BERND HAASLER & ANDREA MAURER
(Pädagogische Hochschule Weingarten & Universität Bremen)

Bewertung von Lösungen gestaltungsoffener Testaufgaben zur Messung berufsfachlicher Kompetenzen: Möglichkeiten und Schwierigkeiten einer internationalen Vergleichbarkeit

Abstract

Die Entscheidung, zur Messung berufsfachlicher Kompetenzen gestaltungsoffene Testaufgaben im Paper-and-Pencil-Design einzusetzen, zieht letztendlich die größte Herausforderung, das Bewerten der Lösungen (Rating) nach sich. Um das Rating nach wissenschaftlich haltbaren Anforderungen durchzuführen, sind umfangreiche Raterschulungen und Testratings mit den Bewertern notwendig. In diesem Beitrag wird aus zwei empirischen Modellvorhaben zur Messung berufsfachlicher Kompetenzen von angehenden Elektronikern berichtet. Der Schwerpunkt der Darstellung konzentriert sich auf Praxiserfahrungen der Ratings. Besonders erwähnenswert an diesen Vorhaben ist der Kontrast. Das ursprünglich in Deutschland konzipierte und durchgeführte Projekt, konnte als Parallelvorhaben auch in Peking realisiert werden. Die Erfahrungen aus Deutschland und China zeigen Möglichkeiten und Probleme auf, die international vergleichende Vorhaben zur Messung beruflicher Kompetenzen in sich tragen. Wenn die Ratings der Lösungsvarianten der Probanden in den beteiligten Ländern durch lokal gebundene Raterteams durchgeführt werden, treten auch länderspezifische Bewertungstendenzen zu Tage, die sich auf kulturellen Bindungen der Bewertergruppen gründen. Nicht zuletzt dieser Gesichtspunkt muss bei Ideen, ein international vergleichendes Berufsbildungs-PISA durchzuführen, berücksichtigt werden.

1 Gestaltungsoffene Testaufgaben

Wenn man als Bezugspunkt für die Messung beruflicher Kompetenz die konkrete Facharbeit heranzieht, kommt man nicht darum herum, gestaltungsoffene Testaufgaben einzusetzen. In zwei Forschungs- und Entwicklungsvorhaben zur Messung beruflicher Kompetenzen1 wurde dieser Aufgaben-Ansatz verfolgt, aus dem in diesem Beitrag berichtet werden soll. Die Grundlagen zur Entwicklung der Testaufgaben in diesen Projekten gehen vor allem auf die Arbeiten von ANDREAS GRUSCHKA zurück, der Evaluationsaufgaben als nachgestellte Entwicklungsaufgaben im Kollegschulprojekt in Nordrhein-Westfalen zur Evaluation der (schulischen) Erzieherausbildung entwickelte (GRUSCHKA 1985). Eine Kernarbeitsgruppe der Universität Bremen hat in den letzten zehn Jahren im Rahmen mehrerer Forschungsvorhaben am Beispiel verschiedener gewerblich-technischer Berufe als zentrales Instrument für

[1] Schul-Modellversuch KOMET der Bundesländer Hessen und Bremen (2007–2011) und KOMET/China: Competence assessment for vocational students in Beijing (TVET-PISA) (2008–2009). Die wissenschaftliche Begleitung beider Vorhaben wurde unter der Leitung von Felix Rauner, Universität Bremen, durchgeführt.

die Erfassung der beruflichen Kompetenzentwicklung Paper & Pencil-Aufgaben entwickelt und erprobt. Zusammenfassend können folgende Leitlinien genannt werden, nach denen gestaltungsoffene Testaufgaben entwickelt wurden (siehe Tabelle 1):

Tabelle 1: **Leitlinien zur Entwicklung gestaltungsoffener Testaufgaben (RAUNER/ HAASLER/ HEINEMANN/ GROLLMANN 2009, S. 101)**

Die Testaufgabenstellung

- erfasst ein realistisches Problem beruflicher und betrieblicher Arbeitspraxis.
- inkorporiert die charakteristischen beruflichen Arbeitsaufgaben des Berufes und die darauf bezogenen Ausbildungsziele.
- steckt einen berufsspezifischen – eher großen – Gestaltungsspielraum ab und ermöglicht damit eine Vielzahl verschiedener Lösungsvarianten unterschiedlicher Tiefe und Breite.
- ist gestaltungsoffen, d. h., es gibt nicht die eine „richtige“ oder die „falsche“ Lösung, sondern anforderungsbezogene Varianten.
- erfordert bei ihrer umfassenden Lösung außer fachlich-instrumentellen Kompetenzen die Berücksichtigung von Aspekten wie Wirtschaftlichkeit, Gebrauchswertorientierung und Umweltverträglichkeit.
- erfordert bei ihrer Lösung ein berufstypisches Vorgehen. Die Bewältigung der Aufgabe konzentriert sich auf den planerisch-konzeptionellen Aspekt und wird dokumentiert unter Verwendung einschlägiger Darstellungsformen (Paper-and-pencil-Design).
- muss nicht praktisch gelöst werden, da die Testaufgabe berufliche Kompetenzentwicklung auf der Konzeptebene misst und nicht auf der Ebene konkreten beruflichen Könnens (Performanz).
- ist keine Lernerfolgskontrolle; die Testaufgaben sind nicht Input-related.
- fordert den Probanden dazu heraus, die Aufgabe im Sinne beruflicher Professionalität (auf dem jeweiligen Entwicklungsniveau) zu lösen, zu dokumentieren und zu begründen.
- stellt auch für eine Fachkraft eine ernstzunehmende Herausforderung dar, gleichwohl muss auch einem Berufsanfänger eine Zugangsmöglichkeit zur Aufgabe geboten werden, die ihm die Bearbeitung ermöglicht.

