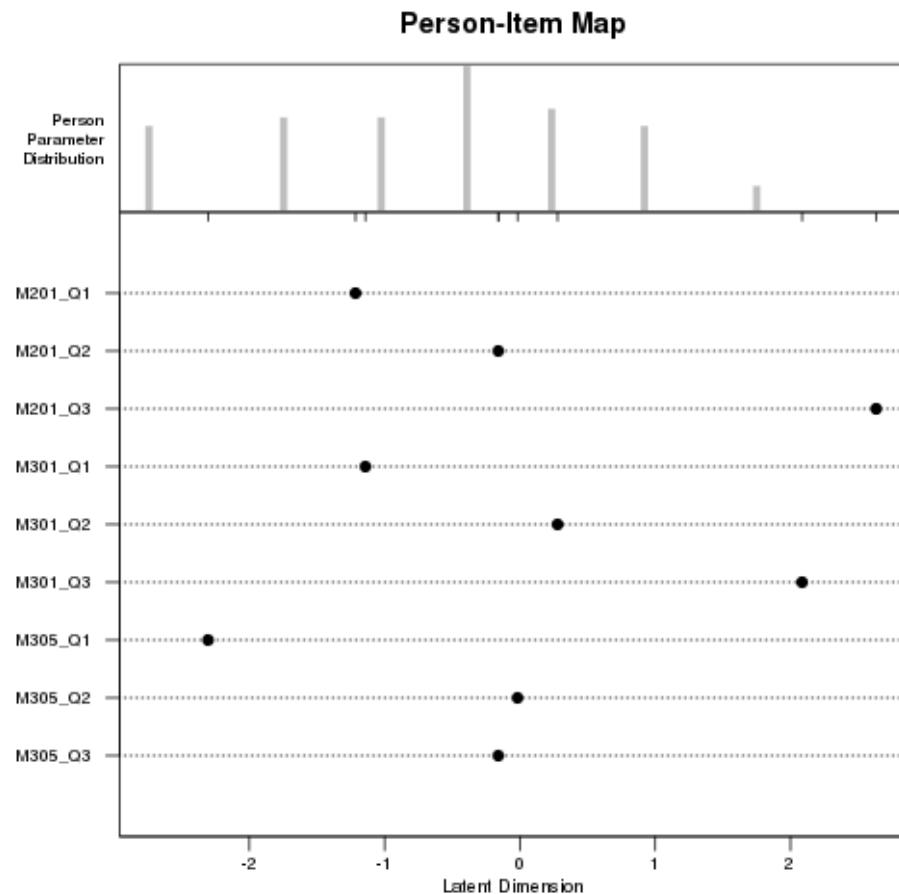
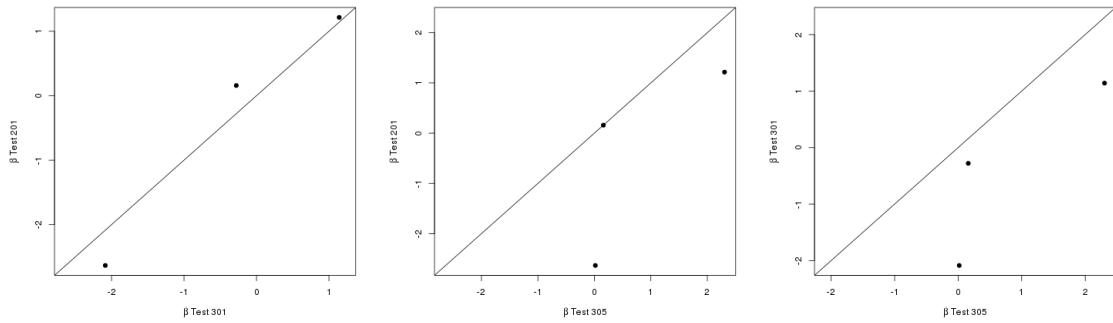


Abbildung 6: Person-Item-Map auf welcher die Verteilung der Personen basierend auf der latenten Skala ersichtlich ist und die Lage der Aufgabenparameter auf der latenten Skala. Anhand dieser Darstellung kann man sehen, dass z.B. der Qualitätsstandard 1 im Test 201 und im Test 301 einen sehr ähnlichen 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 7: Vergleich der Aufgabenparameter zwischen den einzelnen Tests. Wenn die Schwierigkeiten der Qualitätsstandards übereinstimmen würden, müssten alle Punkte auf der Winkel-Halbierenden liegen.

