Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen von 12- bis 15-jährigen Jugendlichen

Christoph Gut*, Susanne Metzger[+]*, Pitt Hild[+], Josiane Tardent[+]

*MINT-Cluster, Pädagogische Hochschule Zürich
[+]Zentrum für Didaktik der Naturwissenschaften, Pädagogische Hochschule Zürich
christoph.gut@phzh.ch, susanne.metzger@phzh.ch, pitt.hild@phzh.ch, josiane.tardent@phzh.ch

Kurzfassung

Im Rahmen der Entwicklung nationaler Bildungsstandards wurde in der Schweiz 2008 ein interdisziplinärer, auf der Unterscheidung von Teilprozessen aufbauender large-scale-Experimentiertest bei 807 Schülerinnen und Schülern der Jahrgangsstufen 6 und 9 durchgeführt (Projekt HarmoS). Eine Analyse der Testvalidität zeigte betreffend Itemabhängigkeiten, kompetenzirrelevanten Anforderungen, Generalisierbarkeit und Interpretierbarkeit der Experimentieraufgaben beeinträchtigende Mängel auf. Im derzeit laufenden Folgeprojekt ExKoNawi (Experimentelle Kompetenzen in den Naturwissenschaften) wird versucht, diese Validitätsprobleme mit einer problemtypenbasierten Kompetenzmodellierung in zufriedenstellender Weise zu lösen. Dabei werden hands-on Aktivitäten in den Fächern Biologie, Chemie und Physik als fächerübergreifende experimentelle Problemtypen wie «Beschreibung qualitativer Beobachtungen», «Messung quantitativer Grössen», «Durchführung experimenteller Vergleiche» und «Untersuchung korrelativer Zusammenhänge» modelliert. 2012/13 wurden Pilottests mit insgesamt 12 Aufgaben bei rund 450 Schülerinnen und Schülern der Jahrgangsstufen 7 bis 9 validiert. Die Auswertung eines Teiltests mit Untersuchungs- und Vergleichsaufgaben belegt, dass mit einer problemtypenbasierten Kompetenzmodellierung ein interdisziplinäres Messinstrument mit guten Item-Fitwerten entwickelt werden kann, das auf tiefem Leistungsniveau signifikante und interpretierbare Leistungsdifferenzen misst. Dabei konnten im Vergleich zum HarmoS-Experimentiertest in allen vier die Validität beeinflussenden Aufgabenaspekten (Itemabhängigkeit, kompetenzirrelevante Anforderungen, Generalisierbarkeit, Interpretierbarkeit) Verbesserungen erzielt werden. Trotz positiver Resultate bleibt die Sicherstellung einer ausreichend hoher Testreliabilität im Hinblick auf die geplante Modellvalidierung die vordringlichste Aufgabe des Projekts ExKoNawi.

1. Einleitung

Mit der zu Beginn des Jahrhunderts einsetzenden Diskussion über nationale Bildungsstandards und kompetenzorientierte Lehrpläne wurden in der Schweiz Forschungsaktivitäten zur Modellierung und Messung naturwissenschaftlicher Kompetenzen initiiert. Dabei konnten in den vergangenen Jahren verschiedene Erfahrungen mit interdisziplinären large-scale Experimentiertests gesammelt werden. 2008 wurde im Rahmen des mittlerweile abgeschlossenen HarmoS-Projekts (Harmonisierung der obligatorischen Schule) ein large-scale Experimentiertest durchgeführt [1], bei dem u. a. auf Erfahrungen mit dem TIMSS-Experimentiertest zurückgegriffen wurde [2, 3]. Die Analyse dieses Experimentiertests deckte verschiedene Schwierigkeiten mit dem Messformat auf, die mit einem Mangel an Validität einhergehen [4, 5]. In einem derzeit laufenden Folgeprojekt ExKoNawi (Experimentelle Kompetenzen in den Naturwissenschaften) wird derzeit versucht, mit Hilfe eines auf der Unterscheidung experimenteller Problemtypen basierenden Modellierungsansatzes verschiedene Validitätsprobleme auf zufriedenstellende Weise zu lösen [6, 7]. In Vorbereitung auf eine Validierung des Kompetenzmodells von ExKoNawi wurden 2013 Pilottests entwickelt und durchgeführt. Die Ergebnisse dieser Tests geben Antwort auf die Frage, inwieweit die erwähnten Validitätsprobleme gelöst werden können.

Im folgenden Beitrag werden zunächst die beiden Kompetenzmodellierungen von HarmoS und ExKoNawi theoretisch eingeordnet und verglichen. Anschließend werden am Beispiel des HarmoS-Experimentiertests die erwähnten Validitätsprobleme besprochen, auf deren Basis das interdisziplinäre Kompetenzmodell von ExKoNawi begründet wird. Abschließend werden ausgewählte Ergebnisse der Pilottests vorgestellt und in Bezug auf Validitätsfragen diskutiert.

2. Rahmentheorie zur Modellierung und Messung experimenteller Kompetenzen

Zur Modellierung und Messung experimenteller Kompetenzen mit Hilfe von Experimentiertests gibt es umfangreiche Literatur (vgl. [4, 8]). Die vielfälti-

gen empirischen Befunde sind jedoch häufig wenig vergleichbar, da den jeweiligen Ansätzen unterschiedliche Modellierungsentscheidungen zugrunde liegen. Diese Entscheidungen werden im Hinblick auf normative Ziele a priori gefällt und unterliegen daher nicht der direkten empirischen Überprüfung.

Kompetenzumfang: Modellierungsansätze unterscheiden sich oft bereits in der *äußeren Abgrenzung experimenteller Kompetenzen*. Im naturwissenschaftlichen Unterricht werden unterschiedlichste Tätigkeiten dem Experimentieren zugeordnet. Dazu gehören neben Fähigkeiten und Fertigkeiten im Umgang mit Experimentiermaterial auch rein kognitive Aufgaben, die sich bei der Vor- und Nachbereitung von Experimenten ergeben, wobei mit dem Experimentieren nicht nur die Gewinnung von Erkenntnis, sondern auch die Erreichung technischer Ziele verbunden wird. Die Modellierung experimenteller Kompetenzen bedingt zuallererst die Abgrenzung des Kompetenzumfangs. Es muss a priori entschieden werden, welche dieser Tätigkeiten und Aufgaben zu den Kompetenzen gezählt werden.