Lösungsvarianten von gestaltungsoffenen Testaufgaben können nur von Ratern bewertet werden, die diese fachlichen Lösungen verstehen, zu deuten wissen und sie in den berufstypischen Kontext professioneller Akteure der Facharbeit einordnen können. In den hier illustrierten Projekten waren als Rater Lehrkräfte aus dem Berufsfeld Elektrotechnik-Informatik beteiligt, die in Berufsschulen und Studienseminaren tätig sind. Nachfolgend ist zur Veranschaulichung exemplarisch eine der Testaufgaben aus dem Untersuchungs-Set von insgesamt vier Testaufgaben dargestellt (siehe Abbildung 1). Alle Testaufgabenstellungen und ihre Lösungsrahmen werden im Band III der KOMET-Publikationsreihe veröffentlicht, der sich im Erscheinen befindet.

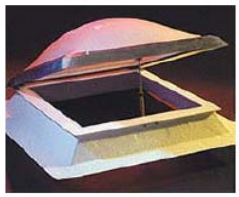
Aufgabe 1

Dachfenster-Steuerung

Situationsbeschreibung

Die Firma Gut & Pünktlich GmbH produziert im Zwei-Schicht-Betrieb (Mo. bis Fr. von 6:00 Uhr bis 22:00 Uhr, Sa. von 6:00 Uhr bis 14:00 Uhr) Einrichtungen für Flugzeugküchen. Die vier Dachfenster einer beheizten Montagehalle wurden bisher dezentral von vier Stellen manuell per Handkurbel geöffnet bzw. geschlossen (siehe Abbildung 1). Durch diese zeitaufwändige Art der Dachfensterbetätigung kam es u. a. dazu, dass abends vergessen wurde, die Dachfenster zu schließen, bzw. bei Sturm wurden offene Dachfenster beschädigt.
Die Werkleitung wünscht eine neue, komfortablere und sichere Steuerung der Dachfenster.
In einem Mitarbeitergespräch werden weitere Anforderungen formuliert:

- *„Die Dachfenster sollen zentral geöffnet und geschlossen werden."*
- *„Wenn die Temperatur im Arbeitsbereich der Halle zu hoch ist, müssen die Fenster öffnen."*
- *„Für das kommende Jahr ist eine Vergrößerung der Montagehalle geplant."*



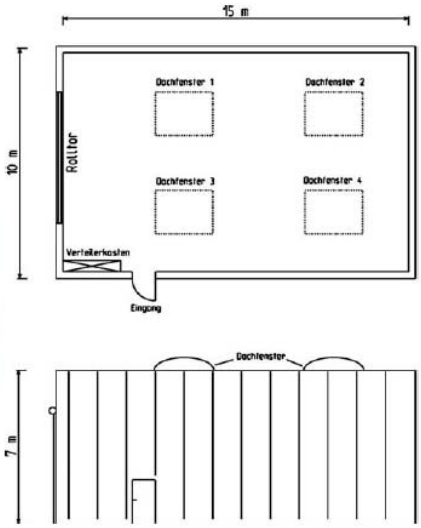


Abbildung 1: Detailaufnahme Dachfenster und Skizze der Montagehalle

Aufgabenstellung

Erstellen Sie möglichst vollständige Unterlagen für die Änderung der Anlage. Falls Sie noch zusätzliche Fragen, z. B. an den Auftraggeber, die Nutzer oder Fachkräfte anderer Gewerke haben, schreiben Sie diese bitte zur Vorbereitung von Abstimmungsgesprächen auf.

Begründen Sie Ihren Lösungsvorschlag umfassend und detailliert.

Arbeits- und Hilfsmittel

Zur Bearbeitung der Aufgabenstellung sind alle schulüblichen Hilfsmittel, wie z. B. Tabellenbücher, Fachbücher, eigene Mitschriften und Taschenrechner, zulässig.

Abb. 1: Testaufgabe „Dachfenstersteuerung" aus den Hauptuntersuchungen

2 Raterschulungen

Es liegt auf der Hand, dass auch ein derartig qualifiziertes Ratingteam ohne eingehende Schulungsmaßnahmen nicht ohne Weiteres in der Lage ist, Bewertungen in Kompetenzmessverfahren vorzunehmen, die wissenschaftlichen Kriterien gerecht werden. Die Güte eines Messinstrumentes zur Messung beruflicher Kompetenz und Kompetenzentwicklung wird entscheidend dadurch geprägt, inwieweit die Bewertungen der Aufgabenlösungen der Probanden durch die einzelnen Beurteiler (Rater) übereinstimmen oder voneinander abweichen (Interra-

ter-Reliabilität). Damit die Rater als Grundlage und Bezugslinie ihrer Bewertungen ein gemeinsames Verständnis der Erwartungshaltung an die Aufgabenlösungen erreichen, wurden Raterschulungen durchgeführt. Das Konzept der Auftaktschulung umfasst ein zwölfstündiges Programm, welches zuerst mit Ratern in Bremen und Hessen realisiert wurde (vgl. GROLLMANN/ HAASLER 2009; HAASLER 2010). Dies fokussierte drei Kernpunkte:

1. Bewertungskriterien zum Rating
2. Testaufgaben
3. Ratingpraxis an empirischem Material

Für die Raterschulung wurde eine didaktisch aufbereitete Handreichung für die Teilnehmer erstellt, die im Schulungsseminar als Arbeitsunterlage diente.

Zu 1) Bewertungskriterien zum Rating

Eingangs wird eine Einführung in das Kompetenzmodell mit seinen Bewertungskriterien zur Beschreibung und Messung beruflicher Kompetenz positioniert. Dieser Impulsvortrag dient dazu, das Vorhaben in der aktuellen Diskussion zu verorten und die zentrale Frage zu beantworten, was man grundsätzlich zur Kompetenzmessung benötigt:

- einen Begründungsrahmen,
- ein Kompetenzmodell,
- das Testverfahren und
- das Auswertungsverfahren.

Eingehend werden vor allem die Bewertungskriterien vorgestellt, mit deren Hilfe das Rating der Lösungsvarianten erfolgt (vgl. RAUNER/ HAASLER/ HEINEMANN/ GROLLMANN 2009). Der Bewertungsbogen umfasst 40 Items, die zur Bewertung einer Lösungsvariante eines Probanden herangezogen werden. Da diese Items weder berufs- noch aufgabenspezifisch formuliert sind, kommt es in der Rater-Schulung vor allem darauf an, eine Bindung zum Kontext herzustellen (vgl. BECK 1980). Da das Bewertungsinstrument domänenspezifisch eingesetzt wird, muss verdeutlicht werden, wie jedes einzelne Item im Kontext des zugrunde liegenden Berufsbildes und des Erwartungshorizontes der einzelnen Testaufgaben interpretiert werden kann.