11.4. 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 8 und die wichtigsten Test Parameter in Tabelle 13.

Code erhältlich auf:

GitHub

<http://git.io/FVjL>

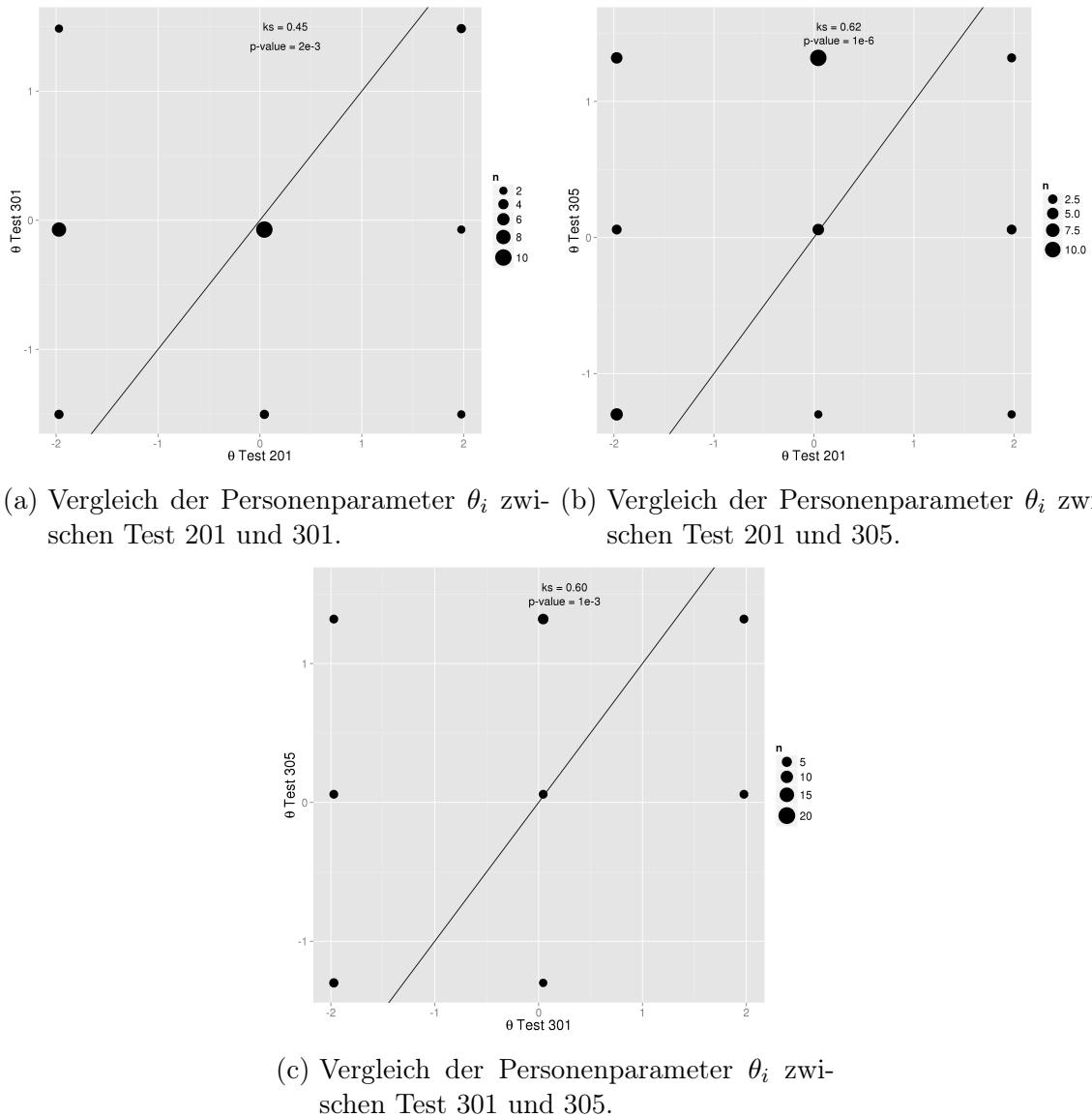


Abbildung 8: Vergleich der Personenparameter für die drei Tests. Zusätzlich sind der p-Wert des Kolmogorow-Smirnow-Test und die Test-Statistik ks angegeben.

