Rohwer, Götz

# Bemerkungen zu einem Testverfahren für Lernfortschritte

*Journal for educational research online 7 (2015) 2, S. 147-156*

Quellenangabe/ Reference:
Rohwer, Götz: Bemerkungen zu einem Testverfahren für Lernfortschritte - In: Journal for educational research online 7 (2015) 2, S. 147-156 - URN: urn:nbn:de:0111-pedocs-114943 - DOI: 10.25656/01:11494

*https://nbn-resolving.org/urn:nbn:de:0111-pedocs-114943*
*https://doi.org/10.25656/01:11494*

Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 7 (2015), No. 2, 147–156


Götz Rohwer

# Bemerkungen zu einem Testverfahren für Lernfortschritte

## Zusammenfassung

*Der Beitrag diskutiert ein Testverfahren für Lernfortschritte, das von Strathmann und Klauer (2010) vorgeschlagen wurde. Das Verfahren beruht darauf, dass man Mengen äquivalenter Testaufgaben definieren kann. Für die Interpretation werden eine deterministische, eine probabilistische und eine generische Konzeption der zu erfassenden Kompetenz unterschieden. Es wird gezeigt, dass sich die generische Konzeption am besten eignet, um zu begründen, dass „der gleiche Test" wiederholt werden kann, um Kompetenzveränderungen zu erfassen.*

## Schlagworte

*Lernverlaufsdiagnostik; Kriteriumsorientierte Tests; Personenorientierte Verfahren; Item-Sampling*

# Remarks on a test procedure for long-term learning progress

## Abstract

*The paper discusses a test procedure for the assessment of long-term learning progress proposed by Strathmann and Klauer (2010). The procedure is based on the assumption that one can define sets of equivalent tasks. For interpreting the procedure, the paper distinguishes a deterministic, a probabilistic, and a generic conception of competencies. It is shown that the generic conception allows one to justify the view that, in order to asses changes of competencies, "the same test" can be repeated.*

## Keywords

*Measuring the development of learning; Criterion referenced tests; Person-oriented procedures; Item-sampling*

Prof. Dr. em. Götz Rohwer, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaft, Universitätsstr. 150, 44780 Bochum, Deutschland
E-Mail: goetz.rohwer@rub.de

## 1. Vorbemerkungen

Ich beziehe mich auf Beiträge von Strathmann und Klauer (2010), Klauer (2011) und Strathmann, Klauer und Greisbach (2010), in denen ein Testverfahren zur Lernfortschrittsmessung vorgestellt wird. Ich versuche, die formale Gestalt des Verfahrens (also unter Absehung von allen Fragen, die die Auswahl von Items für eine bestimmte Anwendung betreffen) möglichst genau und ausführlich zu beschreiben und auf dieser Grundlage einige der von den Autoren angesprochenen Probleme zu diskutieren. Ich möchte zeigen, dass sich einige der Schwierigkeiten bei der Erfassung von Lernfortschritten vermeiden lassen, wenn man Kompetenzen nicht auf einzelne Aufgaben, sondern generisch auf „Arten von Aufgaben" bezieht. Dies betrifft insbesondere die Idee, dass Lernfortschrittsmessung erfordert, dass „der gleiche Test" wiederholt werden kann.

## 2. Die Durchführung der Tests

Als Kontext wird ein Lernprozess vorausgesetzt, durch den Personen, die zu einer als gleichbleibend vorausgesetzten Gesamtheit (beispielsweise eine Schulklasse) gehören, bestimmte Kompetenzen entwickeln können. Die Personen werden durch $i = 1, \ldots, N$ indiziert. Es wird angenommen, dass zur Bezugnahme auf die zu erfassenden Kompetenzen $K$ Mengen von Aufgaben definiert werden können. Jede Menge besteht aus einer endlichen Anzahl von Aufgaben:

$$M_k := \{(k,r) | r = 1, \ldots; R_k\} \; (k = 1, \ldots, K)$$

Jede Aufgabe $(k, r)$ kann durch die Personen in der vorausgesetzten Personengesamtheit bearbeitet werden, und als Ergebnis kann festgestellt werden, ob sie richtig oder nicht richtig gelöst wurde.