Zu 2) Testaufgaben

Zur Kompetenzmessung wurde ein Set von vier Testaufgaben entwickelt, mit denen die Probanden im Verlauf des Vorhabens konfrontiert werden. Es geht in der Rater-Schulung nicht darum, mit den Ratern einen Konsens über die Anlage und Ausgestaltung der Testaufgaben zu erreichen. Die Rater sind nicht die Experten für die Testaufgabenentwicklung, sie sollen vielmehr befähigt werden, ihre Rolle als Bewerter von Lösungsvarianten professionell und gewandt auszuüben. Die Bewertung von Lösungsvarianten der Testaufgaben erfordert ein tiefes Verständnis vom Berufsbild und den Anforderungen, die Facharbeit an die Fachkräfte im

Berufsalltag stellt. Die von den Probanden erarbeiteten Lösungsvarianten müssen vor diesem Hintergrund interpretiert und in den Kontext eingebettet werden. Einzelne Rater werden nicht als Spezialisten für das Rating von Teilaufgaben des Testaufgabensets vorbereitet. Jeder Rater wird als Bewerter für Lösungsvarianten eingesetzt, die aus dem gesamten Aufgabenset stammen; er muss also in der Lage sein, Ratings an allen vier Testaufgaben durchzuführen (vgl. HAASLER/ RAUNER 2010).

Zu 3) Ratingpraxis an empirischem Material

Um den Umgang mit dem Bewertungsbogen einzuüben und eine möglichst hohe Übereinstimmung der Beurteilungen der einzelnen Rater zu erreichen, wird methodisch die praktische Arbeit an empirischem Material favorisiert. Aus anderen Untersuchungen stammende Lösungsvarianten von Probanden wurden daher in Einzelarbeit und in Kleingruppen von den Ratern bewertet. Zunächst wurde jeder Rater in Einzelarbeit mit einer Testaufgabe und zugehörigen Lösungsvarianten dreier Probanden konfrontiert, die mit dem Bewertungsbogen einem unabhängigen Ad-hoc-Einzelrating unterzogen werden sollte.

Die Übereinstimmung der Bewerterurteile wies im Vorfeld der Raterschulung erwartungsgemäß vergleichsweise schlechte Reliabilitäten auf. Im Verlauf der Raterschulung wurde begleitend zur Arbeit am empirischen Material fortlaufend die Interrater-Reliabilität berechnet. So zeigt sich zeitnah die Wirkung der Schulung auf die Verbesserung der Übereinstimmung der Ratings (Näheres dazu siehe KOMET Band I).

Im Anschluss an das unabhängige Einzelrating im Ad-hoc-Verfahren wurde die Arbeit in Kleingruppen mit jeweils vier Ratern fortgesetzt. Die Gruppenarbeit thematisierte das zuvor realisierte Einzelrating. Jeder Rater beschrieb seine Herangehensweise an die Bewertung, seine Interpretation der Items, seinen Erwartungshorizont an die Aufgabenlösung und die Grundlage seiner individuellen Wertung einer Probandenlösung. Diese Reflexion in der Kleingruppe führte im diskursiven Prozess zu einer Nivellierung der Rater-Urteile. So wurde ein gemeinsamer Erwartungshorizont der Kleingruppe als Referenzniveau an die Lösungsvarianten der Testaufgaben entwickelt.

In der nächsten Phase der Raterschulung stellten die Kleingruppen im Plenum aller Rater ihren Diskussionsprozess, die Problemlagen und die gemeinsamen Verabredungen vor. Auch dieser Austausch führte zur weiteren Verfestigung und der Herausbildung eines gemeinsamen Verständnisses als Grundlage des Ratings.

Das dreischrittige Verfahren – Einzelrating, Kleingruppenreflexion, Plenum – wurde fortgesetzt, bis die vier Testaufgaben der beiden Sets vom gesamten Raterteam inhaltlich klar gefestigt verstanden waren und anhand von Probandenlösungen ein praktisches Rating durchgeführt worden war.

Neben der Herausbildung eines gemeinsamen Verständnisses der Rater ist auch der Aspekt der praktischen Rating-Erfahrung ein nicht zu unterschätzendes Ergebnis der Raterschulung. Die Konfrontation mit selbst durchzuführenden praktischen Ratings in der Raterschulung bil-

det bereits einen Vorgriff auf die später relevanten Ratings der Hauptuntersuchung, die jeder Rater in Einzelarbeit autonom in großer Anzahl vornehmen wird.

2.1 Absicherung der Interrater-Reliabilität als Ergebnis der Raterschulung

Die Rekrutierung von Ratern für ein derartiges Vorhaben ist ein nicht unproblematisches Unterfangen: Fachdidaktiker des Berufsfeldes, die im Rahmen eines methodisch anspruchsvollen Evaluationsdesigns in der Lage sind, als Rater tätig zu werden, sind in ihrer Verfügbarkeit eine rare Zielgruppe. Im Rahmen der hier vorgestellten Pilotprojekte konnten Lehrkräfte Berufsbildender Schulen durch Entlastung von ihrer wöchentlichen Unterrichtsverpflichtung zur Mitwirkung gewonnen werden. Da das Rating einer einzelnen Aufgabenlösung eines Probanden rund 15 Minuten Bearbeitungszeit für den Rater umfasst, werden in der Summe einige Arbeitstage für jeden Rater veranschlagt. Als Richtwert gilt folgende Überschlagsrechnung: Für das KOMET-Setting mit 100 Probanden, die jeweils 4 Testaufgaben bearbeiten, ergeben sich 400 Lösungsvarianten, die einem Rating unterzogen werden müssen. Da jede Probanden-Lösung im unabhängigen Doppelrating bewertet wurde, galt es folglich, 800 Ratings durchzuführen. Bei einem Bearbeitungszeitrahmen von 15 Minuten pro Rating ergeben sich summarisch 200 Stunden Arbeitsbelastung für das Raterteam.