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

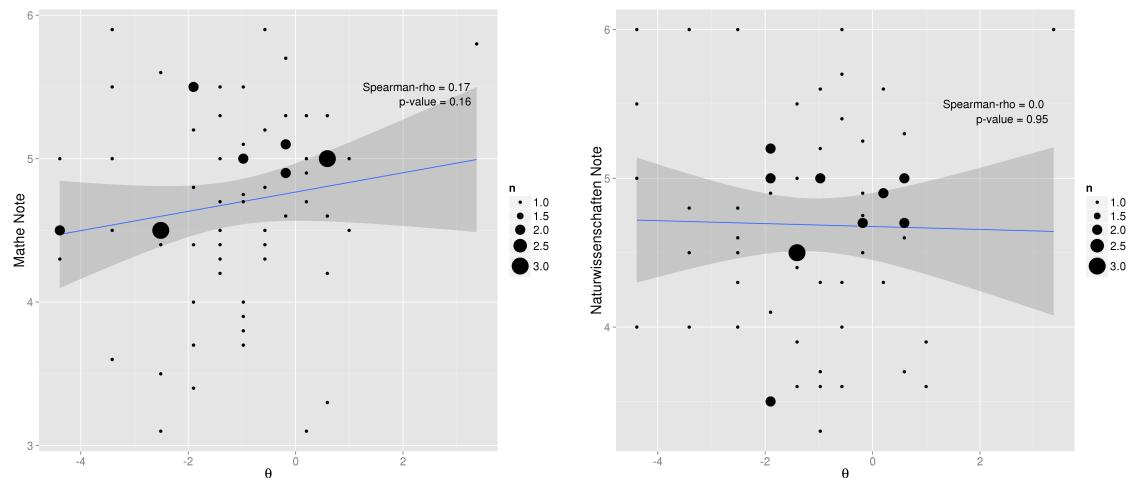
Tabelle 13: Resultate des Kolmogorow-Smirnow-Test für die Übereinstimmung der Personenparameter zwischen den drei Tests. Wobei ks die Test-Statistik des Kolmogorow-Smirnow-Test ist. Mit n wird die Anzahl an Personenparametern angegeben, welche für den Test verwendet werden konnten.

11.5. 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 14.

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 14: 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>

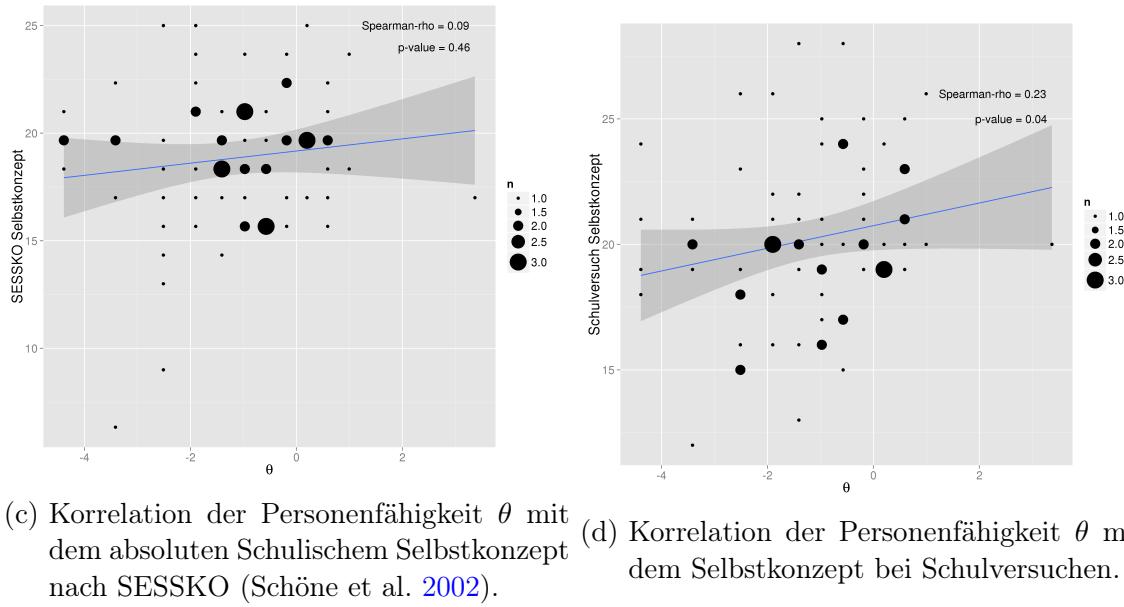


Abbildung 9: Korrelation zwischen der Personenfähigkeit θ und verschiedenen per Fragebogen erhobenen Daten. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die blaue Gerade ist die lineare Regression der zugrunde liegenden Daten, der dunkel graue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