Um Lernfortschritte zu erfassen, werden die Tests in Zeitstellen $t = 1, \ldots, T$ wiederholt. Für jede Person $i$ und Zeitstelle $t$ wird ein Test $T$ $(i, t)$ durchgeführt. Der Test verläuft folgendermaßen: Zunächst wird eine Testvorlage erstellt, die die Aufgaben enthält, die die Person $i$ in der Zeitstelle $t$ bearbeiten soll. Die Erstellung erfolgt dadurch, dass aus den Aufgabenmengen Teilmengen zufällig (mit vorab definierten Inklusionswahrscheinlichkeiten) ausgewählt werden. Aus der Aufgabenmenge $M_k$ werden $m_k$ Aufgaben ausgewählt, so dass jede Testvorlage aus $m := \sum_k m_k$ Aufgaben besteht. Werden dann diese Aufgaben bearbeitet, erhält man einen Ergebnisvektor $x_i(t) := (x_{i1}(t), \ldots x_{im}(t))$, wobei die Werte $x_{ij}(t)$ entweder 0 oder 1 sind.

## 3. Quantifizierung der Kompetenzen

Es bleibt zu überlegen, wie die durch die Tests erfassten Kompetenzen quantitativ repräsentiert werden sollen. Ich nehme an, dass die Quantifizierung eine kriteriumsorientierte Interpretation erlauben soll: Der quantitative Kompetenzwert soll zeigen, wie gut eine Person in einer gegebenen Zeitstelle Aufgaben der Art, wie sie in den Aufgabenmengen definiert worden sind, lösen kann. Es erscheint dann sinnvoll, sich zunächst auf einzelne Aufgabenmengen zu beziehen. Diese Aufgabenmengen bestehen jedoch, wie die Autoren an einem Beispiel ausführen, aus Aufgaben unterschiedlicher Art. Ich nehme also an, dass $M_k$ aus Teilmengen $M_{kl}\,(l = l, \ldots, L_k)$ besteht.

Eine grundlegende Annahme besteht nun darin, dass die Aufgaben innerhalb dieser Teilmengen äquivalent sind, das soll heißen: Jede Aufgabe in $M_{kl}$ eignet sich gleichermaßen, um die Kompetenz einer Person zum Lösen von Aufgaben der durch $M_{kl}$ exemplifizierten Art zu erfassen.[1] Diese Annahme liefert zunächst eine Begründung dafür, die Kompetenz einer Person $i$ zum Lösen von Aufgaben der durch $M_{kl}$ exemplifizierten Art in der Zeitstelle $t$ durch

$$q_{ikl}\,(t) := \text{Anteil der Aufgaben in } M_{kl}\text{, den } i \text{ in } t \text{ lösen kann}$$

zu definieren (wie der Ausdruck „Anteil“ genauer zu verstehen ist, bespreche ich im nächsten Abschnitt). Dann kann die Fähigkeit zum Lösen von Aufgaben aus $M_k$ durch

$$q_{ik}(t) := \sum_{l=1}^{L_k} w_{kl} q_{ikl}\,(t)$$

quantifiziert werden, wobei $w_{kl}$ Gewichte sind ($\sum_l w_{kl} = 1$). Diese Gewichte werden als ein fester Bestandteil des Testverfahrens fixiert. Schließlich werden die gruppenspezifischen Kompetenzen durch

$$q_i\,(t) := \sum_{k=1}^{K} \frac{m_k}{m} q_{ik}\,(t)$$

zu einem Indikator für die Gesamtkompetenz zusammengefasst. Dabei bestimmen die Auswahlsätze $m_k$/m, mit welchem Gewicht die Aufgabenmengen $M_k$ zum Indikator für die Gesamtkompetenz beitragen.