Sowohl der HarmoS-Experimentiertest als auch beim Projekt ExKoNawi werden experimentelle Kompetenzen interdisziplinär interpretiert und umfasst nicht nur Erkenntnisgewinnungsprozesse, sondern auch konstruktive Elemente des Experimentierens.

Strukturmodelle: Eine weitere Diskrepanz von Forschungsansätzen besteht in der *inneren Abgrenzung von Teilkompetenzen*. Im Zusammenhang mit Experimentiertests werden in der Literatur im Wesentlichen zwei Ansätze diskutiert [7, 9]. Zum einen wird das Experimentieren als Zusammenspiel unterscheidbarer Teilprozesse verstanden [4, 8]. Der idealisierte Experimentierprozess zeichnet sich in dieser Denkweise durch eine klare Abfolge von Teilprozessen ab, die häufig wie folgt beschrieben wird: Fragestellung, Hypothesenformulierung, Planung und Durchführung des Experiments, Auswertung der Daten, Reflexion. Wird diese Abfolge in Experimentieraufgaben berücksichtigt, können zu den „science processes“ korrespondierende „skills“ verglichen werden (z. B. [2, 10, 11, 12, 13]). Der Ansatz erlaubt zudem, rein kognitive Teilprozesse zu isolieren und mit weniger aufwendig zu entwickelnden Papier-und-Bleistift-Tests zu evaluieren (z. B. [14, 15, 16, 17]).

Ein zweiter Ansatz beschreibt das Experimentieren als integralen Problemlöseprozess, bei dem Objekte oder Zusammenhänge klassifiziert, verglichen, konstruiert, gemessen oder untersucht werden. Dabei wird auf unterschiedliches Wissen wie „conceptual knowledge“ oder „procedural knowledge“ zurückgegriffen [9, 18, 19]. Nach dieser Sichtweise unterscheiden sich verschiedene Experimentieraufgaben nach dem Typ des Problems, das durch das Experimentieren gelöst werden soll [20].

Der wesentlichste Unterschied zwischen HarmoS und ExKoNawi besteht in der inneren Abgrenzung von Teilkompetenzen. Während mit HarmoS der Teilprozessansatz umgesetzt wurde, werden bei ExKoNawi Problemtypen unterschieden.

Progressionsmodelle: Ein dritter wesentlicher Unterschied von Forschungsansätzen betrifft die *Modellierung der Kompetenzprogression*. Eine Kompetenz kann grundsätzlich in mehreren Richtungen weiterentwickelt werden [4, 19, 21]. Eine Person kann kompetenter werden, indem sie lernt, komplexere Probleme zu lösen (*Problemkomplexität*), bestimmte Probleme qualitativ besser (*Lösungsqualität*), eigenständiger (*Eigenständigkeit*), in mehr fachlichen Kontexten (*Transfervermögen*) oder stabiler zu lösen (*Performancestabilität*). Mit einem so genannten Progressionsmodell wird a priori festgelegt, welche der fünf Kompetenzprogressionen (Problemkomplexität, Lösungsqualität, Eigenständigkeit, Transfervermögen und Performancestabilität) mit welchen Kriterien gemessen werden sollen. Dabei bedarf das Modell je nach modellierter Progression einer internen oder externen Validierung a posteriori [7]. Bei large-scale Assessments werden jedoch die verschiedenen Progressionen nicht gleichwertig berücksichtigt. Während zum Beispiel in der Schulpraxis der Förderung der Eigenständigkeit beim Experimentieren viel Gewicht zukommt, wird die Eigenständigkeit im large-scale höchstens implizit erfasst, was als Mangel an inhaltlicher Validität solcher Assessments gewertet werden kann. Auch das Transfervermögen und die Performancestabilität werden nicht explizit erhoben. Vielmehr wird vorausgesetzt, dass die zu vermessende Schülerstichprobe über ausreichend Transfervermögen und Performancestabilität verfügt, denn andernfalls könnte kein psychometrisches Konstrukt gemessen werden. In Bezug zu den zwei verbleibenden Progressionen Problemkomplexität und Lösungsqualität unterscheiden sich die Modellierungen fundamental: Wird die Problemkomplexität a priori modelliert, müssen dazu Items von unterschiedlicher Komplexität entwickelt werden. Mit einem Vergleich der empirisch erfassten Itemschwierigkeiten und den angenommenen Komplexitätsniveaus wird das Progressionsmodell intern validiert. Wird hingegen die Lösungsqualität a priori modelliert, werden standardisierte Aufgaben entwickelt. Die Kompetenzeinschätzung erfolgt sodann anhand der a priori gesetzten Qualitätsstufen. Mit einem Vergleich dieser Stufen mit anderen Skalen wird das Modell extern validiert.

Sowohl bei HarmoS als auch bei ExKoNawi werden experimentelle Kompetenzen mittels der Lösungsqualität erfasst. Während der HarmoS-Experimentiertest jedoch ohne ein a priori-Progressionsmodell entwickelt wurde, baut das Projekt ExKoNawi explizit auf einer solchen Progressionsmodellierung auf.

Messinstrument: Modellierungsansätze unterschei-

den sich letztlich im verwandten Testformat (Papier-und-Bleistift-, Experimentier- oder Simulationstest), in der Art, wie gemessen wird (Eigen- oder Fremdrapportierung von Experimentieraktivitäten), und bezüglich der Auswertung der Daten (klassische oder probabilistische Testtheorie).

Sowohl bei HarmoS als auch bei ExKoNawi wird mit hands-on Aufgaben gearbeitet, wobei die Testscores rasch-skalierbar sein sollen.

3. Validitätsprobleme interdisziplinärer large-scale Experimentiertests

Der HarmoS-Experimentiertest wurde mit dem Ziel entwickelt, die Experimentierfähigkeiten Schweizer Jugendlicher im Hinblick auf die Formulierung von Basisstandards für den integrierten Naturwissenschaftsunterricht zu ermitteln. Um den heterogenen Unterrichtserfahrungen der Schülerschaft gerecht zu werden, wurden zu 14 Kontexten aus Biologie, Chemie und Physik authentische, für die Schulpraxis relevante und möglichst vielfältige Experimentieraufgaben entwickelt. Über die in Bezug auf die Aufgabenstellungen und die Kodierschemen heterogenen Aufgaben wurde eine gemeinsame Teilprozessstruktur gelegt, wobei die 14 Aufgaben in 95 Items zu verschiedenen Teilprozessen zerlegt wurden [4]. Der HarmoS-Experimentiertest wurde in der Deutschschweiz mit 807 Schülerinnen und Schülern der Jahrgangstufen 6 und 9 evaluiert. Die Auswertung zeigt verschiedene Validitätsprobleme des Tests auf [4, 5].