12. 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 Schülern und Schülerinnen analysiert werden. So konnten von 8 Schülerinnen und Schülern die Aktionen per Video analysiert werden.

12.1. Qualitätsstandards

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

12.1.1. 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.

12.1.2. Messung wiederholen

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 15.

12.2. 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 10 und in Tabelle 16. In keinem der Korrelationstests wird die Signifikanzschwelle überschritten, daher gibt

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

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

12.3. Messzeitpunkte und Messdauer

Zusätzlich zu den zwei Merkmalen wurde für jede Messung noch erhoben, wann die Messung begonnen hatte und wann die Messung 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 werden 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 Thermometer über das Video nicht unterscheidbar waren. Daher

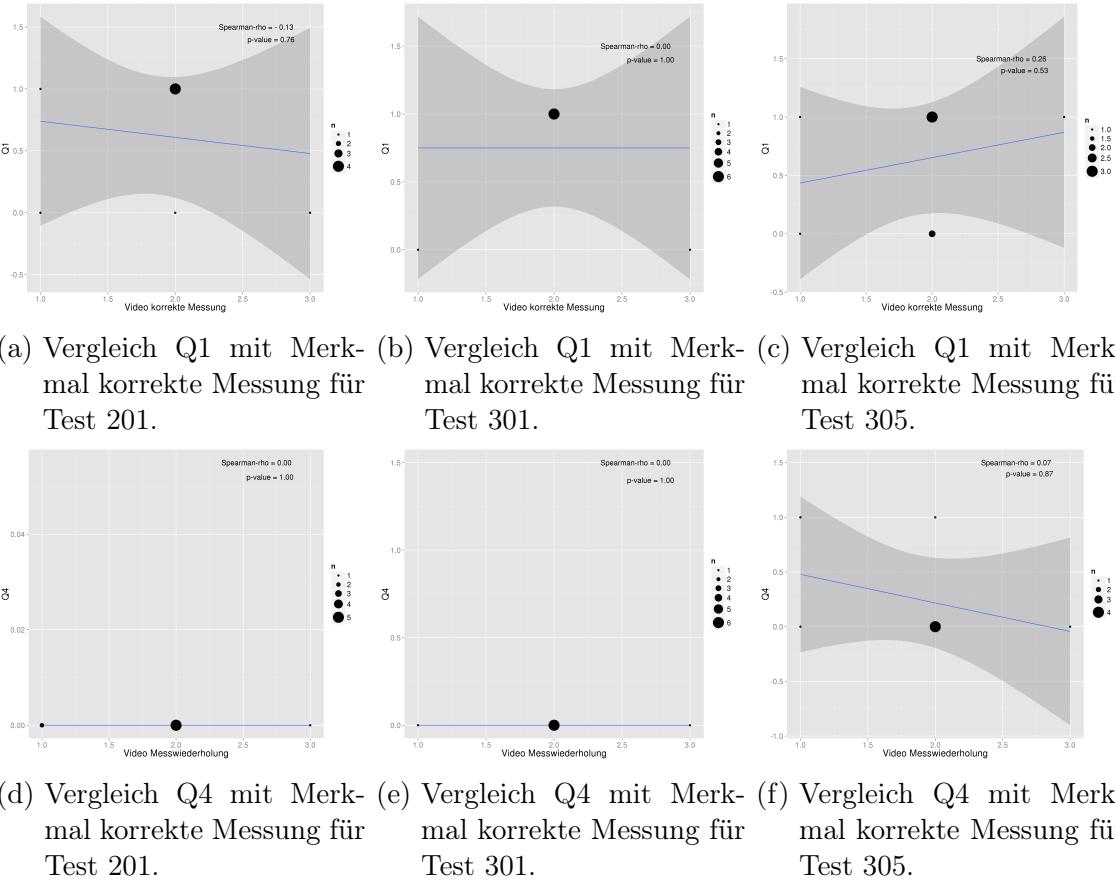


Abbildung 10: Vergleich der Merkmale der Videokodierung mit den Qualitätsstandards 1 und 4. Der Durchmesser der Punkte ist ein Mass für die Anzahl an Datenpunkten, welche an dieser Position liegen. Die blaue Gerade ist die lineare Regression der zugrunde liegenden Daten, der dunkel graue Bereich stellt das Vertrauensintervall (95%) der linearen Regression dar. Zusätzlich sind noch Spearmans ρ und der p-Wert des Signifikanztests angegeben.

wurden dort die Messinstrumente mit 1 und 2 kodiert. Die Resultate finden sich in Darstellung 11.

Code erhältlich auf:

GitHub

<http://git.io/bvQS>

13. Diskussion

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

14. Kodierung

14.1. 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 3 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 prozedurale 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 hatten. 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 ist und die Kodierung korrekt und nachvollziehbar ist.



Abbildung 11: Übersicht über alle Messungen der videografierten Schülerinnen und Schüler. In der ersten Spalte ist der Identifizierungs-Code. In der Spalte 2 welcher Test wo 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)

14.2. 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 Qualitätsstandards basiert auf den Arbeiten von Metzger et al. (2013) und Hild, Metzger und Parchmann (2014b). 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 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 (2015a). Dieser Unterschied in den erreichten Qualitätsstandards deckt sich jedoch mit den Ergebnissen von Metzger et al. (2013).

14.3. Niveaus

Dieses schlechte Abschneiden der Klassen spiegelt sich auch in den erreichten Niveaus wieder. So sieht man in Tabelle 5, 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.