Soweit handelt es sich um eine Definition der zu ermittelnden Kompetenz. Aus der Annahme der Äquivalenz der Aufgaben in den Teilgruppen $M_{kl}$ folgt indessen auch, dass man zum Schätzen von $q_{ikl}(t)$ den Anteil der richtig gelösten Aufgaben in einer einfachen Zufallsauswahl von Aufgaben aus $M_{kl}$ verwenden kann. Für die

1 Klauer (2011) spricht mit einer ähnlichen Intention davon, dass die Aufgaben „homogen“ sein sollten. Er verbindet diesen Begriff jedoch mit Überlegungen zu Aufgabenschwierigkeiten, die ich vermeiden möchte. Das wird am Ende von Abschnitt 5 näher begründet.

praktische Durchführung kann man sich auch sogleich auf die Aufgabenmengen $M_k$ beziehen. Formal wird dann eine geschichtete Zufallsauswahl verwendet, wobei die Anzahl der aus der Teilgruppe $M_{kl}$ mit gleichen Wahrscheinlichkeiten auszuwählenden Aufgaben durch $m_{kl} := m_k w_{kl}$ definiert wird. Dann ist der Anteil der gelösten Aufgaben, also $s_{ik}(t)/m_k$ ein sinnvoller Schätzwert für $q_{ik}(t)$.[2] Schließlich liefert

$$s_i^*(t) := \sum_{k=1}^{K} \frac{m_k}{m} s_{ik}(t)$$

einen Schätzwert für $q_i(t)$.

## 4. Interpretation der Kompetenzen

Ausgangspunkt für die im vorangegangenen Abschnitt eingeführte Quantifizierung ist die Fähigkeit, Aufgaben in den Teilgruppen $M_{kl}$ lösen zu können. Die zu schätzende Größe $q_{ikl}(t)$ wurde als Anteil der Aufgaben in $M_{kl}$, den die Person $i$ in der Zeitstelle $t$ lösen kann, definiert. Genauere Formulierungen hängen von der theoretischen Konzeption der zu erfassenden Kompetenzen ab. Ich unterscheide drei Möglichkeiten. Die ersten beiden knüpfen an herkömmliche Interpretationen von Binomialtests an, die dritte wird durch den hier diskutierten Lernfortschrittstest nahe gelegt.

I-1) Deterministische Konzeption. In diesem Fall wird für jede Aufgabe $(k, r)$ angenommen: Entweder kann die Person $i$ die Aufgabe in der Zeitstelle $t$ lösen, dann ist $\pi_{i(k,r)}(t) = 1$, oder sie kann sie nicht lösen, dann ist $\pi_{i(k,r)}(t) = 0$. Dieser Ansatz erlaubt eine einfache Explikation:

$$q_{ikl}(t) := \sum_{(k,r)\in M_{kl}} \pi_{i,(k,r)}(t)/R_{kl} \qquad (1)$$

wobei $R_{kl}$ die Anzahl der Aufgaben in $M_{kl}$ ist. Der Ansatz ist jedoch kaum damit vereinbar, dass das Lösen von Aufgaben eine Tätigkeit ist, bei der es situationsabhängig zu Fehlern kommen kann, insbesondere während der Zeit, in dem die Fähigkeit noch erlernt werden muss. Hier wird auch relevant, dass es zwei Aspekte des Lernprozesses gibt (Klauer 2011, S. 219): Einerseits wird gelernt, neue Arten von Aufgaben zu lösen, andererseits wird gelernt, Aufgaben, die man schon ansatzweise lösen kann, besser und sicherer zu lösen. Entsprechend der Konzeption der

2 Wegen der Äquivalenz der Aufgaben in $M_{kl}$ kann $q_{ikl}(t)$ durch $s_{ikl}(t)/m_{kl}$ (den Anteil der richtig gelösten Aufgaben in der Teilgruppe) geschätzt werden. Also:

$$\frac{s_{ik}(t)}{m_k} = \frac{\sum_l s_{ikl}(t)}{m_k} \approx \frac{\sum_l m_{kl} q_{ikl}(t)}{m_k} = \sum_l w_{kl} q_{ikl}(t) = q_{ik}(t)$$

Teilgruppen $M_{kl}$ kommt hauptsächlich der zweite Aspekt in Betracht; aber gerade hierfür erscheint die deterministische Konzeption wenig plausibel.