Problem lokaler Itemabhängigkeiten: Die Abgrenzung von Teilprozessen konnte im HarmoS-Experimentiertest statistisch nicht verifiziert werden. Rasch-analytische Dimensionsvergleiche ergaben zwar den besten Modellfit, wenn die Teilprozesse, bei denen mit Experimentiermaterial gearbeitet wird, den rein kognitiven Teilprozessen gegenübergestellt werden (Tab. 1). Die mangelhaften Reliabilitäten und die sehr hohe Korrelation zwischen den Dimensionen lassen jedoch keine Mehrdimensionalität des Tests vermuten. Grund hierfür wird u. a. in *lokalen Itemabhängigkeiten* vermutet, die immer dann entstehen, wenn eine zusammenhängende Experimentieraufgabe in Teilschritte zerlegt wird.

	Teilprozesse	
	kognitiv-manipulativ	**rein kognitiv**
kognitiv-manipulativ	***Rel.* = 0.55**	***r* = 0.99**
rein kognitiv		***Rel.* = 0.56**

Tab. 1: Rasch-analytische dimensionale Modellvergleiche zum HarmoS-Experimentiertest (95 Items, $N = 807$): Bester Modellfit in Bezug auf Teilprozesse mit EAP/PV-Reliabilitäten und Korrelation.

Generalisierbarkeitsproblem: Dimensionale Modellvergleiche in Bezug auf Kontexte ergaben den besten Fit für ein zweidimensionales Modell mit biologischen und stofflichen Items (Biologie, Chemie) als eine Dimension, mechanischen und elektrischen Items (Physik) als andere Dimension (Tab. 2). Die geringen Reliabilitäten und die kleine Korrelation zwischen den Dimensionen deuten auf eine mangelhafte Generalisierbarkeit der Aufgaben hin.

	Kontexte	
	biologisch, stofflich	**mechanisch, elektrisch**
biologisch, stofflich	***Rel.* = 0.40**	***r* = 0.49**
mechanisch, elektrisch		***Rel.* = 0.50**

Tab. 2: Rasch-analytische dimensionale Modellvergleiche zum HarmoS-Experimentiertest (95 Items, $N = 807$): Bester Modellfit in Bezug auf Kontexte mit EAP/PV-Reliabilitäten und Korrelation.

Interpretierbarkeitsproblem: Mit dem HarmoS-Test wurden signifikante Leistungsunterschiede zwischen den Jahrgangsstufen und Schulniveaus gemessen. Zum Beispiel wurde auf der Jahrgangstufe 9 zwischen dem Gymnasium und den nicht gymnasialen Sekundarschulen B/C und A (entsprechen der Hauptschule und Realschule in Deutschland) ein großer und hochsignifikanter Leistungsunterschied ermittelt (Tab. 3). Leider lässt der Test post hoc keine Aussage darüber zu, worin gymnasiale Klassen besser experimentieren als nichtgymnasiale Klassen. Der Mangel an Interpretierbarkeit hat drei Ursachen: Erstens verunmöglicht die Aufgabenheterogenität eine sinnvolle Strukturmodellierung. Zweitens wurde der HarmoS-Test ohne a priori Modellierung der Kompetenzprogression entwickelt. Drittens verliert man bei der in large-scale Assessments üblichen Bewertung der Lösungsqualität mit einem Gesamtscore Informationen über dessen Zusammensetzung. Dies gilt insbesondere für Rasch-skalierte Tests, bei denen aus der Qualität von facettenreichem Experimentierverhalten auf eine latente Fähigkeitsvariable geschlossen wird.

Jahrgangsstufe 9					
Sek B/C ≈ Hauptschule	**Δ**	**Sek A** ≈ Realschule	**Δ**	**Gym** Gymnasium	**alle**
467 (95)	16† (*d* = 0.2)	**483** (96)	71*** (*d* = 0.7)	**554** (87)	**500**$^{+}$ (100)

Tab. 3: Eindimensionale Rasch-Skalierung des HarmoS-Experimentiertests (95 Items, $N = 807$): Mittelwerte mit Standardabweichung der Fähigkeitsvariable in PISA-Metrik^{+}, Mittelwertvergleiche von Schulniveaus mit Signifikanzen († t-Test, ansonsten Mann-Whitney-Test).

Problem kompetenzirrelevanter Aufgabenanforderungen: Letztlich konnte gezeigt werden, dass die Messung mit dem HarmoS-Experimentiertest nicht

unwesentlich durch nicht vernachlässigbare kompetenzirrelevante Anforderungen der Experimentieraufgaben beeinflusst wird [4, 13].

4. Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen

Mit dem Projekt ExKoNawi wird der Versuch unternommen, die beschriebenen Validitätsprobleme des HarmoS-Experimentiertests – sprich das Generalisierungsproblem, das Problem kompetenzirrelevanter Aufgabenanforderungen, das Interpretierbarkeitsproblem von Testscores bzw. latenter Fähigkeitsparameter sowie das Problem lokaler Itemabhängigkeiten bei zusammenhängenden Experimentieraufgaben – auf möglichst zufriedenstellende Weise zu lösen. Hierzu wurde ein interdisziplinäres Kompetenzmodell entwickelt und ein Testinstrument bestehend aus Experimentieraufgaben zu 12 verschiedenen Kontexten erarbeitet. Mit Hilfe zweier Pilottests wurde die Güte der Experimentieraufgaben überprüft. Im Folgenden werden das Kompetenzmodell und das Messinstrument dargestellt, anhand der Validitätsprobleme begründet sowie die Ergebnisse der Vorvalidierung vorgestellt.

