Prüfungen in Naturwissenschaft und Informatik konzipieren aus der Sicht eines Experiments

Prof. Dr. Thomas Lehmann

HAW Hamburg, 2021-11-11

Einleitung

Prüfen ist ein wesentlicher Eckpunkt im Dreieck des Constructive Alignment [Biggs, 2007]. Das Learning Outcome, die Lehrveranstaltung und die zugehörige Prüfung sollen konsistent aufeinander abgestimmt sein. Dabei werden im Rahmen der Hochschullehre Prüfungen zum einen als Basis für ein qualifiziertes Feedback über den Lernfortschritt verwendet, zum anderen summativ für eine abschließende Bewertung der erworbenen Kompetenz.

Auch wenn man die Ziele eines Moduls auf unterschiedliche Weise beschreiben kann, so bewährt sich gerade mit Blick auf die Prüfung eine Formulierung in Form eines kompetenzorientierten Learning Outcome [DAAD, 2008/Kennedy, 2007]. Nach Kennedy in [Kennedy, 2007] soll ein Learning Outcome beschreiben, was die Studierenden „[...] are expected to demonstrate [...]“, also was die Studierenden prinzipiell und somit auch in einer Prüfungssituation in der Lage sind zu demonstrieren. Der Gedanke der Demonstration der Kompetenz bestimmt das Ziel der Prüfungskonstruktion. Die Prüfung soll ermitteln, ob und in welcher Ausprägung die Studierenden am Ende des Lernprozesses in der Lage sind, die geforderte Kompetenz zu demonstrieren; sei es einen Entwurf durchzuführen, eine Software zu erstellen oder eine wissenschaftliche Ausarbeitung zu schreiben. Die Prüfung muss dabei passend und stimmig zum Learning Outcome und zum Lernangebot sein. Weiterhin soll die Prüfung fair und transparent bezüglich der Bewertung sein, insgesamt valide mit einem hohen Qualitätsanspruch.

Im ingenieurwissenschaftlichen Bereich ist das Experiment eine typische Untersuchungsmethode, sei es zur Stützung von Hypothesen, zur systematischen Erfassung von Systemparametern oder zur Prüfung von Qualitätsmerkmalen (vgl. beispielsweise [Siebertz, 2017]). In anderen Fachkulturen ist die Untersuchungsmethode Experiment natürlich ebenso verbreitet.

Im Folgenden soll die Ähnlichkeit zwischen einem Experiment und Prüfungen aufgezeigt werden. Aus dem Vergleich können Ansätze für die Entwicklung der eigenen Prüfungen gezogen werden.

Das Experiment

Ein wissenschaftliches Experiment dient im Allgemeinen dazu, eine Hypothese zu falsifizieren oder dient der quantitativen Bestimmung von Systemgrößen mittels Messung. Die

ermittelten Systemgrößen sind dann eigentlich Koeffizienten in einem Modell des Systems. Innerhalb des Experiments kommen die dafür fachlich erforderlichen Methoden zum Einsatz.

Eine Prüfung soll das Vorhandensein einer Kompetenz oder den Erfüllungsgrad der Kompetenz respektive von Teilkompetenzen bei den zu Prüfenden bestimmen. Im Vergleich bedeutet das, man entwickelt ein Modell für die zu erreichende Kompetenz und versucht dann durch Beobachtung die Ausprägung der Kompetenz bei dem jeweils beobachteten System „zu Prüfende“ zu bestimmen. Die Ermittlung der Teilausprägung einer Kompetenz kann mit der Bestimmung von Systemparameter verglichen werden.

Vergleichbar der Messung im Experiment muss eine Kompetenz mit geeigneten Mitteln erfasst und bewertet werden. In beiden Fällen wird eine Messung einer oder mehrerer Größen im betrachteten System durchgeführt, die eine Aussage über das System ermöglichen; die Zielsetzung ist also vergleichbar. Im Folgenden soll deshalb das Experiment mit seinen verschiedenen Aspekten als Leitbild für die Konzipierung von Prüfungen verwendet werden.

Vorgehen bei der Konzipierung eines Experiments

Prinzipiell erfolgt die Konzipierung und Durchführung eines Experiments in den folgenden Schritten:

Ziel definieren: Welcher Aspekt soll in dem Experiment untersucht werden? Welche Fragestellung soll durch das Experiment geklärt werden? Welche Größe im betrachteten System soll quantitativ bestimmt werden?

Design der Versuchsanordnung: Das Design des Experiments muss den oder die zu untersuchenden Aspekte (die abhängige Variablen) durch die Wahl der Messmethoden klar sichtbar machen. Weiterhin müssen die Randbedingungen innerhalb der Messmethode definiert (feste Parameter) und gegebenenfalls die Variation der freien Variablen für jeden Durchlauf des Experiments bestimmt werden. Für die abhängigen Variablen werden die geeigneten Messmittel bestimmt und der Prozess der Datenerhebung festgelegt.

Als Beispiel kann die Salzkonzentration (abhängige Variable) im Wasser durch Widerstandsmessung bestimmt werden. Weitere andere Ionen dürfen nicht vorhanden sein (Randbedingung) und das Experiment kann mit verschieden Mengen (Prozess) von zugeführtem Salz (freie Variable) durchgeführt werden. Um zufällige Messfehler auszuschließen, werden die Experimente mehrfach durchgeführt.

Durchführung: Das Experiment wird jeweils nach Plan durchgeführt. Die durchgeführten Messungen liefern zunächst wertungsfreie Daten als Ergebnis des Experiments.

Auswertung: Die erfassten Daten werden in einem getrennten Schritt ausgewertet, oftmals mit mathematischen Methoden. Das Resultat der Auswertung sind dann kumulierte quantitative Werte oder ist ein qualitativer Report auf Basis der erfassten Daten.

Diese Schritte können analog auf die Konzipierung, Durchführung und Auswertung einer Prüfung übertragen werden. Auch für eine Prüfung muss das Ziel definiert sein, die Prüfung muss in Prüfungsauftrag, Prüfungsform (Methode), Aufgabenstellung entworfen werden, durchgeführt werden und es erfolgt eine Auswertung der Artefakte mit einer Bewertung. Letzteres mündet in einer Note oder in einer qualitativen Betrachtung im Rahmen eines Feedbacks, wie zum Beispiel als Kompetenzgraph [Lichtenberg 2016].