I-2) Probabilistische Konzeption. Bei diesem Ansatz ist $\pi_{i(k,r)}(t)$ eine Größe zwischen 0 und 1, die als Wahrscheinlichkeit, mit der die Person $i$ die Aufgabe $(k, r)$ in der Zeitstelle $t$ lösen kann, interpretiert wird. Bei dieser Interpretation liefert (1) zwar keinen bestimmten Anteil, kann aber immer noch als ein Erwartungswert für den Anteil der Aufgaben, den die Person lösen kann, betrachtet werden (van der Linden, 1979). Diese Konzeption scheint besser dazu zu passen, dass das Lösen von Aufgaben gerade in der Lernphase eine unsichere Tätigkeit ist. Dem steht jedoch der Nachteil gegenüber, dass den postulierten Wahrscheinlichkeiten keine operationale Bedeutung gegeben werden kann.

Im Unterschied zu den Wahrscheinlichkeiten für die Auswahl von Aufgaben handelt es sich um theoretische Fiktionen, die empirisch nicht ermittelt werden können. Außerdem kann eingewendet werden, dass die Bearbeitung einer Testaufgabe eine menschliche Aktivität ist, die sich grundsätzlich von der Aktivierung eines Zufallsgenerators unterscheidet.

Eine weniger problematische Variante der probabilistischen Konzeption entsteht, wenn man annimmt, dass die Lösungswahrscheinlichkeiten für alle Aufgaben in $M_{kl}$ den gleichen Wert $\pi_{ikl}(t)$ haben. Dann kann man die Definition $q_{ikl}(t) := \pi_{ikl}(t)$ verwenden, und diese Größen können durch $s_{ikl}(t)/m_{kl}$ geschätzt werden.

I-3) Generische Konzeption. Den bisher skizzierten Konzeptionen ist gemeinsam, dass sie sich auf Fähigkeiten zum Lösen einzelner Aufgaben beziehen. Eine Alternative besteht darin, sich auf „Arten von Aufgaben“ zu beziehen, die durch die Teilmengen $M_{kl}$ exemplifiziert werden. Die zu ermittelnde Kompetenz bezieht sich dann von vornherein auf alle Aufgaben in $M_{kl}$, ohne dabei auf einzelne Aufgaben Bezug zu nehmen; sie kann also als Anteil an allen Aufgaben aus $M_{kl}$ definiert werden, den eine Person in einer bestimmten Zeitstelle lösen kann.

Zwar entspricht dieser Definition zunächst nur ein hypothetisches Gedankenexperiment, das darin besteht, dass eine Person alle Aufgaben in $M_{kl}$ bearbeitet. Aber dieses Gedankenexperiment kann bis zu einem gewissen Grad praktisch umgesetzt werden, indem man die Person einen Teil der Aufgaben bearbeiten lässt. Die dafür verwendeten Aufgaben sollten möglichst repräsentativ sein und werden deshalb zufällig ausgewählt. Die auf diese Weise operationalisierte Idee einer generischen Kompetenz setzt nicht voraus, dass die Aufgaben in $M_{kl}$ äquivalent sind. Die Äquivalenz wird dagegen relevant, um einen Test zu konzipieren, der wiederholt werden kann, und zwar so, dass man sagen kann, dass *der gleiche Test* wiederholt wird.