Kompetenzmodellierung: Um das Generalisierungsproblem anzugehen, wurde ein Kompetenzstrukturmodell entwickelt, das auf dem Problemtypenansatz aufbaut. Der derzeit stark erforschte Teilprozessansatz wurde aus vier Gründen als ungeeignet erachtet: 1. Teilprozesse stellen keine charakteristischen Merkmale bestimmter Experimente dar. Gleiche Teilprozesse in unterschiedlichen Experimentieraufgaben tragen unterschiedlich viel zur Problemlösung bei [12] und werden auch nicht mit gleichen Kriterien beurteilt. Experimentelle Teilprozesse sind daher über verschiedene experimentelle Problemtypen hinweg kaum oder nicht vergleichbar. Eine Kompetenzstruktur, die auf Teilprozessen aufbaut, ermöglicht keine einheitliche Aufgabenkonstruktion. 2. Theoretische Überlegungen legen den Schluss nahe, dass die zu den Teilprozessen gehörenden „skills“ oft hauptsächlich von Fachwissen abhängen und daher keine transferfähigen Fähigkeiten darstellen [22, 23]. 3. Teilprozesse von zusammenhängenden Experimentieraufgaben können nur unter Inkaufnahme lokaler Itemabhängigkeiten modelliert werden. 4. Letztlich wird mit der Annahme einer idealisierten Abfolge von Experimentierprozessen ein falsches Bild vermittelt, naturwissenschaftliche Erkenntnis könne aufgrund einer standardisierten Prozedur gewonnen werden [24]. Demgegenüber erlaubt der gewählte Problemtypenansatz, die Vielfalt an experimentellen Aktivitäten in der Schulpraxis fächerübergreifend und umfassender zu beschreiben. In diesem Sinne wurden bei ExKoNawi vorerst folgende vier fächerübergreifende Problemtypen modelliert (Abb. 1): *kategoriengeleitetes Beobachten* [25, 26, 27], *skalenbasiertes Messen* [28, 29, 30], *fragengeleitetes Untersuchen* [31, 32, 33, 34, 35] und *effektbasiertes Vergleichen* [36, 37, 38].

Problemtypen		
	kategorien-geleitetes **Beobachten**	Phänomene anhand gegebener Kategorien (Fragen) beschreiben und vergleichen
	skalen-basiertes **Messen**	quantitative Grössen mit gegebenen Messinstrumenten (Skala) genau messen
	fragen-geleitetes **Untersuchen**	korrelative Zusammenhänge zwischen gegebenen Variablen (Frage) untersuchen
	effekt-basiertes **Vergleichen**	Objekte anhand einer gegebenen Eigenschaft experimentell (ohne direkte Messung) vergleichen

Abb. 1: Strukturmodell der Problemtypen von ExKoNawi

Mit Hilfe der über die Fächer hinweg übertragbaren Problemtypen soll ein erhöhter Transfer erreicht werden [39]. Der erhoffte Transfer beruht je nach Problemtyp einerseits auf dem gemeinsamen Verständnis des zu lösenden Problems (im Sinne eines „understanding of nature and purpose of task“ in [19], 92f) und andererseits auf experimentellem Strategiewissen (im Sinne von „concepts of evidence“ in [18], 795ff, wie Kontrollvariablenansatz, Messwiederholung, Fairness bei Vergleichen etc.), das zur Lösung des Problems benötigt wird. Das Modell kann prinzipiell durch Hinzunahme weiterer bereits angedachter Problemtypen, wie z. B. funktionsgeleitetes Konstruieren oder kriteriengeleitetes Klassifizieren, erweitert werden.

Die Kompetenzprogression wird für jeden Problemtyp separat modelliert, wobei maximal fünf Leistungsniveaus unterschieden werden, die mit der Qualität der Problemlösung zunehmen. Mit typenspezifischen Progressionsmodellen können spezifische, aus der Theorie und Erfahrung hergeleitete Qualitätsstandards einheitlich bewertet werden. Post hoc kann zudem überprüft werden, ob eine empirische Schwierigkeitsprogression der verschiedenen Qualitätsstandards festgelegt werden kann (Progressionsmodellierung). Experimentieraufgaben desselben Problemtyps sind so über die Kontexte hinweg vergleichbar. Für den Problemtyp «Vergleichen» beispielsweise betreffen diese Standards die Problemlösungen „2 Objekte qualitativ vergleichen“, „Bedingungen für fairen Vergleich nennen“, „drei Objekt in eine Rangfolge bringen“, „Differenzen von Objektmerkmalen vergleichen“ (Abb. 2).

Messinstrument: Im Rahmen einer Vorvalidierung des Kompetenzmodells wurde für jeden der oben beschriebenen Problemtypen je eine Experimentieraufgabe zu einem biologischen, chemischen und physikalischen Kontext entwickelt. Alle Aufgaben sind ausschließlich hands-on Experimente für die Sekundarstufe I (Jahrgangsstufen 7 bis 9), welche von den Schülerinnen und Schülern in 20 Minuten Einzelarbeit zu lösen sind. Um die kompetenzirrelevanten Aufgabenanforderungen konstant zu halten, wurde auf größtmögliche Homogenität der schriftlichen Aufgabenstellungen und Antwortformate innerhalb der Problemtypen geachtet. Die Aufgaben bestehen je nach Problemtyp aus zwei bis drei Teilaufgaben, die meist aufbauend gelöst werden (Abb. 2 und 3).

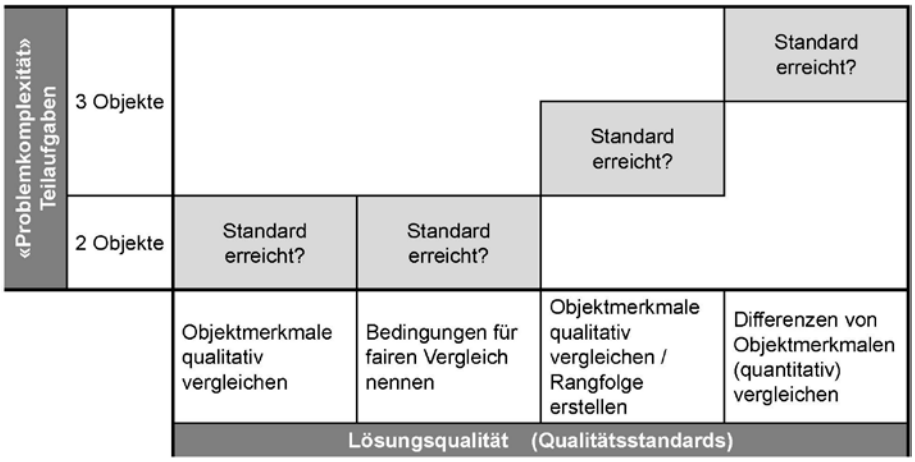

Abb. 2: Aufgabenkonstruktion mit Abfolge der Teilaufgaben für den Problemtyp «effektbasiertes Vergleichen».