Um den Nachweis der Interrater-Reliabilität auf eine solide Basis zu stellen, wurde im Anschluss an die Auftaktschulung der Rater vorab eine Stichprobe aus den Probandenlösungen der Hauptuntersuchung gezogen, die allen Ratern zur Bewertung vorgelegt wurde. Aus dem Set, bestehend aus den vier Testaufgaben, wurden jeweils zwei Probandenlösungen zum Rating herangezogen. Jeder Rater aus dem Team wurde folglich mit 8 Lösungsvarianten von Probanden konfrontiert, die zu durchdringen und zu bewerten waren.

Dieses Vorab-Rating bot die Möglichkeit, durch weitere Raterschulungen die Übereinstimmung der Bewerterurteile zu verbessern, falls dies erforderlich gewesen wäre. Erst als eine zufriedenstellende Interrater-Reliabilität durch das Vorab-Rating der Stichprobe nachgewiesen war, konnte die gesamte Datenbasis der Hauptuntersuchung zum Rating freigegeben werden. Die nachfolgende Abbildung zeigt den Interrater-Reliabilitätskoeffizienten, der im Testrating zum Abschluss der Auftaktraterschulung erreicht wurde, bevor das Raterteam die Ratings der Hauptuntersuchung vornahm. Erreicht wurde durch das Ratertraining eine akzeptable Interrater-Reliabilität, die die Ratingfähigkeit des Teams dokumentiert (siehe Tabelle 2).

Tabelle 2: **Interrater-Reliabilität im Anschluss an die Auftakt-Raterschulung (n=18 Rater, Teilprojekt Hessen)**

Proband	Aufgabe	Intra-Class-Koeffizient (ICC)
H0124	Dachfenstersteuerung	.852
H0265	Dachfenstersteuerung	.902
H0225	Signalanlage	.930
H0282	Signalanlage	.879
H0176	Trockenraum	.819
H0234	Trockenraum	.851
H0134	Kieselaufbereitungsanlage	.799
H0047	Kieselaufbereitungsanlage	.929

Es zeigte sich, dass die Koeffizienten ausnahmslos im Bereich hoher Reliabilität liegen, das für diese Untersuchung definierte Zielkriterium von 0.7 also erreicht bzw. überstiegen wird. Insgesamt können die Interrater-Reliabilitäten somit als zufriedenstellend bezeichnet werden (vgl. ASENDORPF/ WALBOTT 1979; SHROUT/ FLEISS 1979; WIRTZ/ CASPAR 2002). Ein Kernziel der Raterschulungen war somit erreicht. Details zur Auswahl des Reliabilitätskoeffizienten und der Berechnungen der Bewerterübereinstimmungen sind im Band I und Band II der KOMET-Publikationsreihe ausführlich dargelegt.

2.2 Nachschulung der Rater

Nach intensiver Ratingpraxis der Rater wurde eine Nachschulung als notwendig erachtet. Dieses von vornherein geplante „update“ diente der Absicherung der Rating-Qualität (Interrater-Reliabilitäten) für die weiter bevorstehenden Ratings. Das Ratingteam verstand die zweitägige Rater-Nachschulung als Absicherungs-Validierung des Ratings und des erwarteten Lösungsraums/Problemlösungshorizontes. Anlass für die Nachschulung waren prototypische Rating-Fehler, die sich im „Einschleichen“ einseitiger Rating-Tendenzen zur „Strenge“ oder „Milde“ dokumentierten. Vermutliche Ursachen dafür waren:

- zu positives Rating (Rating als „Lehrer“ mit sehr wohlwollendem didaktischem Verständnis für die Schüler) oder
- zu strenge Beurteilung (Lösungsraum gleichgesetzt/verwechselt mit idealtypischer Musterlösung, oft aus ingenieurwissenschaftlicher Perspektive).

Im Rating der Hauptuntersuchung, in dem jeder Rater des Teams rund 200 Ratings von Lösungsvarianten der Probanden aus allen vier Testaufgaben vornahm, wurde von der Wissenschaftlichen Begleitung Datenmaterial eingespeist, welches der Interrater-Reliabilitätskontrolle dient. Im Regelfall der Hauptuntersuchung wurden alle Probandenlösungen im unabhängigen Doppelrating bewertet. Zur Reliabilitätskontrolle der Ratings wurden zusätzlich 12 Probandenlösungen allen Ratern ins Material gemischt, die folgende Interrater-Reliabilitäten nach der Bewertung ergaben (siehe Tabelle 3):

Tabelle 3: **Interrater-Reliabilitäten im Rating der Hauptuntersuchung 2009 (Teilprojekt Hessen)**

Proband	Aufgabe	Intra-Class-Koeffizient
H0004	Dachfenstersteuerung	.843
H0006	Signalanlage	.859
H0008	Dachfenstersteuerung	.794
H0105	Signalanlage	.876
H0128	Trockenraum	.790
H0262	Kieselaufbereitungsanlage	.867
H0424	Kieselaufbereitungsanlage	.704
H0523	Trockenraum	.889
H0845	Dachfenstersteuerung	,839
H0850	Trockenraum	,812
H0865	Kieselaufbereitungsanlage	,781
H0866	Signalanlage	,779

Das Ergebnis zeigt, dass die Übereinstimmung der Raterurteile auf eine bedenkliche Tendenz des Interrater-Reliabilitätskoeffizienten hinweist. Werte, die nur noch eng am Zielwert 0.7 liegen, drohen, wenn der Trend sich fortsetzt, keine zufriedenstellenden Ratings in der Vergleichbarkeit der Rater-Urteile mehr zu bieten. Belastbare Ergebnisse der Kompetenzmessung sind damit nicht mehr erreichbar. Dieses erwartbare Ergebnis der Ratingpraxis zeigte, empirisch untermauert, deutlich die Notwendigkeit einer Raternachschulung an.