## 5. Personenorientierte Auswertungen

Die Tests des hier besprochenen Verfahrens sollen dazu dienen, Lernfortschritte bei den teilnehmenden Personen zu erfassen und zu beschreiben. Die Tests müssen also kriteriumsorientiert interpretierbar sein, wobei das Kriterium darin besteht, wie gut ein vorgegebenes Lernziel erreicht wird. Damit Lernfortschritte personenorientiert beschrieben werden können, muss eine weitere Bedingung erfüllt sein: Die Quantifizierung der Testergebnisse muss verteilungsunabhängig erfolgen, womit gemeint ist: Der Indikator für die Kompetenz einer Person darf nur von ihren Testergebnissen abhängen, nicht von den Testergebnissen anderer Personen. Diese Bedingung ist offenbar erfüllt, wenn man den Indikator $q_i(t)$ bzw. seinen Schätzwert $s_i^*(t)$ verwendet; und somit kann die zeitliche Folge dieser Indikatoren als eine Beschreibung der Lernfortschritte der Person $i$ interpretiert werden.

Damit personenorientierte Auswertungen möglich werden, ist es nicht erforderlich, dass jede Person eine eigene Auswahl von Testaufgaben erhält. Es würde genügen, in jeder Zeitstelle eine zufällige Auswahl der Aufgaben vorzunehmen, die dann von allen Personen bearbeitet werden. Die Bedingung der Verteilungsunabhängigkeit wäre aber z. B. verletzt, wenn man bei der Berechnung der Kompetenzindikatoren Aufgabenschwierigkeiten verwendet, die aus der Verteilung der Testergebnisse in der jeweiligen Personengesamtheit ermittelt werden. Das ist insbesondere dann der Fall, wenn zur Skalierung der Testergebnisse ein *Item-Response-Theory*-Modell (IRT) verwendet wird. Um ein IRT-Modell für den hier besprochenen Test einzusetzen, könnte man die aus den Teilgruppen $M_{kl}$ ausgewählten Items als komplexe Items betrachten und die Testergebnisse bzgl. dieser Items durch $s_{ikl}(t)$ erfassen. Dann könnte man ein *Partial-Credit*-Modell (Masters, 1982) verwenden, um die Testergebnisse zu skalieren und den Personen zurechenbare Kompetenzwerte zu berechnen. Diese Kompetenzwerte wären dann allerdings von der Verteilung der Testergebnisse in der Personengesamtheit abhängig und könnten nicht als Grundlage für personenorientierte Auswertungen dienen.

Bei dem oben im Abschnitt 3 besprochenen Verfahren wird die Verteilungsunabhängigkeit dadurch erreicht, dass die Gewichte $w_{kl}$ und $m_k$ unabhängig von den Testergebnissen definiert werden. Sie müssen auch zeitunabhängig definiert werden, damit die erfassten Veränderungen nicht von zeitlichen Veränderungen des Skalierungsverfahrens abhängig werden. An dieser Stelle kann auch eine Überlegung erwähnt werden, die von Strathmann und Klauer (2010) so formuliert wird:

> [...] dass Veränderungen nur dann zu erfassen sind, wenn die Tests stets das Gleiche messen, sich also auch durch gleiche Schwierigkeit auszeichnen. Wenn aber zwischenzeitlich Lernen stattfindet und in den Tests nachweisbar sein soll, so müssen die Tests notwendigerweise leichter werden, ihre Mittelwerte also ansteigen. Es wird darauf ankommen, diesem Dilemma in geeigneter Form Rechnung zu tragen. (Strathmann und Klauer, 2010, S. 113)

Aus meiner Sicht handelt es sich nur scheinbar um ein Dilemma. Denn wenn man die Schwierigkeit eines Tests empirisch durch den durchschnittlichen Anteil der nicht gelösten Aufgaben bestimmt, kann man nicht fordern, dass die Tests im Zeitablauf gleich schwierig sein sollen. Um Veränderungen zu erfassen, ist es vielmehr erforderlich, dass *der gleiche Test* wiederholt wird; das Messverfahren darf sich im Zeitablauf nicht verändern. Damit man sagen kann, dass der gleiche Test wiederholt wird, ist allerdings nicht nur die Verteilungs- und Zeitunabhängigkeit des Skalierungsverfahrens relevant. Gleichermaßen wichtig ist, dass die Aufgaben in den Teilgruppen $M_{kl}$ äquivalent sind, d. h. sich gleichermaßen eignen, um Kompetenzen bzgl. der Aufgaben dieser Art zu ermitteln.