15. 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 10 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 6 und Tabelle 7). 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.

16. 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 nicht geeignet ist (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 8, 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.

17. 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 9 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, das beim bedingten Niveau nur sehr wenig hohe Werte erreicht wurden (siehe Tabelle 5). Daher kommt es zu einer schlechten Datenmenge bei Niveaus über 2, dies kann man auch sehr gut in den Darstellungen 2 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 abhängig ist in welchem Test dies erreicht wurde, sondern rein von der Kompetenz des skalenbasierten Messens abhängt.

18. 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.

18.1. Parameterschätzung

Ein grosses Problem bei der Rasch-Analyse ist die Parameterschätzung. Das grösste Problem 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 zwar die Aufgabenschwierigkeiten β übereinstimmen geschätzt werden (siehe Darstellung 3). Nach Rost (2004) ist diese Schätzung jedoch deutlich unkritischer, wie die der Personenparameter. Bei den Personenparametern θ gibt es jedoch Unterschiede zwischen beiden Schätzern. Bei der bedingten 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. Desse[n] 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.

18.2. 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). 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), als auch mit dem Wald-Test (siehe Tabelle 10). 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.

18.3. 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 5. 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 in der Schwierigkeit. Dies liegt höchstwahrscheinlich daran, dass dieser Test im Vergleich zu den anderen beiden Test leichter ist (Sichau 2015a). Dies sieht man auch in der Darstellung 6.

Zusätzlich zu der qualitativen Überprüfung wurde noch ein Kolmogorow-Smirnow-Test durchgeführt, um festzustellen ob die Unterschiede in den Aufgabenparametern (siehe Tabelle 11) signifikant sind. Die Testergebnisse in Tabelle 12 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 7.

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

18.4. 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 13 und Darstellung 8 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 13, 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 Grosssteils auf Problemen mit dem Parameterschätzer. Auch der marginale Maximum-Likelihood Schätzer hatte massive Probleme mit dem Datensatz und war noch schlechter, daher wurden dessen Ergebnisse nicht präsentiert.

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

18.5. 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 14 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 sehr 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.