Der Ausweg aus den leicht auseinanderdriftenden Ratingtendenzen des Ratingteams wurde methodisch in der diskursiven Validierung des Erwartungshorizontes der Probandenlösungen in der Rating-Gruppe gesehen. Für die zweitägige Nachschulung des Ratingteams wurde daher ein Programm entwickelt, welches sich eng an die Auftaktveranstaltung der Raterschulung lehnt (siehe Tabelle 4):

Tabelle 4: **Ablaufplan der Rater-Nachschulung**

Arbeitsphase	Ort
Bilden von Arbeitsgruppen mit 4-5 Kollegen. Testrating in den Arbeitsgruppen in vier Schritten: • Jeder Rater bewertet die Lösung individuell. • Die Raterergebnisse werden miteinander in den Gruppen verglichen. Die unterschiedlichen Bewertungen werden analysiert. • Es wird ein Gruppenrating durchgeführt. Die Schwierigkeiten beim Finden gemeinsamer Bewertungen werden in einem Kurzprotokoll festgehalten. • Die Ergebnisse des Ratings **(individuell und Gruppe)** werden elektronisch zur Datenauswertung erfasst.	Arbeits-gruppe
• Präsentation der Rating-Ergebnisse der Arbeitsgruppen und der beim Rating aufgetretenen Schwierigkeiten. • Analyse aller Rating-Ergebnisse im Plenum. Dabei werden auffällige Ratingwerte (von einzelnen Ratern bzw. zu einzelnen Items) analysiert und diskutiert.	Plenum

Gearbeitet wurde wiederum an empirischem Material, welches diesmal aus der Hauptuntersuchung entnommen wurde. Als Hauptergebnis der Rater-Nachschulung bleibt festzuhalten, dass sich die Interrater-Reliabilität durch die diskursive Validierung wieder homogener zeigt als zuvor. Die Ratingfähigkeit des Teams konnte wieder hergestellt werden.

Als eine Erkenntnis der Ratingpraxis bleibt festzuhalten, dass selbst ein sehr erfahrenes Ratingteam mit langer Ratingpraxis nicht dauerhaft uneingeschränkt ratingfähig bleibt (in Bezug auf die Übereinstimmung der Bewerterurteile). Im Verlauf von Längsschnittuntersuchungen ist es daher ratsam, neben gründlichen Eingangstrainings des Ratingteams auch regelmäßige Nachschulungen einzuplanen, um über längere Zeiträume akzeptable Übereinstimmungen der Bewerterurteile sicherzustellen.

3 Raterschulung im Teilprojekt in China

Die Übertragung des KOMET-Settings auf internationaler Ebene in andere Kontexte beruflicher Bildung konnte in einem Teilprojekt in Peking erprobt werden (RAUNER 2009; RAUNER/ HEINEMANN u. a. 2009). Während Kompetenzmodell, Testaufgaben und Auswertungsroutinen von vornherein für einen derartigen Transfer entwickelt wurden, war klar, dass das Rating von einheimischen Experten in China durchgeführt werden muss. Eine Übersetzung der Lösungsvarianten chinesischer Probanden ins Deutsche und das Rating durch deutsche Rater wurde schnell verworfen.

Das Ratertraining am Institut für berufliche Bildung der Akademie für Erziehungswissenschaften in Peking (durchgeführt im April 2009) lehnte sich in der Grundstruktur an das Ratertraining der deutschen Rater (Hessen und Bremen) an. Mitwirkende bei der Schulung waren 35 chinesische Rater, ebenfalls Lehrer für Elektrotechnik aus der Berufsbildungspraxis. Als Arbeitsunterlage wurde das in Deutschland eingesetzte Manual ins Chinesische übersetzt.

Die Übersetzung der Testaufgaben, der Lösungsräume und der beispielhaft ausgewählten Lösungsvarianten von Probanden erwies sich als unproblematisch, da die Testaufgaben von beruflichen Aufgaben abgeleitet sind, die international beruflicher Praxis entsprechen. Beachtet werden mussten Differenzen bei technischen Normen und Vorschriften für den Betrieb und die Einrichtung elektrischer Anlagen. Kulturelle Unterschiede spielen dagegen in diesem Feld der Technik und der beruflichen Arbeitspraxis keine nennenswerte Rolle.

Im Rahmen eines einwöchigen Vorbereitungsseminars in Peking im Dezember 2008 zu diesem international vergleichenden Kompetenzerhebungsprojekt wurde unter anderem mit einer Arbeitsgruppe von Lehrern und Fachleitern der Fachrichtung Elektrotechnik/Elektronik die curriculare und berufliche Validität der vier Testaufgaben diskutiert. Drei der vier Testaufgaben wurden ohne größeren Diskussionsbedarf als valide eingestuft. Eine der vier Aufgaben löste dagegen eine längere Diskussion aus, da die Aufgabe in der Situationsbeschreibung aus der Sicht der chinesischen Lehrer ein fachfremdes Element enthielt. Dieses bezog sich auf die Wärmedämmung der Wände eines Raumes, in dem eine elektrische Heizung eingerichtet

werden soll. Eine Erläuterung der Aufgabenstellung unter expliziter Bezugnahme auf das Kompetenzmodell führte schließlich zur Zustimmung zu dieser Testaufgabe. Auslöser für die zunächst abweichende Einschätzung der beruflichen Validität dieser Aufgabe war die aus der Sicht der chinesischen Lehrer fachfremde Einbeziehung der Wärmedämmung in die Aufgabenstellung. Dies bestätigt die Erfahrung, dass im Kontext einer berufsfachschulischen Ausbildung die Fachperspektive stärker gewichtet wird als die Berufsperspektive. Die Einbeziehung des Lernfeldkonzeptes in diese Diskussion erleichterte die Überbrückung dieser Differenzen, die sich unter Berücksichtigung des Aspekts der Arbeits- und Geschäftsprozessorientierung in der Regel als Scheinprobleme herauskristallisieren.

Das Kompetenzmodell und die davon abgeleiteten Items zum Rating fanden engagierte Zustimmung sowohl bei den chinesischen Lehrern als auch bei den Experten der Bildungsverwaltung und Bildungsforschung. Dieses zunächst überraschende Ereignis liegt vermutlich darin begründet, dass die Konzepte zur entwicklungslogischen Systematisierung beruflicher Curricula und beruflicher Bildungsprozesse sowie die Konzepte Arbeitsprozesswissen und gestaltungsorientierte Berufsbildung mittlerweile auch in China zum Standardrepertoire der Fortbildung von Berufsschullehrern im Bereich der Curriculumentwicklung gehören.