Könnte dagegen eingewendet werden, dass man doch zufällig „unterschiedlich schwierige" Aufgaben ziehen könnte? Diese Frage wird auch von Klauer (2011, S. 217) diskutiert. Die Antwort hängt davon ab, welche der oben unterschiedenen Auffassungen der zu ermittelnden Kompetenz zugrunde gelegt wird. Konzeptionen, die sich auf einzelne Aufgaben beziehen, erlauben folgende Definition:

> Zwei Aufgaben $(k, r)$ und $(k', r')$ sind für eine Person $i$ in der Zeitstelle $t$ *unterschiedlich schwer*, wenn $\pi_{i,(k,r)}(t) \neq \pi_{i,(k',r')}(t)$ ist.

Sowohl bei der deterministischen als auch bei der probabilistischen Konzeption kann dies als Möglichkeit vorgestellt werden; aber die Vorstellung ist in beiden Fällen empirisch kaum gehaltvoll: Bei der deterministischen Konzeption kann man nur hinterher sagen, dass eine Aufgabe, die die Person lösen konnte, offenbar leicht für sie war und eine Aufgabe, die sie nicht lösen konnte, offenbar schwer; und bei der probabilistischen Konzeption kann man nicht einmal dies sagen. Geht man dagegen von dem Spezialfall der probabilistischen Konzeption aus, bei dem alle Aufgaben aus $M_{kl}$ mit der gleichen Wahrscheinlichkeit gelöst werden können, werden die Aufgaben von vornherein als „gleich schwierig" (im Sinne der o. a. Definition) angenommen.

Anders verhält es sich bei der generischen Konzeption. Da sich bei dieser Konzeption die zu erfassende Kompetenz nicht auf einzelne Aufgaben beziehen lässt, kommt der Vorstellung, dass einzelne Aufgaben „unterschiedlich schwierig" sein können, keine relevante Bedeutung zu. Es ist im Rahmen dieser Konzeption weder erforderlich noch ohne Weiteres überhaupt möglich, unterschiedliche Aufgabenschwierigkeiten zu definieren. Auch aus diesem Grund erscheint mir die generische Konzeption zur Explikation des hier diskutierten Testverfahrens am Besten geeignet.

## 6. Genauigkeit der Kompetenzschätzungen

Kann man etwas über die Reliabilität des Testverfahrens sagen? Wie insbesondere von Klauer (2011) ausführlich diskutiert wird, sind die üblichen Methoden kaum sinnvoll anwendbar. Einerseits sind offenbar alle Methoden problematisch, die an-

nehmen, dass ein Test wiederholt werden könnte, ohne dass sich zwischenzeitlich die zu erfassende Kompetenz verändert hat. Andererseits können Indikatoren für die interne Konsistenz eines Tests (wie z. B. Varianten von Cronbachs Alpha) zwar mit den Daten aus einer Zeitstelle berechnet werden; diese Indikatoren beziehen sich jedoch stets auf die Gesamtheit der am Test beteiligten Personen und sind infolgedessen für personenorientierte Auswertungen kaum relevant.

Wie von Strathmann und Klauer (2010) betont wird, zielt das von ihnen vorgeschlagene Testverfahren auf personenorientierte Auswertungen, so dass es auch wichtig ist, wie genau Aussagen über einzelne Personen getroffen werden können. Die Fragestellung lautet dann: Wie genau die Größen $q_i(t)$, die man schätzen möchte, durch $s_i^*(t)$ geschätzt werden können. Die folgenden Überlegungen beziehen sich auf diese Frage. Um die Überlegungen zu vereinfachen, nehme ich an, dass die Aufgaben zufällig mit Zurücklegen gezogen werden. Dann kann man sich auf Zufallsvariablen $X_{ikl}(t)$ beziehen, die ihre Werte dadurch annehmen, dass zufällig eine Aufgabe aus $M_{kl}$ gezogen und dann durch die Person $i$ in der Zeitstelle $t$ bearbeitet wird; sie nimmt den Wert 1 an, wenn die Lösung richtig ist, und andernfalls den Wert 0. Die Zufallsvariablen $S_{ikl}(t)$, die die Anzahl der richtigen Lösungen bei der Bearbeitung von $m_{kl}$ Aufgaben aus $M_{kl}$ angeben, ergeben sich dann als Summe von $m_{kl}$ unabhängigen $X_{ikl}(t)$-Variablen. Die weiteren Überlegungen hängen davon ab, welche der oben unterschiedenen Konzeptionen der zu erfassenden Kompetenz zugrunde gelegt wird.