Das in naturwissenschaftlichen Fächern bekannte Vorgehen der Konzeption von Experimenten kann nun auf das Problem „Konzeption von Prüfungen“ angewandt werden.

Die Analogien helfen bei der Betrachtung von Einflussfaktoren und Zusammenhängen im Bereich der Prüfungskonzeption.

Ziele der Prüfung durch Learning Outcomes definieren

Ausgangspunkt eines Experiments ist eine Hypothese oder eine unbeantwortete Frage. Die Fragestellung einer Prüfung ist ob eine Mindestkompetenz erreicht wurde oder in welcher Ausprägung wurde die Kompetenz insgesamt erreicht? Die binäre Entscheidung, „bestanden oder nicht bestanden", stellt dabei in der Auswertung nur einen Schwellwert dar, sodass die eigentliche Fragestellung in der Ausprägung der Kompetenz liegt. Ausgangspunkt der Prüfungskonzeption ist das Learning Outcome des Moduls. Wurde das Learning Outcome in einem Satz mit einem einzigen Verb formuliert [nach Reis 2014], so kann der grobe *Prüfungsauftrag* und damit das Ziel der Messung durch eine einfache Umformulierung mit Verberststellung, d.h. das Verb steht am Anfang, erreicht werden. Hier exemplarisch gezeigt an einem Learning Outcome aus der Programmierausbildung:

„Die Studierenden können Programme unter Verwendung prozeduraler Paradigmen entwickeln."

Der reduzierte Kern der zu demonstrierenden Kompetenz dieses Learning Outcome ist „[...] Programme [...] entwickeln." In der Verberststellung ergibt sich daraus der Arbeitsauftrag für die Prüfung im Imperativ:

„Entwickeln [...] Programme [...]!" oder vollständig
„Entwickeln Sie Programme unter Verwendung prozeduraler Paradigmen!"

Durch die Bearbeitung dieses Prüfungsauftrags haben die Studierenden die Möglichkeit ihre Kompetenz zu demonstrieren.

Nach dieser Umformung eines Learning Outcome in einen Prüfungsauftrag können bereits zwei Aspekte als Qualitätssicherungsmaßnahme geprüft werden. Zum einen kann man hier bereits abschätzen, welche Ressourcen für die Durchführung dieses Experiments erforderlich sind. So erfordert zum Beispiel der Prüfungsauftrag: „Leiten Sie ein Projekt!", dass ein Projekt bereitsteht, welches über einen Zeitraum geleitet wird. In diesem Zeitraum muss durch den Prüfenden eine Beobachtung dieses Prozesses der Projektleitung durch den zu Prüfenden mit entsprechendem Zeitaufwand stattfinden. Zum anderen muss geprüft werden, ob überhaupt mehrere messbare Aspekte (abhängige Variablen) vorhanden sind, ob also eine Beobachtung zur Datenerhebung überhaupt möglich ist[1]. Die Demonstration der Kompetenz muss einen beobachtbaren Prozess darstellen oder ein beobachtbares/untersuchbares Artefakt liefern. Ein Gegenbeispiel ist die Formulierung eines Learning Outcome unter Verwendung der Verben „kennen" oder „verstehen", bei der sinnlose Prüfungsaufträge entstehen:

„Die Studierenden haben das Programmieren unter Verwendung prozeduraler Paradigmen verstanden."

In Verberststellung ergibt sich daraus der Prüfungsauftrag:

„Verstehen Sie das Programmieren unter Verwendung prozeduraler Paradigmen!"

[1] Hier können Verbindungen zum Begriff der „Beobachtbarkeit" aus der Systemtheorie gezogen werden.

Dieser Satz macht nur als Frage und nicht als Auftrag Sinn und die naheliegende Antwort auf die (Prüfungs-)Frage ist einfach: „Ja, verstehe ich.“ Derartig formulierte Learning Outcomes müssen dringend überarbeitet werden!

Somit sind einige Prüfungsaufträge prinzipiell nicht durchführbar, da innere Prozesse (verstehen) betrachtet werden müssten, die sich einer externen Beobachtbarkeit während ihrer Demonstration entziehen. Weiterhin kann sich ergeben, dass nötige Ressourcen für ein adäquates Prüfungssetting nicht verfügbar sind, wie zum Beispiel ein zu leitendes Team. In diesen Fällen muss, um im Constructive Alignment zu bleiben, das Learning Outcome angepasst werden.

Sind die Formulierungen des Learning Outcomes komplexer, so ergeben sich entsprechend komplexere Prüfungsaufträge.

Design der Prüfung

Messmethode (Prüfungsform)

Die aus dem Prüfungsauftrag zu entwickelnde genaue Aufgabenstellung und der Rahmen der Prüfung müssen dann die Demonstration der Kompetenz zum einen ermöglichen und zum anderen die Beobachtbarkeit weiterhin gewährleisten. Das führt dazu, dass vergleichbar mit der Wahl der Messmittel im Experiment die passende Prüfungsform gewählt werden muss. Eine umfangreiche Übersicht über verschiedene Prüfungsformen findet man beispielsweise in [Gerick, 2017]. Oftmals ist die Auswahl durch die Vorgaben der Prüfungsordnung bereits auf eine Menge oder auf eine einzelne Prüfungsform eingeschränkt. Trotzdem ergeben sich noch Freiheiten innerhalb der Prüfungsformen, die man kreativ nutzen kann.

Nach Gabi Reinmann [Reinmann, 2019] kann man die Prüfungsformen zunächst in symbolische Formen und in enaktive (demonstrierende) Formen unterteilen. Dabei wird bei der symbolischen Form noch zwischen mündlicher und schriftlicher Form unterschieden. Bei der Wahl der Prüfungsform kann man somit generell zwischen mündlicher Form, der Betrachtung von Artefakten oder dem Beobachten eines Prozesses unterscheiden. Aus dem Prozess heraus können sich ebenso bewertbare Artefakte ergeben. Zeigt sich somit die Kompetenz im Fachgespräch, führt die zu beobachtende Kompetenz zu einem schriftlichen oder gegenständlichen Artefakt oder zeigt sich die Kompetenz in einem zu beobachtenden Prozess?