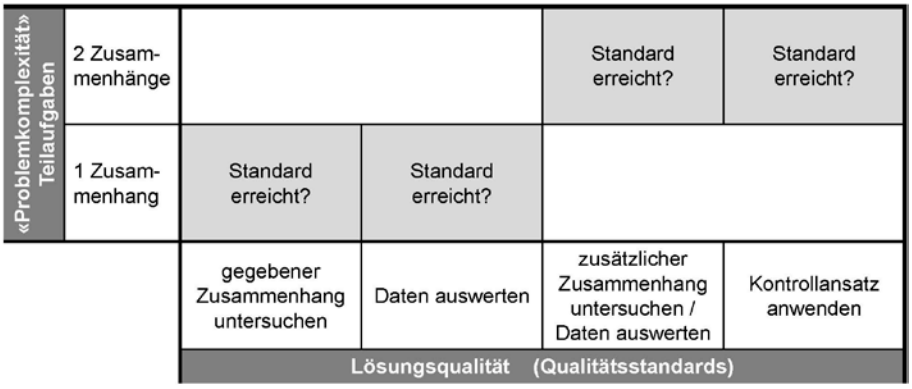

Abb. 3: Aufgabenkonstruktion mit Abfolge der Teilaufgaben für den Problemtyp «fragegeleitetes Untersuchen».

Die Konstruktion von Vergleichsaufgaben (Abb. 2) soll am Beispiel der adaptierten TIMSS-Aufgabe «Magnete» [4, 37, 38] erläutert werden: Die Schülerinnen und Schüler erhalten drei Magnete und eine Auswahl an magnetisierbaren und nicht magnetisierbaren Experimentiermaterialen (Abb. 4). In einer ersten Teilaufgabe sollen die Stärken von zwei Magneten qualitativ verglichen werden, dann soll der dritte Magnet hinzugenommen und eine Rangfolge der Magnete ermittelt werden. Zuletzt soll herausgefunden werden, welche zwei Magnete „ähnlicher“ sind, d. h. man muss die Differenzen zwischen den Magnetstärken vergleichen.

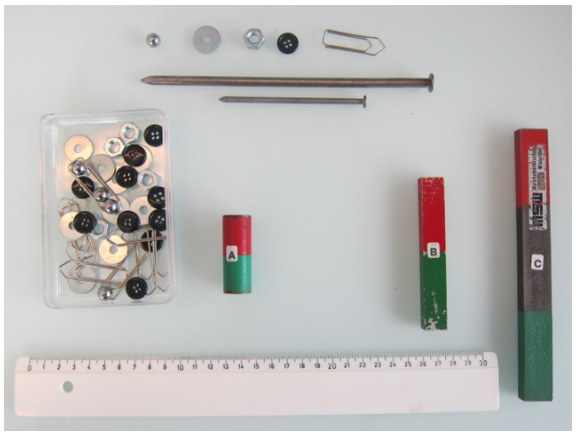

Abb. 4: ExKoNawi-Vergleichsaufgabe «Magnete».

Die Ergebnissicherung erfolgt mittels Eigenrapportierung in gedruckte Testhefte. Dabei werden die Jugendlichen zunächst jeweils aufgefordert, ihre Beobachtungen, Messungen, Untersuchungen oder Vergleiche zu protokollieren. Diese Protokolle werden durch explizite Fragen nach den Resultaten, Reflexionen über die Durchführung und das erhaltene Ergebnis sowie abschließende Kontrollfragen ergänzt. Jede Experimentieraufgabe wird als ein Item kodiert (Vermeidung lokaler Itemabhängigkeiten), wobei je nach Problemtyp drei bis fünf Qualitätsstandards bewertet werden (Abb. 2 und 3). Jeder Qualitätsstandard setzt sich in der Regel aus mehreren dichotom kodierten Kriterien zusammen, zu deren Beurteilung die Antworten zu allen Teilaufgaben berücksichtigt werden. Ein Qualitätsstandard ist erreicht, wenn eine a priori normativ festgelegte Zahl der Kriterien (zwischen der Hälfte und zwei Drittel der möglichen Punktzahl) erfüllt sind.

5. Validität der Pilottests

Pilotierung: Für die Vorvalidierung des Testinstruments wurden zwei Pilottests durchgeführt: Pilottest 1 (Herbst/Winter 2012) zu *kategoriengeleitetem Beobachten* und *skalenbasiertem Messen*, Pilottest 2 (Frühjahr/Sommer 2013) zu *effektbasiertem Vergleichen* und *fragengeleitetem Untersuchen*. Insgesamt nahmen 465 Schülerinnen und Schüler der Jahrgangstufen 7, 8 und 9 aus jeweils unterschiedlichen

Schulniveaus an den Tests teil, wobei jede Experimentieraufgabe insgesamt von mindestens 120 Jugendlichen bearbeitet wurde. Alle Lösungen wurden von zwei Personen kodiert, wobei die Rater-Übereinstimmung der 129 kodierten Kriterien der Antwortqualitäten größer als 0.68 war (mit Ausnahme von 10 Kriterien war sie größer als 0.8). Beide Pilottests bestehend aus jeweils 6 Items (Experimentieraufgaben) zu zwei Problemtypen und drei Kontexten (Biologie, Chemie, Physik) wurden separat ausgewertet. Die Itempunktzahlen (= Anzahl erreichte Qualitätsstandards) wurden jeweils mit dem Programm ConQuest 2.0 [40] mit einem eindimensionalen partial-credit-Modell Rasch-skaliert. Aufgrund des nicht hinreichenden Testdesigns wurde auf mehrdimensionale Skalierungen und somit auf eine Vorvalidierung des Strukturmodells verzichtet. Im Weiteren beschränken wir die Validitätsanalyse auf den Piltottest 2, die Auswertung des ersten Pilottests wurde bereits unter [6] veröffentlicht.

Vorvalidierung der Progressionsmodelle: Der Vergleich der Häufigkeiten der erreichten Qualitätsstandards zeigt sowohl für den Problemtyp «Untersuchen» als auch für den Problemtyp «Vergleichen» eine mehrheitlich konsistente Progression der „Schwierigkeiten“ (Abb. 5 und 6). B/C-Schüler erreichen Qualitätsstandards grundsätzlich weniger häufig als A-Schüler. Dies gilt auch für den Vergleich der Jahrgangsstufen 7 und 9, mit Ausnahme der sehr guten Ergebnisse der 7. Sekundarklassen A bei den Vergleichsaufgaben. Diese Abweichung könnte als Effekt des Curriculums interpretiert werden, wonach die besonders einfachen Aufgaben jüngeren Schülerinnen und Schülern präsenter sind als älteren Schülerinnen und Schülern.