3.1 Ablauf der Raterschulung in Peking

Auch das Ratertraining in China wurde so aufgebaut, dass schrittweise das typische Bewerten von Schülerlösungen (bei offenen Testaufgaben) in der Berufsbildungspraxis abgelöst wird durch ein standardisiertes Rating auf der Grundlage von Lösungsräumen für jede Testaufgabe sowie durch eine Liste von Items, mit denen die Kompetenzkriterien operationalisiert werden. Für die Aneignung von Raterkompetenz ist entscheidend, dass in den Arbeitsgruppen (während der Raterschulung) die individuellen Bewertungen Item für Item miteinander verglichen und Abweichungen analysiert werden. Daran schließt sich ein Gruppenrating an, bei dem die Gruppenmitglieder einen Konsens für jedes Item erzielen müssen. Die Schwierigkeiten, die dabei auftreten, werden in einem Kurzprotokoll festgehalten. Daran schließt sich – gegebenenfalls – das Rating einer zweiten Lösungsvariante eines anderen Probanden an. Durch diese Form der Gruppenarbeit wird ein gemeinsames Verständnis der Items und ihrer Anwendung bei der Bewertung von Lösungen erreicht.

Die Plenumssitzungen haben das Ziel, von Anfang an Differenzen, die sich zwischen den Gruppenratings ergeben, zu analysieren, um damit letztendlich eine ausreichende Interrater-Reliabilität zu erreichen. Die Präsentation der Ratingdaten und ihre statistische Auswertung ist dabei von grundlegender Bedeutung, da das individuelle Ratingverhalten für jeden Rater im Kontext aller anderen Rater transparent wird. Das Einblenden der durchschnittlichen Ratingwerte zu den einzelnen Items sowie des Gesamtdurchschnittswertes von ausgebildeten Ratern dient als ein Referenzwert, an dem Abweichungen gespiegelt und im Fortgang des Ratings korrigiert werden können.

Das abschließende Rating von Lösungsvarianten aller vier Testaufgaben auf der Grundlage der erworbenen Raterkompetenz hat die Funktion

1. den letztlich erreichten Wert der Interrater-Reliabilität zu ermitteln,
2. den ausgebildeten Ratern zurückzumelden, wie sie im Vergleich zu anderen Ratern gewertet haben und ob es noch auffällige Abweichungen gibt und
3. die erreichte Ratingqualität mit der anderer Ratergruppen zu vergleichen (z. B. in Form einer Interrater-Reliabilität, die auch in internationalen Projekten erreicht werden muss).

Die Raterergebnisse zur ersten und zweiten Schülerlösung (Testaufgabe 1) weichen charakteristisch von den Ratings der ausgebildeten deutschen Rater ab. In der Tendenz wird die erste Lösungsvariante um einen Punktwert besser bewertet als von den deutschen Raterteams. Die Gruppenauswertung der ersten Schülerlösung hat bereits während der ersten Gruppensitzung zu Einsichten über stark abweichende Bewertungen einzelner Rater bzw. zu spezifischen Items geführt. Dieser „Lerneffekt“ schlägt sich beim Bewerten der zweiten Lösungsvariante eines Probanden in einem leicht höheren Koeffizienten nieder. Beim Rating der zwei Lösungen zur ersten Testaufgabe haben sich die chinesischen Lehrer in ihrer großen Mehrheit unmittelbar dem Rating zugewandt – ohne sich vorab mit dem Lösungsraum zu beschäftigen und entgegen der Empfehlung, die dazu bei der Einführung in das Ratingverfahren im Plenum gegeben wurde.

Als Ursache für dieses Verhalten stellte sich heraus, dass die Teilnehmer überwiegend bereits am „Lehrertest“ teilgenommen hatten und daher mit den Testaufgaben sehr gut vertraut waren. Daraus wurde – implizit – ihr eigenes Testverhalten zum Maßstab für die Bewertung der Schülerlösungen. Der Lösungsraum wurde als eine überflüssige fachliche Hilfestellung eingestuft und beim Rating weitestgehend außer Acht gelassen. Ihr Rating basiert daher auf ihrer subjektiven und individuellen Lehrerkompetenz und einem entsprechenden Problemlösungshorizont. Es ist naheliegend, dass sich dieser von dem ihrer Schüler/Studenten allenfalls graduell unterschied. Im Ergebnis fielen daher die Bewertungen recht positiv aus. Das durch ihre Lehrerpraxis geprägte implizite Kompetenzmodell reduziert vermutlich die berufsfachliche Kompetenz im Wesentlichen auf kontextfreies fachkundliches und fachtheoretisches Wissen. Die Heterogenität der Ratings wurde ebenfalls dadurch verstärkt, dass die objektivierende Funktion, die dem Lösungsraum zukommt, nicht zur Wirkung kam. Anhand der Auswertung der Einzel- und Gruppenratings im Plenum unter Bezugnahme auf die projizierten Ratingtabellen wurden das Raterverhalten und seine Ursachen von den Teilnehmern selbst analysiert. Die Ratinganalyse der ersten Raterrunde wurde abgeschlossen mit der Verabredung, die zu jeder Testaufgabe entwickelten Lösungsräume grundsätzlich dem Rating zugrunde zu legen, da es sich um ein wesentliches Element der Standardisierung und Objektivierung des Testverfahrens handelt.

Das Rating der zwei Schülerlösungen zur zweiten Testaufgabe unterschied sich auffällig vom ersten Rating. Es herrschte insgesamt ein spürbar höherer Grad an Konzentration, und die Regel, während des individuellen Ratings sich nicht mit anderen Gruppenmitgliedern auszutauschen, wurde konsequent eingehalten. Alle Rater studierten zunächst den vorliegenden

Lösungsraum zur Testaufgabe und nahmen ihn noch während des Ratings gelegentlich zur Hilfe.