Die grundlegende Entscheidung der prinzipiellen Prüfungsform leitet sich auch wieder aus dem Learning Outcome ab. Fordert das Learning Outcome beispielsweise das Lösen von Gleichungen mit einem Computer Algebra-System, so ist das produzierte Artefakt die ermittelte Lösung des Gleichungssystems. Wird weiterhin die Verwendung eines Computer Algebra-System gefordert, so muss die Prüfung auch den Einsatz an einem Computer ermöglichen und nicht eine schriftliche Papierklausur als Prüfungsform verwendet werden. Es wird bereits durch das Learning Outcome eine Anforderung an den Prozess der Bearbeitung, somit der Demonstration, gestellt. Wird der Einsatz eines Computer Algebra-Systems nicht gefordert, so kann man als Prüfungsform auch eine schriftliche Papierklausur wählen. Eine rein mündliche Prüfung würde den Lösungsprozess in den Vordergrund rücken, da im Fachgespräch der Werdegang der Lösung Kernthema sein kann. Genauso wenig sollte man Computerprogramme im Rahmen einer schriftlichen Papierklausur erstellen lassen, sondern die Prüfung ebenso am Computer unter Verwendung der entsprechenden Tools durchführen.

Ebenso sollte die Kompetenz der Durchführung einer chemischen Analyse in einem Laboraufbau geprüft werden und nicht nur auf Papier erfragt werden. Hier kann man noch entscheiden, ob das Endergebnis hinreichend ist oder ob der gesamte Prozess mit seinen einzelnen Schritten beobachtet werden muss. Diese Frage wird später noch bei der Diskussion von systembedingten Fehlmessungen wieder aufgegriffen.

Als Ergebnis der Auswahl der Messmethode steht der Prüfungsauftrag sowie die prinzipielle Messmethodik durch die Prüfungsform fest.

Messpunkte

Ist der Rahmen der Prüfung durch die Prüfungsform grob definiert, so folgt die Wahl der Messpunkte, d.h. welche Aspekte sollen durch die Prüfung beobachtet und als abhängige Variable als Maß für die Kompetenz erfasst werden? Diese Auswahl ist gekoppelt mit der späteren Auswertung und wird in dem Zusammenhang noch einmal betrachtet. Die Messpunkte bzw. Aspekte können sehr unterschiedlicher Art sein und sind entweder (dokumentierte) Schritte im Prozess oder Artefakte als Ergebnis der Demonstration der Kompetenz.

Die abhängigen, zu messenden Variablen können zum Beispiel Textqualität, Struktur des Berichts, Seitenzahl, Coding-Style, Methodenauswahl, Programmfunktion, Modellierung, Zwischenergebnisse, Prozessbeschreibungen, Kommunikationsverhalten in Projekt-Meetings, Berechnungsergebnis, Vollständigkeit von Diagrammen, Bearbeitungsdauer, Berechnungsergebnisse oder vieles mehr sein. Die Fachlichkeit des Themas bestimmt hier die Auswahl der Aspekte. Dem gegenüber können auch Fehler gemessen werden, wie zum Beispiel Anzahl der Rechtschreibfehler oder unsinnige Diagrammteile. Eine typische Frage in der Mathematik oder bei Berechnungen in den Ingenieurswissenschaften ist „Wird der Weg bewertet oder nur das Endergebnis?“ Beide Standpunkte werden kontrovers diskutiert und als Prüfer*in muss man hier eine Entscheidung zwischen Prozess und Artefakt treffen.

Zu diesem Zeitpunkt unbekannte Aspekte, wie alternative Lösungswege welche eine unbekannte abhängige Variable darstellen, können hier naturgemäß noch nicht berücksichtigt werden. Sie treten erst nach der Durchführung der Prüfung als Anomalie in Erscheinung und führen später zu einer nachgelagerten Adaption des Auswerteverfahrens.

Wurde das Learning Outcome in der sogenannten „indem“-Form [HAW, 2021] formuliert, so liefern die aufgelisteten Schritte Anhaltspunkte für wichtige abhängige Variablen. Ein Beispiel für eine Formulierung mit Schritten der Bearbeitung ist folgendes Learning Outcome:

„Die Studierenden können Methoden und Techniken des Systems und Software Engineerings zur systematischen Entwicklung eines Software-Konzeptes für ein mechatronisches System anwenden, indem sie dazu
1.) die Anforderungen und Randbedingungen systematisch erfassen,
2.) das System in Struktur und Verhalten entwerfen und modellieren,
3.) das Softwaresystem in Struktur und Verhalten entwerfen, modellieren und implementieren,
4.) Modelle geeignet transformieren und ergänzen
5.) und Qualitätssicherungsmaßnahmen auf den Entwicklungsebenen durchführen.“

Aus dem ersten Teilschritt lässt sich ableiten, dass als ein Aspekt geprüft wird, ob und wie gut die Anforderungen und Randbedingungen erfasst wurden und dort die in Punkt 5 geforderten Qualitätssicherungsmaßnahmen durchgeführt wurden. Entsprechend kann man

bei den weiteren Teilschritten ebenfalls die zu beobachtenden Hauptaspekte ableiten. Innerhalb der grob beschriebenen Aspekte können weitere feinere Aspekte definiert werden.

Alternativ kann eine Learning Outcome durch weitere Teilkompetenzen verfeinert werden [Lehmann, 2015]. Hier definiert jede Teilkompetenz jeweils einen Aspekt der Gesamtkompetenz und ist ein Hinweis auf einen Messpunkt innerhalb der Prüfung.

Wie bei allen Experimenten kann man schon in der Design-Phase die Frage stellen, ob es Korrelationen zwischen den abhängigen Variablen gibt und ob somit wirklich alle erfasst werden müssen. Beispielswiese muss der Umgang mit Programmiertools nicht separat geprüft werden, wenn die korrekte Bedienung zur Erstellung von Software zwingend erforderlich ist. Hier bilden sich gegebenenfalls sogar Taxonomien aus, d. h. eine Kompetenz kann nur bei Vorhandensein einer anderen demonstriert werden. Diese Ordnung in den Abhängigkeiten kann später bei der Auswertung zu Vereinfachungen des Bewertungsschemas führen.

Generell sollte die Auswahl der zu beobachtenden Aspekte klein gehalten werden, da für alle identifizierten Aspekte in der Prüfung Daten erhoben werden müssen. Im eingangs beschriebenen Fall mit dem Learning Outcome

„Die Studierenden können Programme unter Verwendung prozeduraler Paradigmen entwickeln."

wird gemessen, ob die Studierenden Programme erstellen können und welche prozeduralen Elemente Sie in der Lage sind bei der Entwicklung zu verwenden. Die zu erfassenden Aspekte werden sich auf die geforderte Funktionalität und auf die prozeduralen Elemente im Artefakt des erstellten Programms beziehen.