Testqualität: Die eindimensionale Rasch-Skalierung ergab für die 6 Experimentieraufgaben des Piltotests 2 (Items) sehr gute Fitwerte (Tab. 4): Die standardisierten Abweichungen der Infit- und Outfitwerte vom Idealwert 1 betragen nicht mehr als 1.2, Item-Separations-Reliabilität und Trennschärfe liegen ausreichend hoch. Die Reliabilitäten von 0.51 bzw. 0.65 deuten zwar wenig Generalisierbarkeit der Performanz über verschiedene Problemtypen an, sind jedoch höher als beispielsweise beim HarmoS-Experimentiertest (Tab. 1 und 2). Es wird erwartet, dass mit der Verdoppelung der Testlänge und der dimensionalen Unterscheidung von Problemtypen hinreichend gute Reliabilitäten erreicht werden.

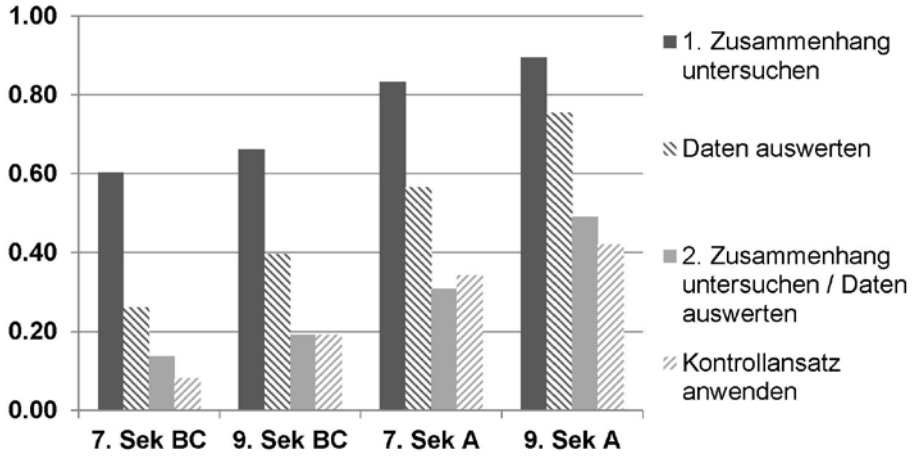

Abb. 5: Häufigkeiten der erreichten Qualitätsstandards summiert über alle drei Aufgaben zum Problemtyp «fragegeleitetes Untersuchen» (Pilottest 2).

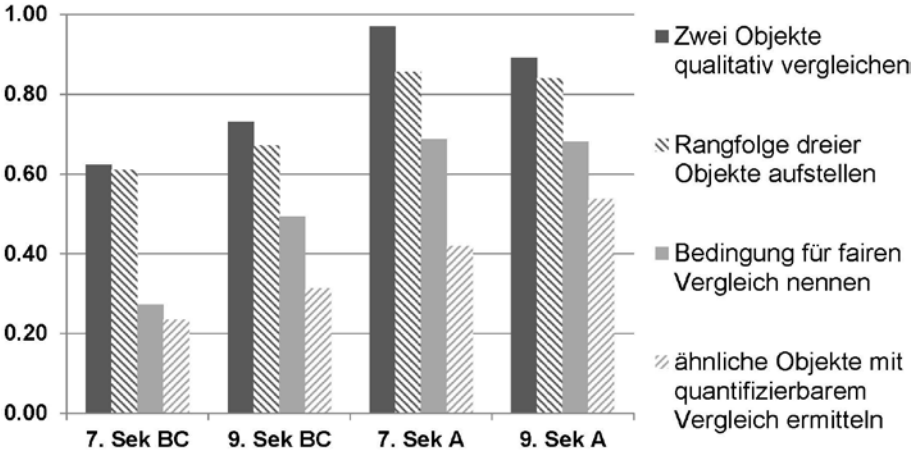

Abb. 6: Häufigkeiten der erreichten Qualitätsstandards summiert über alle drei Aufgaben zum Problemtyp «effektbasiertes Vergleichen» (Pilottest 2).

Die sehr niedrige Varianz könnte auf die geringe inhaltliche Validität der Tests zurückzuführen sein. Die getesteten Schülerinnen und Schülern haben wenig bis keine Erfahrung mit den Anforderungen des Tests (Problemtypen, experimentelle Strategien).

EAP/PV-Reliabilität	0.645
Varianz	0.464
Item-Separations-Reliabilität	0.983
Infit	0.93 – 1.09 ($\lvert T \rvert \leq 0.9$)
Outfit	0.88 – 1.14 ($\lvert T \rvert \leq 1.1$)
Trennschärfe	0.69 – 0.80

Tab. 4: Eindimensionale Rasch-Skalierung des ExKoNawi-Pilottests 2 (6 Items, $N = 331$) zu den Problemtypen «Untersuchen» und «Vergleichen».

Testsensitivität. Ungeachtet der geringen Varianz misst der Test signifikante Leistungsdifferenzen zwischen den Schulniveaus und den Jahrgangsstufen (Tab. 5). Vor allem auch im tiefen Leistungsbereich. Während der HarmoS-Experimentiertest zwischen den nicht gymnasialen Leistungszügen Sekundarschule B/C und A des 9 Schuljahrs keine signifikante Leistungsdifferenz von $d = 0.2$ misst (Tab. 3), differenziert der Pilottest 2 zwischen den beiden Sekundarschulen hochsignifikant mit einem Effekt von $d = 1.0$ (Tab. 5).

	Sek B/C	**Δ**	**Sek A**	**alle**
Jahrgangsstufe 7	**395** (79)	117 *** †	**512** (90)	
Δ	39 * †		25 *	
Jahrgangsstufe 9	**434** (101)	103 *** ($d = 1.0$)	**537** (78)	**500** (100)$^{+}$

Tab. 5: Eindimensionale Rasch-Skalierung des ExKoNawi-Pilottests 2 (6 Items, $N = 331$):
Mittelwerte der Fähigkeitsvariable in PISA-Metrik^{+}, Mittelwertvergleiche von Schulniveaus mit Signifikanzen († t-Test, ansonsten Mann-Whitney-Test).

Interpretierbarkeit: Gestützt auf der in der Vorvalidierung aufgezeigten „Ordinalität" der Progressionsmodelle (Abb. 5 und 6) lassen sich für die beiden Problemtypen «Untersuchen» und «Vergleichen» post hoc Niveaus festlegen. Mit Hilfe der Niveaus, die wir *for the sake of argument* mit der gebotenen Vorsicht annehmen, soll eingeschätzt werden, inwieweit die gemessenen Leistungsunterschiede (Tab. 5) interpretiert werden können.