Die Präsentation der Rating-Ergebnisse der zweiten Raterrunde, einschließlich des Vergleichs zu den Ratings des deutschen Raterteams, löste eine Überraschung aus, da die Ergebnisse der chinesischen Rater jetzt tendenziell signifikant niedriger lagen: Die Rater hatten jetzt „strenger“ bewertet als ihre deutschen Kollegen. Zugleich nahm der Grad an Übereinstimmung beim Rating deutlich zu. Die Interrater-Reliabilitäten erreichten bereits nach dieser zweiten Raterrunde die relativ hohen Werte von .80/.75 (siehe Tabelle 5). Aufzuklären war bei der Auswertung im Plenum der Effekt der zu strengen Bewertung – verglichen mit den Bewertungen erfahrener Rater aus Deutschland.

Tabelle 5: **Entwicklung des Interrater-Reliabilitäts-Koeffizienten im Verlauf der Raterschulung in Peking (n=35 Rater)**

Probanden-Code	Testaufgabe	Tag 1	Tag 2 vormittags	Tag 2 nachmittags	Tag 3 vormittags	Tag 3 nachmittags
		Intra-Class-Koeffizient (ICC)				
H0282	Signalanlage	.41				.82
H0225	Signalanlage	.54				.79
H0176	Trockenraum		.80			.84
H0234	Trockenraum		.75			.80
H0265	Dachfenster-steuerung			.84		.82
H0102	Dachfenster-steuerung			.82		.83
H0336	Kieselaufbereitungsanlage				.86	.85
H0047	Kieselaufbereitungsanlage				.79	.79

3.2 Missverständnisse um den Lösungsraum

Als Ursache für das „zu strenge“ Rating wurde ein Missverständnis identifiziert, nach dem die Mehrheit der Rater den Lösungsraum jetzt als eine idealtypische und maximale Lösung gehandhabt hat. Dem Statement eines Teilnehmers, wonach der Lösungsraum dazu verführe, „auf alle Lösungsaspekte zu achten, die er beinhaltet“, stimmten viele Teilnehmer zu. Die Handhabung des Lösungsraumes als Maßstab und Referenzsystem zur Bewertung der Schülerlösungen hatte sich implizit eingestellt, da bei der Einführung des Testkonzeptes (Eingangsplenum) auf die Möglichkeit dieses Missverständnisses nicht ausdrücklich hingewiesen worden war. Möglicherweise trug auch die Übersetzung des Textes zu einer gewissen Unschärfe bei der Beschreibung der Funktion des Lösungsraums bei. Die Aufklärung dieses Missverständnisses anhand der Raterdaten löste auch bei diesem kritischen Punkt ein „Aha-Erlebnis“ aus. Dass der Lösungsraum für die Testaufgaben möglichst vielfältig zu allen Lösungskriterien Lösungsaspekte und -möglichkeiten zusammenstellt, bedeutet, dass die

Lösungsräume prinzipiell 1. unvollständig sind und 2. auch sehr gute und vollständige Einzellösungen immer nur eine Teilmenge der im Lösungsraum zusammengestellten Lösungsmöglichkeiten umfassen. Diese Definition wurde von den Teilnehmern der Raterschulung als unmittelbar einsichtig angenommen.

Mit dem abschließenden wiederholenden Rating der jeweils zwei Lösungen zu den vier Testaufgaben wurde das Ziel verfolgt, die eingeübte Raterkompetenz zu stabilisieren. Wie zu erwarten, stellte sich am dritten Tag der Raterschulung ein messbar professionelles Raterverhalten ein (siehe Tabelle 5). Der sehr hohe Wert für die Interrater-Reliabilität zeigt, dass in der Summe alle Teilnehmer das Ziel erreichten, anstelle ihrer subjektiven Bewertungsmaßstäbe das objektivierende und standardisierte Ratingverfahren ausreichend sicher anzuwenden.

4 Fazit der Raterschulung in Peking und Transferfähigkeit des Schulungskonzeptes

Die chinesischen Lehrer bewerteten beim ersten Proberating im Rahmen der Raterschulung die Schülerlösungen zur ersten Testaufgabe, ohne sich vorher mit dem vorliegenden skizzierten Lösungsraum der Testaufgabe auseinander zu setzen. Bei der Bewertung der Lösungen dominierte daher ihr eigener Problemlösungshorizont, der vor allem durch ihre subjektive Unterrichtspraxis als Lehrkraft geprägt war. Dieser unterschied sich – wie zu erwarten – kaum von dem ihrer Schüler. Das objektivierende Moment der Bewertung stellen die insgesamt 40 Bewertungsitems sowie der jeweilige Lösungsraum dar. In der Summe fielen die Bewertungen der Kollegen in Peking daher deutlich besser aus als die Ratingergebnisse der deutschen Rater. Außerdem variierten die Ergebnisse im Auftaktrating stark.

- Die Konfrontation der Lehrer mit ihren relativ hohen und sehr unterschiedlichen Ratingwerten sowie der Vergleich mit den Bewertungen der deutschen Rater (im Sinne eines Referenzsystems) trugen erheblich zu der Einsicht bei, dass das Ratingverfahren voraussetzt, sich mit dem Lösungsraum für die offenen Testaufgaben gründlich auseinander zu setzen und die Lösungsräume bei der Bewertung der Schülerlösungen konsequent zu berücksichtigen. Damit soll der subjektive Problemlösungshorizont der Lehrer als Bewertungsmaßstab abgelöst werden durch ein standardisiertes Ratingverfahren.

- Beim Rating der Aufgabenlösungen zur zweiten Testaufgabe nahm die Übereinstimmung der Raterergebnisse sprunghaft zu. Zugleich fiel die Bewertung der Rater jetzt insgesamt erkennbar strenger aus als beim ersten Proberating. Beim Rollenwechsel „vom Lehrer zum Rater“ wurde nun der Lösungsraum in der Tendenz als eine vollständige Beschreibung der Lösung der Testaufgaben interpretiert. Dem liegt das Missverständnis zugrunde, dass es sich beim Lösungsraum um eine idealtypische vollständige Lösung handele. Dieses Missverständnis konnte anhand der Bewertungs-

ergebnisse der Arbeitsgruppen und der einzelnen Rater veranschaulicht und aufgeklärt werden.