Versuchsaufbau (Aufgabenstellung)

Die Prüfungsaufgabe formuliert fachspezifisch konkret was und wie eine Kompetenz zu demonstrieren ist. Weiterhin gibt sie einen Rahmen vor, oder die Aufgabe wird in ein Szenario eingebettet, beispielsweise eine Fallbeschreibung. Dabei muss der Rahmen, die festen Parameter des Experiments, weiterhin die Entfaltung der Kompetenz ermöglichen oder zumindest das Erreichen der Messpunkte. Eine Einschränkung kann beispielsweise die Vorgabe einer anzuwendenden Lösungsmethode für den beschriebenen Fall sein. Messpunkte für die Kompetenz können dann nur noch innerhalb der Lösungsmethode liegen, es kann nur noch die Güte der Anwendung dieser Lösungsmethode gemessen werden. Die Wahl einer geeigneten Bearbeitungsmethode als (Teil-)Kompetenz kann dann entsprechend nicht mehr erfasst werden, da dieser Freiheitsgrad durch die Einschränkung nicht mehr gegeben ist (konstanter Parameter). Demgegenüber verringert das Festlegen von Rahmenbedingungen die Varianz der Lösungen und vereinfacht später die Auswertung.

Die Vorgabe von Zwischenschritten, z. B. durch die Abfrage von bestimmten Zwischenergebnissen bei Berechnungen, kann ebenso schon eine Leitlinie für die Stufen der Bearbeitung darstellen und schränkt somit die Kompetenz in der Auswahl der Bearbeitungsmethoden ein (Prozess vs. Ergebnis). Wird dagegen nur das Endergebnis betrachtet, kann keine Aussage über den Entstehungsprozess gemacht werden. Will man somit die Bearbeitung zu bestimmten Messpunkten hin führen, muss man die Bearbeitungsoptionen entsprechend eng setzen.

Das Einschränken des Lösungsraums kann auch dazu führen, dass sich K.O.-Kriterien für die Bearbeitung ergeben. Wird die Anwendung einer bestimmten Bearbeitungsmethode gefordert und diese ist bei dem zu Prüfenden nicht präsent, so kann gegebenenfalls keine

Bearbeitung erfolgen. Eine vielleicht vorhandene prinzipielle Kompetenz unter Einsatz alternativer Methoden kann/darf nicht demonstriert werden. Aufgabenstellungen sollten durch eine Analyse auf derartige Schwachstellen geprüft werden, hier gezeigt anhand eines Beispiels aus einer Programmierprüfung:

Wieder ein Beispiel aus der Software-Entwicklung: Software kann unabhängig von der Qualität der inneren Struktur vollständig korrekt arbeiten. Eine Teilkompetenz bei der Programmentwicklung zur Herstellung einer guten inneren Qualität ist die Aufteilung des prozeduralen Programmcodes in Funktionen und ist somit ein Aspekt der Prüfung im prozeduralen Programmieren. In der Aufgabenstellung wurde beispielsweise gefordert, dass die Bestimmung von Minimum und Maximum eines Datensatzes durch eine Funktion erfolgen soll, deren Ergebnisse mittels des Verfahrens „Call-by-Reference“ geliefert werden. Somit sollte gemessen werden, ob die Studierenden in der Lage sind, derartige Funktionen zu erstellen. Prinzipiell lässt sich die Gesamtfunktionalität der zu entwickelnden Software auch mit zwei einzelnen Funktionen ohne Verwendung von „Call-by-Reference“ oder sogar ohne die Strukturierung in Funktionen realisieren. Funktionen unter Verwendung von „Call-by-Reference“ stellen hier die komplizierteste Variante dar und erfordern somit ein tieferes Verständnis in der Anwendung. Die Forderung der Aufgabenstellung nach der Verwendung von „Call-by-Reference“ stellt aber ein K.O.-Kriterium dar, da die Funktion ein wesentlicher Zwischenschritt für die Gesamtfunktionalität des Programmes ist. Andere Aspekte der Programmiertätigkeit können nicht demonstriert werden, wenn diese Funktion nicht erstellt wird. Die Software lässt sich allerdings auch ohne diese Vorgabe erstellen. Somit wurde die Aufgabenstellung abgewandelt, sodass die Art der Strukturierung in Funktionen freigestellt, aber wichtiger Teil der Bewertung ist und es wurde auch auf diesen Punkt hingewiesen. Hier zeigt sich auch die Bildung einer inhaltlichen Taxonomie, da man das programmiertechnische Problem ohne, mit mehreren oder mit einer speziellen Art von Funktionen lösen kann. Jedes Vorgehen erfordert eine jeweils fundiertere Kompetenz um die Anwendung der Programmiertechniken.

Systematische Designfehler / Störgrößen / Fehlmessungen

Auch beim Design von Prüfungen muss man sich über systematische Fehler oder Störgrößen auf die Messung Gedanken machen.

Als systematischer Fehler in Prüfungen werden oft systembedingt falsche Rückschlüsse aus den Beobachtungen auf die Kompetenz gezogen. Das *Ergebnis* einer Projektarbeit wird beispielsweise als Maß für die Qualität der Projekt*durchführung* angesehen. Die Qualität der Projektdurchführung kann nur durch die Beobachtung der eigentlichen Arbeit im Projekt erfolgen, da Ergebnisse auch bei nicht systematisch durchgeführten Projekten erzielt werden und umgekehrt nach Lehrbuch durchgeführte Projekte genauso scheitern können. Es wird oftmals eine Korrelation zwischen Entstehungsprozess und Ergebnis unterstellt, die nur begrenzt vorhanden ist. In einem beobachteten Fall wurde die Kompetenz der Projekt*durchführung* durch eine Klausur über das *Wissen der Methoden* geprüft. Hier liegt der systematische Fehler in der Wahl einer nicht adäquaten Prüfungsform.

Ein weiterer Fehler ist die Erfassung von irrelevanten bzw. den falschen Aspekten, die später in der Auswertung zu Fehlern führen. Die Darstellungsform, wie Wortgewandtheit oder die optische Aufmachung eines zu produzierenden Videos, lenkt vom eigentlichen zu berücksichtigendem Inhalt ab oder wertet diesen unangemessen auf (→Blendung) oder ab. In mündlichen Prüfungen, bei Präsentationen oder Abschlussarbeiten wird fälschlicherweise mehr die Beziehung zwischen Prüfenden und zu Prüfenden gemessen und verfälscht die

Auswertung (→Halo-Effekt). Bei Präsentationen werden Aspekte erfasst, die zwar als selbstverständlich angesehen werden, aber nie definiert wurden, wie z.B. das Tragen formeller Kleidung während der Prüfungen („In Jogginghose, das geht ja gar nicht!“).