19. Videoanalyse

In einem letzten Schritt wurden noch die Videos analysiert. Die dabei entwickelten Merkmale wurden mit den Qualitätsstandards korreliert. Wie in Tabelle 16 und Darstellung 10 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 Test Durchführung gemacht wurden, lassen sich diese Ergebnisse jedoch erklären. Viele Schülerinnen und Schüler waren während der Test Durchführung sehr auf die experimentelle Seite

fokussiert und haben insgesamt sehr wenig auf den Datenbögen ausgefüllt. Dies zeigt sich auch im insgesamt eher schlechtem Abschneiden der Schülerinnen und Schüler (siehe Tabelle 5). Daher widerspiegeln die Qualitätsstandards nur den Teil des Experimentes 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 11 gute 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.

20. 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 sollte, müsste die Untersuchungsgruppe massiv vergrössert werden.

Das Resultat dieser Arbeit ist daher, dass:

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

21. Ausblick

22. Datengrundlage

Mit dieser Arbeit wurde ein erster Versuch durchgeführt, zu zeigen, dass bestimmte Kompetenzen kontextunabhängig sind. Für einen 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 bestärkt werden und keine gegensätzlichen Resultate gefunden werden. Die bisherigen Ergebnisse sind jedoch aufgrund der zu geringen Stichprobe nicht generalisierbar.

23. 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 es 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 gehoben 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.

24. 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, treffen. 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 zugrunde liegenden 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 einen 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 (bei der Publikation), damit andere Personen die Resultate nachvollziehen können, wie dies z.B. in der Bio-Informatik Standard ist. Ws wurde versucht diese Forderung in dieser Arbeit umzusetzen, daher hier nochmals der Link zu dem vollständigen Source-Code der Auswertung.

Code erhältlich auf:

GitHub

<http://git.io/buGR>

Literaturverzeichnis

- 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.
- 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.
- 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.
- 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.
- 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.
- Hild, Pitt, Susanne Metzger und Ilka Parchmann 2014. "Individuelle Förderung experimenteller Kompetenzen mit Lernaufgaben". 2014a. In *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht*, herausgegeben von Sascha Bernholt, 477–479.
- Hild, Pitt, Susanne Metzger und Ilka Parchmann 2014. "Using feedback and feed forward to foster experimental competence in student-centered learning environments". 2014b.
- 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.
- Landis, J R, und G G Koch 1977. "The measurement of observer agreement for categorical data." 1977. *Biometrics* 33 (1): 159–174.
- 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.
- 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.
- 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.
- 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.
- 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.

Sichau, David 2015. "Entwicklung eines hands-on Experimentiertests zur Messung der Kompetenz des skalenbasierenden Messens". 2015a. Forschungsarbeit, PH Zürich.

Sichau, David 2015. "Masterarbeit". 2015b. Masterarbeit, PH Zürich.

Strobl, Carolin 2012. *Das Rasch-Modell*. 2012. 132. Rainer Hampp Verlag.

Anhang

A. 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

B. 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>

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

C. Fragebogen

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

C.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 Schulversuche gefallen. Schreibe bitte eine kurze Begründung.

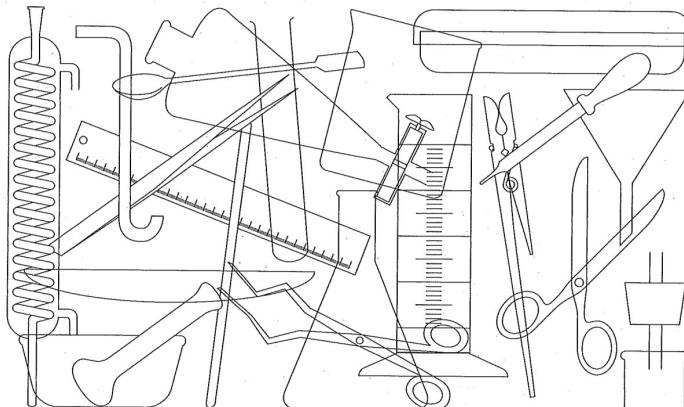
4. Welchen Versuch sollte man unbedingt im Unterricht durchführen und warum?

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

C.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.