Die Progressionen der Qualitätsstandards werden in der Wright-Map (Abb. 7) für jede Aufgabe (Item) in Form von vier Thurstonian thresholds (Schwellenwerten) abgebildet. Dabei markiert ein bestimmter Thurstonian threshold *x* auf der logit-Skala den minimalen Fähigkeitswert, ab welchem mit mehr als 50% Wahrscheinlichkeit *x* oder mehr Qualitätsstandards erreicht werden. Somit lassen sich Niveaus wie folgt definieren: Eine Person erreicht bei einem Problemtyp das Niveau *x*, wenn sie bei der Mehrheit der Aufgaben zum gleichen Problemtyp (in diesem Fall zwei Aufgaben) zu 50 oder mehr Prozent Wahrscheinlichkeit *x* oder mehr Qualitätsstandards erreicht. Für eine gegebene Anzahl Standards $x = 0, \ldots 4$ entspricht damit der jeweils mittlere von den drei Thurstonian thresholds der *Niveauschwelle*, die in der Wright-Map als dreistelliger Wert in PISA-Metrik dargestellt ist (Abb. 7). Die Verteilung der Niveauschwellen auf der logit-Skala bestätigt das bereits aus den Häufigkeitsanalysen der erreichten Qualitätsstandards (Abb. 5 und 6) ersichtliche Resultat, dass bei den Vergleichsaufgaben im Schnitt mehr Qualitätsstandards erreicht werden als bei den Untersuchungsaufgaben.

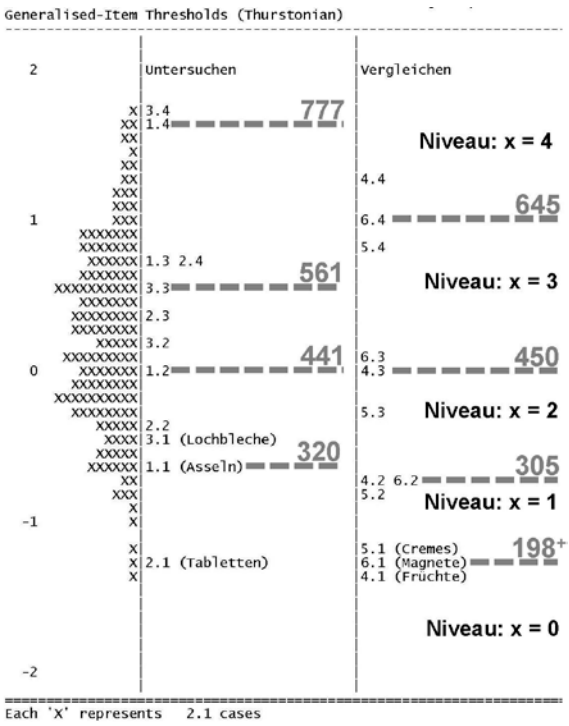

Abb. 7: Wright-Map des eindimensional Rasch-skalierten ExKoNawi-Piltotests 2 (6 Items, $N = 331$) mit Thurstonian thresholds:
post-hoc-Niveaustufen $x = 0, \ldots 4$ mit Niveauschwellen in PISA-Metrik^{+} (fett).

Betrachtet man in den Schulniveau- und Jahrgangskohorten der Tabelle 5 nur Schülerinnen und Schüler mit Fähigkeiten, die nicht mehr als eine halbe Standardabweichung vom Mittelwert abweichen, kann die mittlere Niveauerreichung der Kohorten verglichen werden. Dabei zeigt sich, dass bei beiden Problemtypen die Leistungsunterschiede zwischen den Schulniveaus rund einem interpretierbaren Niveau entsprechen. Während die Leistungsunterschiede zwischen den Jahrgangsstufen weniger als ein interpretierbares Niveau ausmachen (Tab. 6).

	fragegeleitetes Untersuchen		effektbasiertes Vergleichen	
	Sek B/C	Sek A	Sek B/C	Sek A
Jahrgangsstufe 7	**395 ± 39** (= ½ SD)		**512 ± 45** (= ½ SD)	
	Niveau 1	Niveau 2	Niveau 2	Niveau 3
Jahrgangsstufe 9	**434 ± 50** (= ½ SD)		**537 ± 39** (= ½ SD)	
	Niveau 1-2	Niveau 2-3	Niveau 2-3	Niveau 3

Tab. 6: Niveauerreichung von Schülerinnen und Schülern mit mittleren Fähigkeiten in Schulniveau- und Jahrgangsstufenkohorten beim ExKoNawi-Pilottest 2 (6 Items, *N* = 331).

6. Fazit

Die Auswertung des Pilottests zu den Problemtypen «Untersuchen» und «Vergleichen» belegt (wie auch schon die Auswertung des Pilottests 1 zum «Messen» und «Beobachten» [6]), dass mit einer problemtypenbasierten Kompetenzmodellierung ein interdisziplinäres Messinstrument mit guten Item-Fitwerten entwickelt werden kann, das auf tiefem Leistungsniveau signifikante und interpretierbare Leistungsdifferenzen misst. Dabei konnten im Vergleich zum HarmoS-Experimentiertest in allen diskutierten, die Validität beeinflussenden Aspekten – Itemabhängigkeiten, Einfluss kompetenzirrelevanter Anforderungen, Generalisierbarkeit und Interpretierbarkeit – Verbesserungen erzielt werden. Mit Hilfe der Unterscheidung von Problemtypen, der a priori-Modellierung von Kompetenzprogressionen sowie der in Bezug auf kompetenzrelevante und kompetenzirrelevante Anforderungen gut standardisierten Aufgabenentwicklung konnte vor allem die Interpretierbarkeit des Experimentiertests verbessert werden. Trotz aller Maßnahmen konnte jedoch noch keine ausreichend hohe Generalisierbarkeit erreicht werden. Die vordringlichste Aufgabe bleibt daher in der Verbesserung der Testreliabilität. Diese soll im Rahmen einer geplanten Validierung des ganzen Kompetenzmodells (u. a. die dimensionale Analyse des Strukturmodells hinsichtlich Problemtypen) mit Hilfe umfangreicherer Messinstrumente erreicht werden.

7. Literatur

[1] Labudde, P., Metzger, S. & Gut, C. (2009). Bildungsstandards: Validierung des Schweizer Kompetenzmodells. Konferenzbeitrag GDCP, Schwäbisch-Gmünd.