Damit diese Störfaktoren nicht in die Bewertung einfließen, sollten diese natürlich erst gar nicht erfasst werden. Da sie oft nur unterbewusst wahrgenommen werden, ist es hier hilfreich, sich dieser Aspekte bewusst zu sein und sie dann nicht einfließen zu lassen. Oder diese explizit machen, beispielsweise das formelle Kleidung Voraussetzung ist.

Durchführung und Datenerfassung

In der Durchführung der Prüfung auf Basis der Prüfungsaufgaben und des gewählten Settings kommt es zur Demonstration der Kompetenz. Hier erfolgt dann die Datenerhebung durch Beobachtung oder das abschließende Erfassen der erstellten Artefakte. Erstes Ergebnis der Durchführung sind Protokolle, Checklisten oder Stapel an Klausuren und Hausarbeiten. Typischerweise erfolgt keine Wiederholung mit Variation von Parametern für mehr statistische Sicherheit, sondern das Experiment wird pro zu Prüfenden innerhalb einer Prüfungsphase nur einmalig durchgeführt.

Die Datenerhebung muss von einer Bewertung deutlich getrennt werden. Die Korrektur einer Klausur ist noch Teil der Datenerhebung. Es wird erfasst, welche Bearbeitungsmethode ausgewählt wurde (ohne Wertung der Auswahl), ob die Berechnung korrekt ausgeführt wurde, ob alle Elemente in einem Diagramm sinnvoll ausgewählt sind, usw. Das sind noch Schritte der Datenerfassung. Oftmals werden diese Erhebungen direkt mit einer Bewertung vermischt, da Attribute wie gut, mäßig, o.ä. verwendet werden oder direkt eine bewertende Punktzahl vergeben wird. Die Bewertung der Daten ist ein von der Erhebung unabhängiger nachgelagerter Schritt.

Auswertung (Bewertung)

Die Planung der Auswertung ist eng mit der Planung der Messung und somit mit dem Design der Prüfung gekoppelt und sollte entsprechend gemeinsam durchgeführt werden. Aus der Betrachtung der Ergebnisse für die einzelnen Aspekte muss eine Gesamtbewertung abgeleitet werden. Dabei beeinflussen sich die Menge und die Art der Aspekte sowie die Abbildung auf eine Bewertung gegenseitig.

Die Ausprägung einer Kompetenz kann insgesamt als eine kontinuierliche Größe angesehen werden. Sie setzt sich aus vielen Aspekten zusammen, welche die Dimensionen einer Kompetenz aufspannen. Die Datenerhebung über die einzelnen Aspekte liefert somit einen Punkt in einem multidimensionalen Raum, der die (Gesamt-)Kompetenz basierend auf den beobachteten Aspekten A_i darstellt (siehe Abbildung 1). In einer für die Auswertung von Experimenten typischen mathematischen Betrachtung erhält man eine Menge von Aspekten $A = \{A_i\}$, die in eine Bewertung eingehen, eventuell inklusive der ungewollten Aspekte.

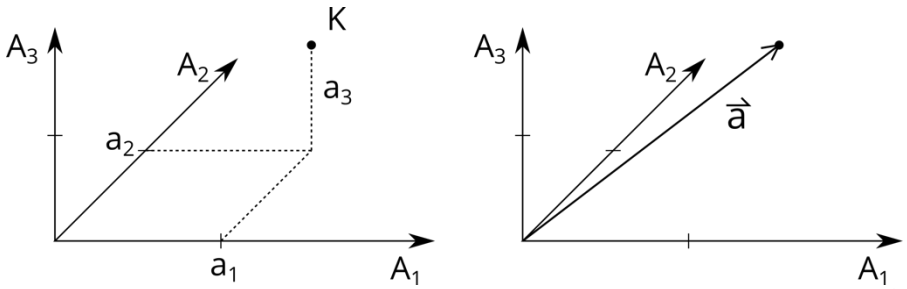

Abbildung 1: Gesamtkompetenz K aus den drei Aspekt-Dimensionen A_1-A_3 bzw. als Vektor $\vec{a}$

Im Rahmen einer formativen Prüfung zum Zwecke des Feedbacks kann man bei diesen einzelnen Aspekten bleiben, da jeder Teilaspekt betrachtet werden kann und den zu Prüfenden einen Hinweis auf Veränderung geben soll [Lichtenberg 2016]. Für eine summative Prüfung soll eine diskrete Gesamtausprägung im Sinne einer Note n ermittelt werden. Die Menge der Noten N ist in den Prüfungsordnungen festgelegt und kann einfache Bezeichnungen N = {„sehr gut“, „gut“, ...} im Sinne einer Ordinalskala oder numerische Werte wie N = {15, 14, 13, ...} enthalten. Ob es sich dann auch wirklich um eine Ratioskala handelt oder nur auf Grund der Verwendung von Zahlen den Anschein macht, muss im Einzelnen betrachtet werden.

Aufgabe der Bewertung und des dafür zu entwickelnden Schemas ist es aus dem Raum der erfassten Aspekte auf die Noten des Bewertungsschemas aus der Prüfungsordnung abzubilden, d. h. möglichst eine Funktion $f: A \rightarrow N$ zu finden. Diese mathematische Betrachtung soll helfen die verschiedenen möglichen Bewertungsschemata einzuordnen und soll wieder Rückschlüsse auf die Auswahl der zu bewerten Aspekte liefern.