I-1) Wenn man die deterministische Konzeption voraussetzt, ist die Wahrscheinlichkeit, dass $X_{ikl}(t)$ den Wert 1 annimmt, gleich $q_{ikl}(t)$, und infolgedessen hat $S_{ikl}(t)$ eine Binomialverteilung mit dem Mittelwert $m_{kl}\,q_{ikl}(t)$ und der Varianz $m_{kl}\,q_{ikl}(t)\,(1\text{-}q_{ikl}(t))$. Jetzt kann man die Variable $S_{ik}(t) := \sum_l s_{ikl}(t)$ betrachten, die die Anzahl der richtigen Antworten für alle aus $M_k$ gezogenen Aufgaben erfasst. Sie hat den Mittelwert $\sum_l m_{kl}\,q_{ikl}(t)$ und die Varianz $\sum_l m_{kl}\,q_{ikl}(t)\,(1\text{-}q_{ikl}(t))$. Und schließlich kann man die Variable $S_i^*(t) := \sum_k S_{ik}(t)/\mathrm{m}$ betrachten, die den Anteil der richtigen Antworten bei allen aus $M$ gezogenen Aufgaben erfasst. Sie hat den Mittelwert

$$\sum_{k=1}^{K} \frac{m_k}{m}\, q_{ik}(t)$$

und die Varianz

$$\frac{1}{m} = \sum_{k=1}^{K} \frac{m_k}{m} \sum_{l=1}^{L_k} w_{kl} q_{ikl}(t)\,(1-\,q_{ikl}\,(t)) \qquad (2)$$

Ein Schätzwert für diese Varianz kann offenbar aus den Daten gewonnen werden.

I-2) Jetzt beziehe ich mich auf die probabilistische Konzeption. Setzt man zunächst den Spezialfall voraus, bei dem die unterstellten Lösungswahrscheinlichkeiten $\pi_{i,(kl)}(t)$ für alle Aufgaben in $M_{kl}$ den gleichen Wert haben, können die Überlegungen, die eben für die deterministische Konzeption angestellt wurden,

ohne Änderung übernommen werden. Anders verhält es sich jedoch, wenn man zulässt, dass sich diese Wahrscheinlichkeiten zwischen den Aufgaben in $M_{kl}$ unterscheiden können. Dann hat $S_{ikl}(t)$ keine Binomialverteilung (wie auch bereits von van der Linden (1979) festgestellt wurde), und es ist nicht mehr auf einfache Weise möglich, die Varianz dieser Zufallsvariablen und mithin von $S_i^*(t)$ zu bestimmen.

I-3) Wie verhält es sich schließlich bei der generischen Konzeption? Eine Möglichkeit besteht darin, auch in diesem Fall die Varianzberechnung (2) zu verwenden. Das könnte auf folgende Weise motiviert werden: Der zu schätzende Indikator ist durch ein Gedankenexperiment definiert, bei dem eine Person alle Aufgaben aus den Mengen $M_{kl}$ bearbeitet. Stellt man sich nun vor, dass dieses Gedankenexperiment realisiert worden ist, dann steht für alle Aufgaben fest, ob sie von der Person gelöst worden sind, und man kann wie im Fall (I-1) die tatsächlich ermittelten Werte $x_i(t)$ als eine Stichprobe aus einer Menge von Werten ansehen, die für jede Aufgabe aus $M$ angeben, ob sie von der Person gelöst worden ist.