[2] Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., . . . Orpwood, G. (1997). Performance Assessment in IEA's Third International Mathematics and Science Study. Chestnut Hill: Boston College.

[3] Labudde, P. & Stebler, R. (1999). Lern- und Prüfungsaufgaben für den Physikunterricht. Erträge aus dem TIMSS-Experimentiertest. Unterricht Physik, 10(54), 23-31.

[4] Gut, C. (2012). Modellierung und Messung experimenteller Kompetenz - Analyse eines large-scale Experimentiertests. Berlin: Logos.

[5] Gut, C. & Labudde, P. (2013). HarmoS-Projekt: Validitätsanalyse des large-scale Experimentiertests. Konferenzbeitrag GDCP, Hannover.

[6] Metzger, S., Gut, C., Hild, P. & Tardent, J. (2014). Modelling and assessing experimental competence: an interdisciplinary progress model for hands-on assessments. Konferenzbeitrag ESERA, Nikosia.

[7] Gut, C., Hild, P., Metzger, S. & Tardent, J. (in Druck). Projekt ExKoNawi: Modell für hands-on Assessments experimenteller Kompetenzen. Konferenzbeitrag GDCP, München.

[8] Emden, M. (2011). Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I. Berlin: Logos.

[9] Gott, R. & Duggan, S. (2002). Problems with the assessment of performance in practical science: which way now. Cambridge Journal of Education, 32(2), 183-201.

[10] Lock, R. (1989). Assessment of practical skills. Part 1. The Relationships between component skills. Research in Science & Technological Education, 7(2), 221-233.

[11] Toh, K.-A. & Woolnough, B. E. (1994). Science process skills: are they generalisable? Research in Science & Technological Education, 12(1), 31-42.

[12] Nawrath, D., Maiseyenka, V. & Schecker, H. (2011). Experimentelle Kompetenz. Ein Modell für die Unterrichtspraxis. Physik in der Schule, 60(6), 42-48.

[13] Theyssen, H., Schecker, H., Gut, C., Hopf, M., Kuhn, J., Labudde, P., . . . Vogt, P. (2014). Modelling and assessing experimental competencies in physics. In C. Bruguière, A. Thibergien & P. Clément (Hrsg.), Topics and trends in current science education. 9th ESERA conference selected contributions (S. 321-338). Dordrecht: Springer.

[14] Germann, P. J. & Aram, R. J. (1996). Student Performances on the science processes of recording data, analysing data, conclusions, and providing evidance. Journal of Research in Science Teaching, 33(7), 773-798.

[15] Hammann, M., Phan, T. H. & Bayrhuber, H. (2007). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? Zeitschrift für Erziehungswissenschaft, 10(Sonderheft 8), 33-49.

[16] Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), Theorien in der biologiedidaktischen Forschung (S. 177-186). Berlin: Springer.

[17] Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., . . . Walpuski, M. (2012). Evaluation der Bildungsstandards - eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. Zeitschrift für Didaktik der Naturwissenschaften, 18, 261-291.
[18] Gott, R. & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. International Journal of Science Education, 18(7), 791-806.
[19] Millar, R., Gott, R., Lubben, F. & Duggan, S. (1996). Childern's performance of investigative tasks in science: a framework for considering progression. In M. Hughes (Hrsg.), Progression in learning (S. 82-108). Clevedon: Multilingual Matters.
[20] Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: an update. Journal of Research in Science Teaching, 33(10), 1045-1063.
[21] Qualter, A., Strang, J., Swatton, P. & Taylor, R. (1990). Exploration: a way of learning science. Oxford: Basil Blackwell.
[22] Millar, R. & Driver, R. (1987). Beyond processes. Studies in Science Education, 14, 33-62.
[23] Millar, R. (1991). A means to an end. The role of processes in science education. In B. Woolnough (Hrsg.), Practical science (S. 43-52). Milton Keynes: Open University Press.
[24] Finley, F. N. (1983). Science processes. Journal of Research in Science Teaching, 20(1), 47-54.
[25] Stevens, P. (1978). On the Nuffield philosophy of science. Journal of Philosophy of Education, 12, 99-111.
[26] Gott, R. & Welford, G. (1987). The assessment of observation in science. School Science Review, 69, 217-227.
[27] Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A., Schultz, S. E. & Wiley, E. W. (1997). On the development and scoring of classification and observation science performance assessments. CSE Technical Report (Vol. 458). National Center for Research on Evaluation, Standards, and Student Testing.
[28] Lubben, F. & Millar, R. (1996). Children's ideas about the reliability of experimental data. International Journal of Science Education, 18(8), 955-968.
[29] Masnick, A. M. & Klahr, D. (2003). Error matters: an initial exploration of elementary school children's understanding of experimental error. Journal of Cognition and Development, 4(1), 67-98.
[30] Munier, V., Merle, H. & Brehelin, D. (2011). Teaching scientific measurement and uncertainty in elementary school. International Journal of Science Education, 1-32, iFirst Article.
[31] Kuhn, D. & Phelps, E. (1982). The development of problem-solving strategies. Advances in Child Development and Behavior, 17, 1-44.
[32] Donnelly, J. F. (1987). Fifteen-year-old pupils' variable handling performance in the context of scientific investigations. Research in Science & Technological Education, 5(2), 135-147.
[33] Schauble, L., Glaser, R., Raghavan, K. & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. The Journal of the Learning Sciences, 1(2), 201-238.
[34] Kanari, Z. & Millar, R. (2004). Reasoning from data: how students collect and interpret data in science investigations. Journal of Research in Science Teaching, 41(7), 748-769.
[35] Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupil's skills in experimentation. Journal of Biological Education, 42(2), 66-72.
[36] Solano-Flores, G. (1994). A logical model for the development of science performance assessments. University of California, Santa Barbara.
[37] Erickson, G. (1994). Pupil's understanding of magnetism in a practical assessment context: the relationship between content, process and progression. In P. Fensham, G. Richard & R. White (Hrsg.), The content of science (S. 80-97). London: Falmer.
[38] Meyer, K. & Carlisle, R. W. (1996). Children as experimenters. International Journal of Science Education, 18(2), 231-248.
[39] Webb, N. M. & Schlackman, J. (2000). The dependability and interchangeability of assessment methods in science. Applied Measurement in Education, 13(3), 277-301.
[40] Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER ConQuest Version 2.0. Camberwell: ACER Press.