Punkteschema und Kriterienraster

Für eine mathematische Auswertung der Ergebnisse müssen die Resultate der einzelnen Aspekte mit einer Ratioskala erfasst oder mindestens auf eine Ratio-/Intervallskala abgebildet werden. Beobachtungen werden oftmals nicht ohne Wertung erfasst, es wird vielmehr mindestens eine Ordinalskala mit Begriffen wie „gut“, „unzureichend“, „umfangreich“ oder ähnliches verwendet. In schriftlichen Klausuren, beispielsweise im Fach Mathematik, werden oftmals direkt Punkte zugeordnet, zum Beispiel gibt es für das erfolgreiche Durchführen einer Berechnung fünf Punkte, wobei durch die Verwendung eines numerischen Wertes direkt eine Ratioskala impliziert wird. Im Übergang von der Ordinalskala zu einer Ratioskala werden den beschreibenden Begriffen Werte zugewiesen, basierend auf einer Wertigkeit von Aufgabenstellungen, Komplexität von Lösungen, Dauer der Durchführung oder Ähnlichem. Die Begründung für die Abbildungen und Zuordnungen auf eine Ratioskala müssen plausibel und transparent sein und sich aus der gegebenen Fachlichkeit ableiten. Für eine Diskussion über den Übergang zwischen Skalen sei beispielsweise auf [Ritschl, 2016] verwiesen.

Für die Mathematik-Affinen: Liegt allen erfassten Aspekten eine Ratioskala zu Grunde, so spannen die Aspekte als Basis einen Vektorraum auf und das Ergebnis der Messung kann als Vektor $\vec{a}$ (siehe Abbildung 1 rechts) aufgefasst werden. Nun wird also eine Funktion gesucht, die den Vektor auf die Note abbildet $f(\vec{a}) \rightarrow n$. Dabei kann man zunächst den Vektor auf einen kontinuierlichen skalaren Wert mittels $f'(\vec{a})$ reduzieren und dann in einem zweiten Schritt mittels $g() \rightarrow n$ auf die diskrete Note abbilden $n = g(f'(\vec{a}))$. Die Funktion $g()$ muss dabei nicht unbedingt auf gleichmäßige Intervalle abbilden.

Umgekehrt kann man fragen, welche Menge an Punkte im Vektorraum der möglichen Bewertungen bildet auf die gleiche Note ab (vgl. Abbildung 2)? Für ein Minimum (Grenze nicht bestanden/bestanden) ergibt sich oft eine (Hyper-)Ebene in den Dimensionen die die Aspekte aufspannen. In Abbildung 2 wird der Aspekt a_2 doppelt gegenüber dem Aspekt a_1 gewichtet und aussummiert. Es ergibt sich eine Gerade für die Grenze zwischen bestehen (oberhalb) und nicht-bestehen (unterhalb). Es zeigt, wie sich die beiden betrachteten Aspekte bei der Bildung der Gesamtbewertung gegenseitig kompensieren. Die Frage kann man bei mehreren Dimensionen dahingehend erweitern, welche Hyperebenen trennen die einzelnen Notenstufen gegeneinander ab?

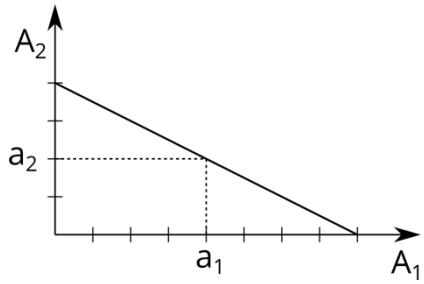

Abbildung 2: Linie trennt Notenstufe als Linearkombination der Aspekte A_1 und A_2. Aspekt A_2 hat doppelte Gewichtung.

Eine Kompensation von Aspekten in der Bewertung ergibt sich durch die Bildung der Summe über alle Dimensionen, das Aufsummieren aller Teilpunkte. Mathematisch wird hier das Skalarprodukt gebildet: $f'(\vec{a}) = \vec{w}^T \circ \vec{a}$ mit dem Gewichtungsvektor $\vec{w}$, hier mit $w_i = 1$. Jeder erzielte Punkt kompensiert einen nicht erhaltenen Punkt, egal welche Art von Kompetenz(-ausprägung) jeweils dafür erforderlich war. Die Betrachtung zeigt auch, dass diese Funktion nicht umkehrbar ist und man somit nicht aus der Gesamtbewertung auf die Ausprägung in den einzelnen Aspekten schließen kann.

Im Bewertungsschema Kriterienraster, beispielsweise angewandt bei der Bewertung von Abschlussarbeiten, werden die Faktoren w_i oftmals unterschiedlich gewichtet, somit oft $w_i \neq 1$, um die Relevanz (Skalierung) der einzelnen Aspekte untereinander zu steuern. Eine Skalierung, und die damit verbundene Gewichtung der Aspekte untereinander, erfolgt allerdings auch versteckt beim Übergang von den Beobachtungen in einer Ordinalskala auf die Ratioskala der zugehörigen Dimension des Vektorraums. Bei diesem Übergang werden der Beobachtung Punkte im Kriterienraster zugeordnet und hier kann zusätzlich eine Wertung des Aspektes einfließen. Dieser Übergang von der Datenerhebung zur Bewertung wird dann bei Einsichtnahmen in die Prüfungsunterlagen gerne als Verhandlungsbasis verwendet: „Kann ich nicht dafür noch einen Punkt mehr bekommen?“

Weiterhin zeigt diese Betrachtung, dass je mehr Aspekte und somit Dimensionen betrachtet werden, je feiner und detaillierter die Messung insgesamt erfolgt, desto schwieriger wird die formelle Definition der Abbildungsfunktion. Ein Ansatz zu Reduktion ist das nicht betrachten von Aspekten („es muss nicht alles geprüft werden“) in Kombination mit der Reduktion zu einem Niveaustufenmodell mit einer einzigen betrachteten dominanten (diskreten) Dimension.

Niveaustufenmodell

Im Niveaustufenmodell erfolgt eine Messung der Kompetenz nur innerhalb einer Taxonomiestufe. Basierend auf dem Ansatz der Taxonomie (beispielsweise nach Bloom[Bloom, 1972]) werden Korrelationen ausgenutzt, d.h. Aufgaben in höheren Taxonomiestufen lassen sich nur bearbeiten, wenn Kompetenzen auf den niedrigen Taxonomiestufen vorhanden sind. Voraussetzung ist auch hier, das die Taxonomiestufe sich aus dem Learning Outcome ableiten lässt. Innerhalb der Taxonomiestufe soll das Niveau der Kompetenz in Stufen durch Messung in nur einer Dimension erfasst werden; es wird somit von einer Dominanz einer Dimension i und somit $w_i \gg w_k$ für $i \neq k$ ausgegangen[2]. Das erreichte Niveau wird in Stufen eingeteilt und beschrieben. Optional wird eine weitere Dimension betrachtet, um die definierten Stufen in der primären Dimension weiter zu

[2] Ergebnisse in der dominanten Dimension sind hoch gewichtet, damit sie in der Summenbildung auch eine Dominanz auf die Note haben.

differenzieren[3]. Die Reduktion auf eine Dimension erleichtert die deskriptive Definition von Schwellen statt durch Hyperebenen mit vielen Aspekten. Entsprechend wird die Beobachtung zunächst deskriptiv erfasst und einer Ordinalskala zugeordnet. In einem zweiten Schritt werden die Stufen, unter Berücksichtigung von differenzierenden Aspekten, auf eine Bewertung abgebildet.