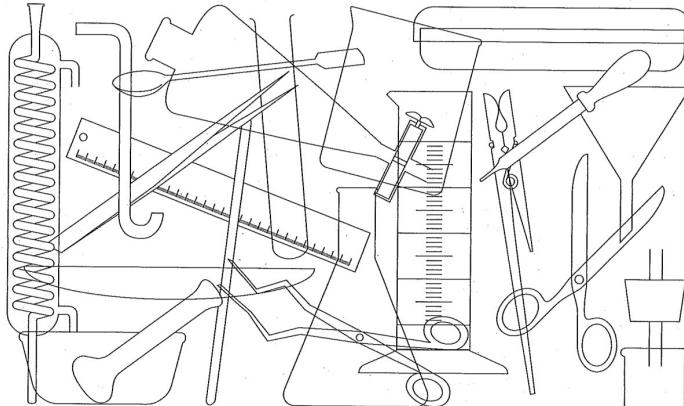
4. Was könnte man an den Versuchen verbessern?

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

D. Aufgabenstellung und Kodierungen

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

D.1. Test 201: Aufgabenstellung

Salz lösen

Problem

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

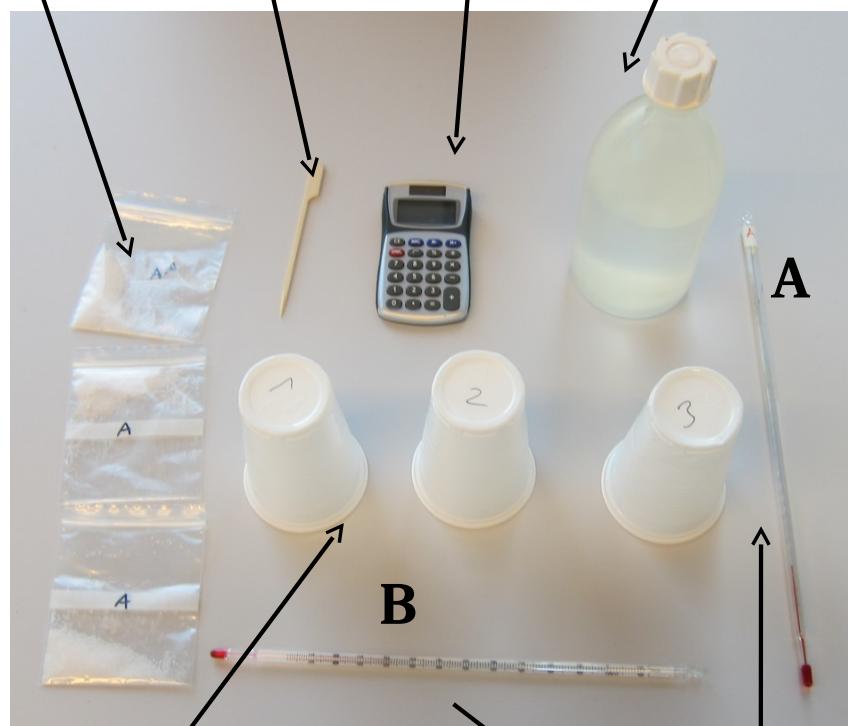
Material

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.**

D.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. <p>Erklärung: Die Raumtemperatur, sowie die Temperatur des Leitungswassers wurden im Vorfeld gemessen und liegen alle in einem Bereich zwischen 19°C und 25°C.</p>
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)