Als Alternative oder Ergänzung kann man ein *Bootstrap-* oder *Jackknife*-Verfahren verwenden. Besonders einfach ist hier das *Jackknife*-Verfahren (Efron & Tibshirani 1993, S. 140 ff). Man lässt jeweils eine Aufgabe aus und berechnet mit den Ergebnissen für die verbleibenden $m - 1$ Aufgaben Schätzwerte $s_{i\,j}^*(t)$ (für $j = 1, \ldots, m$, wobei $j$ der Index der ausgelassenen Aufgabe ist).

Bezeichnet $s_{i,\cdot}^*(t)$ den Mittelwert dieser Schätzwerte, liefert

$$\frac{m-1}{m}\sum_{j=1}^{m}\left(s_{i,j}^*(t)-s_{i,\cdot}^*(t)\right)^2 \qquad (3)$$

einen Schätzwert für die Varianz von $s_i^*(t)$.

## 7. Ergebnis

Der Beitrag diskutiert ein von Strathmann und Klauer (2010) vorgeschlagenes Verfahren zur Lernfortschrittsmessung, das es erlauben soll, Lernfortschritte bei einzelnen Personen zu ermitteln. Wie bei jedem Verfahren, mit dem Kompetenzveränderungen erfasst werden sollen, besteht ein zentrales Problem darin, ein gegenüber den zu erfassenden Veränderungen relativ invariantes Messverfahren zu begründen. Das hier diskutierte Verfahren zeichnet sich dadurch aus, dass bei Testwiederholungen zwar die Aufgaben verändert werden, aber dennoch behauptet werden kann, dass „der gleiche Test“ wiederholt wird. Dies wird in technischer Hinsicht dadurch erreicht, dass Testaufgaben aus vorgegebenen und gleichbleibenden Aufgabenmengen zufällig ausgewählt werden. Aber wichtig ist außerdem, dass der Zusammenhang zwischen den *möglichen* Aufgaben und der zu erfassenden Kompetenz expliziert werden kann. Der Beitrag zeigt, dass dieses Problem dadurch gelöst werden kann, dass der Kompetenzbegriff nicht auf einzelne Aufgaben, sondern generisch auf *Arten von Aufgaben* bezogen wird.

Diese generische Konzeption ermöglicht es auch, ohne einen Begriff der „Schwierigkeit“ von Aufgaben auszukommen. Das ist wichtig, denn die übliche Idee, Aufgabenschwierigkeiten durch durchschnittliche Kompetenzen einer Bezugsgruppe zu definieren, ist bei der Erfassung von Kompetenzveränderungen offenbar nicht sinnvoll.

Aus der Zielsetzung, Kompetenzveränderungen zu erfassen, folgt auch, dass die herkömmlichen Methoden zur Ermittlung der Reliabilität bei dem hier diskutierten Verfahren nicht verwendet werden können. Der Beitrag schlägt deshalb vor, sich stattdessen auf die Frage zu beziehen, mit welcher statistischen Genauigkeit die generisch konzipierten Kompetenzen für einzelne Personen geschätzt werden können. Diese Fragestellung erscheint insbesondere für das hier diskutierte Verfahren angemessen, bei dem personenorientiert interpretierbare Ergebnisse angestrebt werden.

## Literatur

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Klauer, K. J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten. *Empirische Sonderpädagogik, 3*(3), 207–224.

Masters, G. N. (1982). A Rasch Model for partial credit scoring. *Psychometrika, 47*, 149–173.

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*(2), 111–122.

Strathmann, A., Klauer, K. J. & Greisbach, M. (2010). Lernverlaufsdiagnostik – Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule. *Empirische Sonderpädagogik, 2*(1), 64–77.

van der Linden, W. J. (1979). Binomial test models and item difficulty. *Applied Psychological Measurement, 3*(3), 401–411.