Die Schwierigkeit bei diesem Vorgehen ist das Finden der dominanten Dimension und das beschreiben der Stufen in der Ordinalskala. Beispielsweise kann die Bearbeitung eines Falles Ausgangspunkt der Prüfung sein. Die Schritte des Bearbeitungsprozesses sollen dokumentiert werden und es wird bewertet, wie weit der Fall korrekt bearbeitet oder alternativ wie ausdifferenziert der Fall bearbeitet wurde.

Im Bereich der Programmierung kann beispielsweise die Bearbeitung einer Problemstellung den Einsatz von immer komplexeren Programmierkonzepten erfordern (siehe Abbildung 3). Bewertungsmaß ist dann die erreichte Komplexitätsstufe der Lösungen. Im Umkehrschluss kann aus der Bewertung abgelesen werden, bis zu welcher Komplexitätsstufe die Studierenden in der Lage sind, Lösungen in Form von Software abzuliefern (vgl. Umkehrbarkeit der Abbildungsrelation Kompetenz und Note).

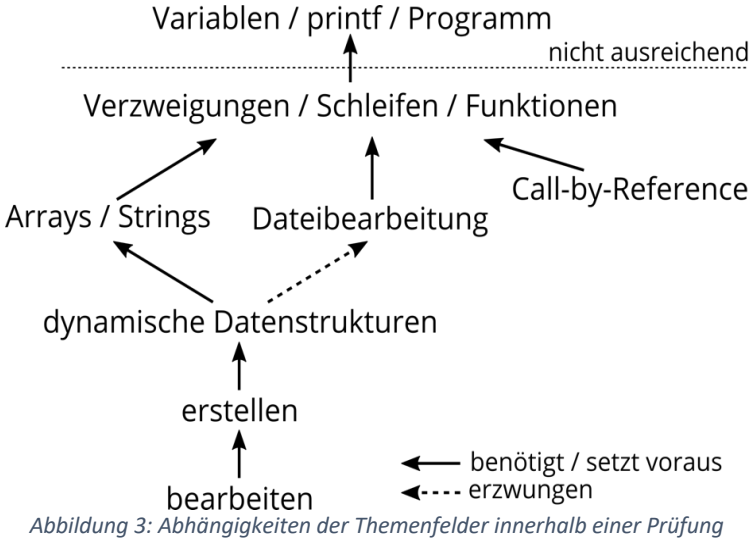

Abbildung 3: Abhängigkeiten der Themenfelder innerhalb einer Prüfung

Beispiel: Stufen einer Programmierprüfung (in der Mechatronik 1. Semester):

A – (Minimum) Die Studierenden erstellen Programme unter Verwendung von Eingaben, Rechenoperationen, Verzweigungen, Schleifen und Funktionen.
B - Die Studierenden erstellen Programme unter Verwendung statischer Arrays oder Strings.
[...]
E - Die Studierenden erstellen Programme unter Verwendung von Datei-Operationen und von dynamischen Strukturen.
F – (Maximum) Die Studierenden erstellen Programme unter Verwendung von Datei-Operationen, dynamischen Strukturen und können Auswertungen/Operationen in den Strukturen durchführen.

Die Aufgaben der Prüfung werden so konstruiert, dass die Studierenden für die Lösung entsprechend dem Niveaustufenmodell immer komplexere Elemente für die Lösung der Problemstellung einsetzen müssen. Die Formulierung der Aufgabenstellung kann dabei die

[3] Das Verfahren ist von der Idee vergleichbar mit der Principal Component Analysis (PCA)[Pearson 1901], nur werden hier keine statistischen Merkmale, sondern systematisches Überlegungen eingesetzt.

Einhaltung fordern, indem die „dynamischen Strukturen“ in Stufe „E“ mittels „Dateioperationen“ aufgebaut werden müssen. Somit wird die Abhängigkeit der beiden Themen in der Niveaustufe „E“ eingehalten (siehe gestrichelte Abhängigkeit in Abbildung 3), andererseits ein K.O.-Kriterium hinein konstruiert.

Insgesamt vereinfacht das Niveaustufenmodell die Bewertung, da nach der Datenerhebung eine Klassifizierung in wenige Stufen notwendig ist, die sich über differenzierende Faktoren auf Noten abbilden lassen. Gefühlt gibt es im Niveaustufenmodell eine höhere Unschärfe in der Abbildung von der Beobachtung auf die Ordinalskala des erreichten Niveaus und auf die Note als beim Punktesystem/Kriterienraster. Bei genauer Betrachtung ist die Unschärfe im Punktesystem genauso vorhanden.

Säulenmodell

Eine Zwischenstufe zwischen Kriterienraster und Niveaustufen stellt das Bewertungsschema Säulenmodell dar. Im Säulenmodell werden mehrere dominante Aspekte und somit Dimensionen nebeneinander betrachtet. Analog zum Niveaustufenmodell sind diese einzelnen Dimensionen in Stufen deskriptiv definiert. Man kann das Säulenmodell auch als mehrere nebeneinanderstehende Niveaustufenmodelle für die einzelnen Dimensionen ansehen.

Die verschiedenen beobachtbaren Aspekte in den Säulen werden bei der Bewertung zusammengefasst. Beispielsweise hat Christian Decker [Decker 2016] in seinem Modell die Bewertung einer Hausarbeit im Bereich wissenschaftliches Arbeiten in die Säulen „Explanation“, „Principles“ und „Conclusion“ unterteilt und in jeder Säule Ausprägungen für die einzelnen erreichten Stufen definiert. Die Definitionen sind so formuliert, dass die Beschreibungen der einzelnen Stufen jeweils einen insgesamt qualitativ beschreibenden Satz als Gesamtbewertung ergeben. Die Gesamtbewertung wird dann in einem weiteren Schritt auf die Note abgebildet. Der beschreibende Satz liefert für die Studierenden ein Feedback sowie eine Begründung für die Note.