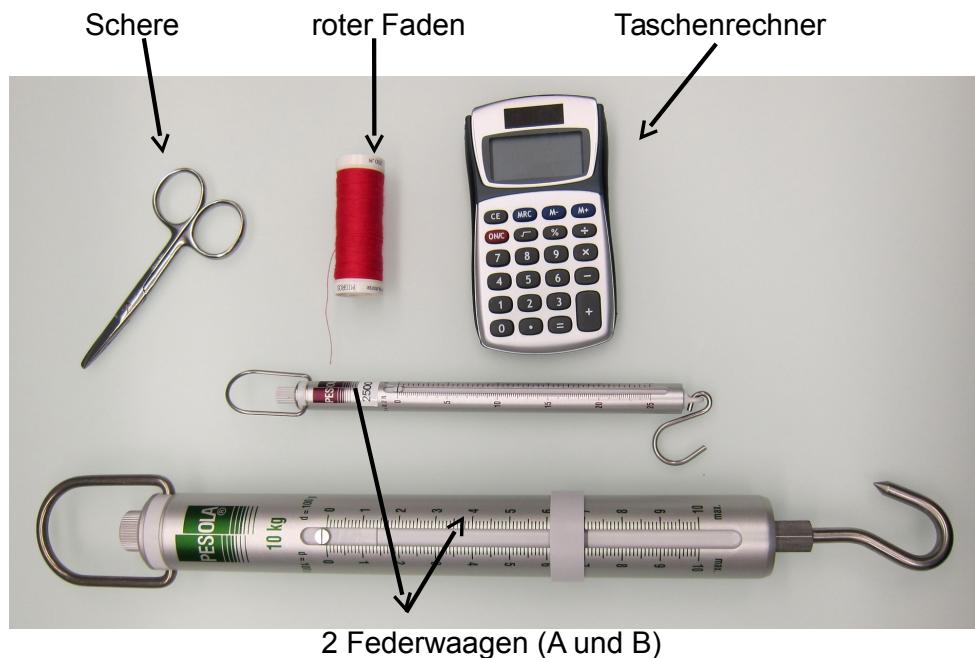
D.3. Test 301: Aufgabenstellung

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.**

D.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)

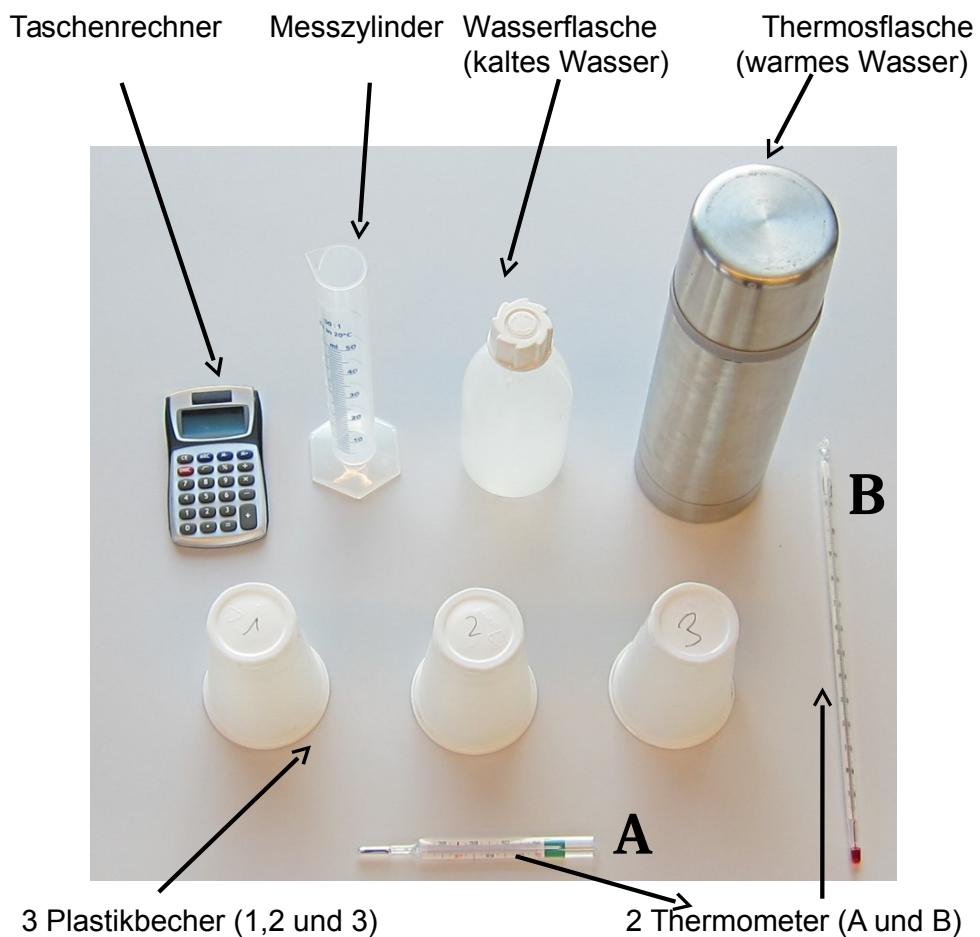
des

D.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.**

D.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 Messinstruments korrekt?	Wahl des Messinstruments mit korrekter Skala: B
3.2	i06	1P Wird die Wahl des Messinstruments 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	Lösungsvorschläge 1. Verbesserungen bei der Messtechnik 2. Messwiederholung und „Mittelwertbildung“ Messwert-Selektion 3. Wahl Messinstrument (Messinstrument mit feinerer Skala)

E. 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

--	--