- Beim Rating der Schülerlösungen zur dritten und vierten Testaufgabe stellte sich ein professionelles Rating mit entsprechend hohen Werten für die Interrater-Reliabilität ein. Die Werte liegen sogar über denen der deutschen Rater. Das kann vermutlich auch darauf zurückgeführt werden, dass im Ratertraining abschließend noch einmal Lösungsvarianten bewertet wurden, die im Verlauf der dreitägigen Veranstaltung schon Gegenstand der Arbeit waren. Unterstützend wirkte sich vermutlich auch die statistische Auswertung der Raterergebnisse unmittelbar im Anschluss an das Rating in den Arbeitsgruppen aus, da auf der Grundlage der tabellarischen Übersichten das Raterverhalten detailliert analysiert werden konnte. Die Möglichkeit, die sich für jeden Rater bot, seine Ratings mit denen aller anderen Rater detailliert zu vergleichen, stieß auf großes Interesse und führte – wie erhofft – schrittweise zu einer kontinuierlichen Anhebung der Qualität der Ergebnisse.

- Die Diskussion in den Arbeitsgruppen sowie im Plenum über auffällige Ratingergebnisse und die jeweiligen Ursachen trug erheblich zur Qualifizierung der Rater bei.

- Den drei aufeinanderfolgenden Schritten des Schulungskonzeptes – Einzelarbeit, Gruppenauswertung und Plenumsdiskussion, auf der Grundlage der eigenen sowie externen Raterergebnisse – kommt eine zentrale methodische Bedeutung im Rahmen der Raterqualifizerung zu.

Das in China eingesetzte Konzept der Raterschulung zeigt, dass auch Raterteams, die nicht an der Entwicklung der Testaufgaben oder des Ratingverfahrens – wie im Pilotprojekt in Deutschland – beteiligt waren, mit dem Schulungskonzept zu Ratern qualifiziert werden können, deren Bewertungen wissenschaftlichen Anforderungen an Kompetenzmessungen genügen.

Literatur

ASENDORPF, J./ WALLBOTT, H. G. (1979): Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. Zeitschrift für Sozialpsychologie, H. 10/1979, 243-252.

BECK, K. (1980): Die Bedeutungsüberschneidung von Beschreibungskategorien als Problem der Unterrichtsforschung – Eine methodenkritische Untersuchung am Beispiel des Ratingverfahrens. Forschungsbericht Nr. 6. Otto-Selz-Institut für Psychologie und Erziehungswissenschaft. Universität Mannheim.

GROLLMANN, P./ HAASLER, B. (2009): Berufliche Kompetenzentwicklung als Maßgabe für die Qualität beruflicher Bildung – Vorstellung eines Instruments. In: MÜNK, H.-D./ WEIß, R. (Hrsg.): Qualität in der Beruflichen Bildung – Forschungsergebnisse und Desiderata. Bielefeld, 69–89.

GRUSCHKA, (1985): Wie Schüler Erzieher werden. Wetzlar.

HAASLER, B. (2010): Berufliche Kompetenzen angehender Elektroniker: Zwischenergebnisse zur Kompetenzdiagnostik aus einem Schul-Modellversuch der Bundesländer Bremen und Hessen. In: BECKER, M./ FISCHER, M./ SPÖTTL, G. (Hrsg.): Von der Arbeitsanalyse zur Diagnose beruflicher Kompetenzen. Frankfurt/M., 177–193.

HAASLER, B./ RAUNER, F. (2010): Messen beruflicher Kompetenz: Konzept einer Large-Scale-Untersuchung und erste empirische Ergebnisse. In: MÜNK, H.-D./ SCHELTEN, A. (Hrsg.): Kompetenzermittlung für die Berufsbildung – Verfahren, Probleme und Perspektiven im nationalen, europäischen und internationalen Raum. Bielefeld, 77–99.

RAUNER, F. (2009): 800 chinesische Auszubildende nehmen am KOMET-Projekt teil. In: Zeitschrift für Berufs- und Wirtschaftspädagogik (ZBW), Heft 2/2009, Stuttgart, 330–331.

RAUNER, F./ HAASLER, B./ HEINEMANN, L./ GROLLMANN, P. (2009): Messen beruflicher Kompetenzen – Band I: Grundlagen und Konzeption des KOMET-Projektes. Münster.

RAUNER, F./ HEINEMANN, L./ PIENING, D./ HAASLER, B./ MAURER, A./ ERDWIEN, B./ MARTENS, T./ KATZENMEYER, R./ BALTES, D./ BECKER, U./ GILLE, M./ HUBACEK, G./ KULLMANN, B./ LANDMESSER, W. (2009): Messen beruflicher Kompetenzen – Band II: Ergebnisse KOMET 2008. Münster.

SHROUT, P. E./ FLEISS, J. L. (1979): Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, Heft 86(2)/1979, 420–428.

WIRTZ, M./ CASPAR, F. (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen.

Zitieren dieses Beitrages

HAASLER, B./ MAURER, A. (2011): Bewertung von Lösungen gestaltungsoffener Testaufgaben zur Messung berufsfachlicher Kompetenzen: Möglichkeiten und Schwierigkeiten einer internationalen Vergleichbarkeit. In: *bwp@* Spezial 5 – Hochschultage Berufliche Bildung 2011, Fachtagung 08.1/2, hrsg. v. SCHWENGER, U./ HOWE, F./ VOLLMER, T./ HARTMANN, M./ REICHWEIN, W., 1-16. Online: http://www.bwpat.de/ht2011/ft08/haasler_maurer_ft08-ht2011.pdf (19-11-2011).

Die AutorInnen:

Prof. Dr. BERND HAASLER

Pädagogische Hochschule Weingarten

Kirchplatz 2, 88250 Weingarten

E-Mail: haasler@phweingarten.de

Homepage: http://bernd-haasler.blogspot.com/

ANDREA MAURER

Universität Bremen

Leobener Straße/NW 2, 28359 Bremen

E-Mail: amaurer@uni-bremen.de

Homepage: http://www.ibb.uni-bremen.de/