Das Säulenmodell kann auch im Kleinen für die Bewertung von einzelnen Teilaufgaben einer Prüfung verwendet werden. Beispielsweise soll im Bereich Softwareentwicklung ein sogenanntes Klassendiagramm als eine Teilleistung auf Basis eines Falls erstellt werden. Die Hauptdimension für die Bewertung ist dabei die Struktur der Elemente im Diagramm mit allen strukturellen Elementen (Klassen und Relationen) und in der zweiten Dimension die korrekte Annotation der Elemente. Die Aufteilung resultiert daraus, dass für eine Annotation die zu annotierende Struktur vorhanden sein muss. Weiterhin geht in die Klassifikation der Lösung mit ein, ob in dem Diagramm unwichtige oder falsche Elemente eingebaut wurden. Für die Bewertung werden die Lösungen im ersten Schritt in vier Gruppen eingeteilt (Stapel nach der Korrektur) und dann im zweiten Schritt wird jeder Stapel in sich differenziert nach den Stufen in der Annotation. Die Position im Stapel (Stapel + Differenzierung) wird dann auf eine erreichte Punktzahl abgebildet. Die Punkte pro Aufgabe werden innerhalb der Prüfung über alle Teilaufgaben aufsummiert, da zu den anderen Teilaufgaben der Prüfungen keine Taxonomie konstruiert werden kann und diese als gleichwertig angesehen werden.

Im obigen Beispiel erfolgte die Abbildung aus erreichten Stufen in zwei Dimensionen wieder auf eine erreichte Punktzahl. Ein anderes Modell für das Zusammenfassen der Säulen verwendet das Minimum in den Säulen für die Abbildung auf die Note und differenziert über das Maximum in den anderen Säulen. Hier muss die zu Grunde liegende Fachlichkeit die Begründung für die Aggregation der Stufen liefern.

Fazit

Zwischen einem Experiment und einer Prüfung finden sich viele Ähnlichkeiten. Die Analogie zwischen der Gestaltung von Experimenten und Prüfungen kann zur Konstruktion von Prüfungen und zur Reflexion der Tätigkeiten genutzt werden. Dabei darf das Vorgehen beim Entwurf eines Experiments nicht streng auf die Entwicklung einer Prüfung übertragen werden. Es kann mehr als ein Leitbild verstanden werden, das unser Verständnis über den Komplex *Prüfungen* verbessert.

Meiner Auffassung nach kann das Experiment gerade für Neueinsteiger*innen in der Lehre ein guter Leitgedanke bei der Entwicklung ihrer Prüfungen sein. Später kann das Leitbild bei der Verbesserung der Prüfungen unterstützen.

Nach wie vor ist die Entwicklung einer Prüfung ein kreativer Prozess, der viele Möglichkeiten der Gestaltung in jedem Abschnitt bietet und fordert. Weiterhin zeigt dieser Artikel, das man auch mit der Brille seiner eigenen Fachlichkeit auf die Lehre schauen sollte um ein besseres Verständnis über das eigene Handeln zu bekommen.

Literatur

[Biggs, 1982] “Evaluating the Quality of Learning - the SOLO Taxonomy”, Biggs, J.B., and Collis, K.F., New York: Academic Press, 1982

[Biggs, 2007] “Teaching for Quality Learning”, Biggs, J. and Tang, C. McGraw-Hill Companies,Inc., 2007

[Bloom, 1972] „Taxonomie von Lernzielen im kognitiven Bereich“, Benjamin S. Bloom, Beltz Verlag, Weinheim und Basel, 1972

[DAAD, 2008] “Lernergebnisse (Learning Outcomes) in der Praxis”, Declan Kennedy, (Terence Mitchell and Volker Gehmlich and Marina Steinmann), DAAD, 2008

[Decker, 2016]: Die fallbasierte Klausur als schriftliche Prüfungsleistung. Christian Decker In: Haag, Johann; Weißenböck, Josef; Gruber, Wolfgang; Freisleben-Teutscher, Christian F. (Hrsg.): Kompetenzorientiert Lehren und Prüfen. Basics-Modelle-Best Practices. Tagungsband zum 5. Tag der Lehre an der FH St. Pölten am 20.10.2016. St. Pölten, S. 75 – 84. URL: http://skill.fhstp.ac.at/wp-content/uploads/2016/11/Tagungsband2016.pdf (Stand: 23.08.2020).

[Gerick, 2017] “Kompetent Prüfungen gestalten”, von Gerick, Julia ; Sommer, Angela ; Zimmermann, Germo, utb, 2017

[HAW, 2021] „Lehre lotsen“, HAW Hamburg, 2021, ISBN 978-3-00-067460-0

[Kennedy, 2007] “Writing and Using Learning Outcomes: A Practical Guide”, Kennedy, D., University College Cork, 2007

[Lehmann, 2015] „Lecture Engineering“, Thomas Lehmann, SEUH15, Dresden, 2015

[Lichtenberg 2016] „Kompetenzgraphen zur Darstellung von Prüfungsergebnissen. Ein Visualisierungsinstrument für individualisierte Leistungsbeobachtungen“, Gerwald Lichtenberg; Reis, Oliver, Aus: Neues Handbuch Hochschullehre. [Teil] H. Prüfungen und Leistungskontrollen. 6. Weiterentwicklung des Prüfungssystems in der Konsequenz des Bologna-Prozesses. Berlin: DUZ Verlags- und Medienhaus (2016) H 6.3, S. 99-120

[Pearson, 1901], „On Lines and Planes of Closest Fit to Systems of Points in Space”, 1901, Philosophical Magazine

[Reinmann, 2019] „Forschendes Lernen prüfen“, Gabi Reinmann, Prüfen hoch 3, HUL Universität Hamburg, 2019

[Reis, 2014] „Systematische Theologie für eine kompetenzorientierte Religionslehrer/innenausbildung“, Band 4 von Theologie und Hochschuldidaktik, Oliver Reis, Verlag LIT Verlag Münster, 2014

[Ritschl, 2016] “Wissenschaftliches Arbeiten und Schreiben“, Valentin Ritschl, Roman Weigl, Tanja Stamm, Springer, Berlin, Heidelberg, 2016

[Siebertz, 2017] „Statistische Versuchsplanung - Design of Experiments (DoE)“, Karl Siebertz and David van Bebber and Thomas Hochkirchen, Springer, 2010