Bericht: Fifth Ottawa International Conference on Assessment of Clinical Competence in Dundee 1992

R. Busse und Ch. Schmidt, Hannover und Berlin

Zusammenfassung: Der folgende Bericht schildert persönliche Eindrücke über eine internationale Tagung, die sich mit Prüfungsverfahren in Hinblick auf klinische Kompetenz befaßte. Unter anderem wird auf Neuerungen zu "objective structured clinical examinations", Simulationspatienten und MC-Prüfungen eingegangen und ein Überblick über gegenwärtige Tendenzen gegeben.

Summary: Personal impressions from the Fifth Ottawa Conference on Clinical Assessment are given. Recent developments of e. g. objective structured clinical examinations, multiple choice questions, and simulated patients are summarized, and present trends characterized.

1. Kurze Beschreibung der Tagung

Die Konferenz war die fünfte internationale Konferenz über die Beurteilung klinischer Kompetenzen. Diese Konferenzen fanden ursprünglich in Ottawa (daher der Name) unter der Leitung des dort tätigen Ian Hart statt. Inzwischen wechselt der Austragungsort der zweijährlich stattfindenden Konferenz zwischen Nordamerika und Europa: nach Dundee 1992 wird sie 1994 in Toronto und 1996 in Maastricht stattfinden.

Diesmal war sie übrigens kombiniert mit den Jahrestagungen von ASME (Association for the Study of Medical Education) und AMEE (Association for Medical Education in Europe), die dabei beschloß, sich von einem reinen Dachverband in einen Verband mit nationalen, institutionellen und individuellen Mitgliedern umzuwandeln.

Die Teilnehmerzahl der Ottawa-Konferenzen ist in der siebenjährigen Geschichte kontinuierlich gestiegen; diesmal nahmen über fünfhundert Personen aus knapp fünfzig Ländern teil. (Weitere Anmeldungen mußten abgelehnt werden!) Am meisten vertreten waren dabei die USA, Kanada, die Niederlande, Großbritannien und Schweden. Die Stellung der Ausbildungsforschung in Deutschland wurde durch seine Teilnehmerzahl eindrucksvoll dokumentiert: vier (die Verfasser eingeschlossen).

Die in Klammern angegebenen Literaturangaben beziehen sich entweder auf die zweibändigen Proceedings der Konferenz [HARDEN RM, HART IR, MULHOLLAND H (eds.): Approaches to the Assessment of Clinical Competence. Dundee, 1992], nachfolgend AA abgekürzt, oder auf das Januarheft von "Medical Education" [1993, Vol. 27, No. 1], abgekürzt ME, in dem die Abstracts der im Rahmen des ASME-Members' Day gehaltenen Vorträge abgedruckt wurden.

2. Bericht über einzelne Veranstaltungen und Themenkomplexe

OSCE: Eine der wesentlichen Neuerungen im Prüfungswesen ist die Einführung von nach Lernzielen strukturierten objektiven klinischen Prüfungen (engl. "objective structured clinical examinations", OSCE). Seit längerer Zeit wird an einer zunehmenden Zahl von medizinischen Fakultäten dieses Prüfungsmodell eingesetzt, bei dem Prüflinge einen Parcours von Stationen durchlaufen, an denen sie theoretische Kenntnisse und praktische Fertigkeiten demonstrieren. Dabei werden sie anhand vorab erstellter Checklisten beurteilt. Eine Reihe von Beiträgen setzte sich mit Erfahrungen und Weiterentwicklung dieses Ansatzes auseinander.

Relativ lange Erfahrungen mit OSCE wurden in der Chirurgie an der Universität von Dundee gesammelt. In fünfzehn Jahren ist dort aus besonders validen und reliablen Aufgaben eine "Bank" erstellt worden. Jedes Jahr werden zwei neue Aufgaben getestet. Ethische Fragen sollen miteingeschlossen werden [PREECE et al. AA: 163].

Am gleichen Ort werden beispielhaft verschiedene Variablen in die Prüfung eines Chirurgie-Blockpraktikums integriert: Bei der ersten Station geht es um praktische Fertigkeiten und die Problemerkennung, was anhand von Anamnese, und körperlicher Untersuchung erfaßt wird. Anschließend gibt es für die Prüflinge eine Rückkoppelung durch Bekanntgeben der korrekten "Antworten". Bei der zweiten Station geht es um das Erstellen einer Hypothese für die Diagnose. Anschließend wiederum Bekanntgeben der korrekten "Antworten". Nach einer Vorbereitungsphase wird bei der dritten Station über die Ursachen des Krankheitsbildes geprüft. Anschließend wie zuvor Rückkoppelung, dann bei einer vierten Station Erstellung eines Plans für das weitere diagnostische und therapeutische Vorgehen durch die Prüflinge sowie schließlich wieder Bekanntgeben der korrekten "Antworten". Durch die jeweilige Rückkoppelung werden Fehler schnell korrrigiert und es kann jede Station unabhängig gemessen werden. So können Defizite einzelner Komponenten erfaßt werden, die bei einer einzelnen Station verloren gehen würden [COLLINS, RICHARDSON. AA: 669ff].

Eine andere Untersuchung beschäftigte sich mit der Prüfung des Lernziels, Patienten in der Notaufnahme herauszufinden, die aufgenommen bzw. nicht aufgenmmen werden müssen. Untersucht wurde dies bei Assistenzärzten in der Inneren Medizin, Chirurgie, Allgemeinmedizin und Pädiatrie, wobei OSCE als Prüfungsmethode eingesetzt wurde. Drei Papierfälle mußten begründet in eine Rangfolge gebracht werden. Problematisch waren dabei die zeitintensive Prüfung (ca. 1-2 Stunden pro Assistenzarzt), der zuvor nicht einheitlichen Unterricht und die fehlende direkte Vergleichsmöglichkeit mit früheren Jahrgängen [BENITEZ, ALLISON. AA: 153ff].

Logistische Probleme entstehen dadurch, daß gleichzeitig mehrere OSCE-Stationen erforderlich sind und ferner durch die kurze "Halbwertzeit" der Assistenten als Prüfende von Studierenden [ROSLANI et al. 141ff].

Daher wurde vorgeschlagen, Studierende höherer Semester als Prüfende in OSCEs einzusetzen: Dabei wurden ähnliche Prüfungser-

gebnisse wie bei Lehrenden als Prüfern erzielt, die Akzeptanz bei Geprüften war gut und zudem die Kosten geringer; allerdings fühlten sich 10% der Geprüften unfair beurteilt und 15% lehnten Studierende als Prüfende ab [GLEESON et al. AA: 659ff].

"Standardized Patients" (SP): SP sind Laien oder Schauspieler, mit oder ohne besondere körperliche Besonderheiten, die trainiert werden, um als Ausbildende und Prüfende eingesetzt werden zu können. Unterschiedliche Gruppen können als SP eingesetzt werden:

1. Die Vorteile von tatsächlichen, nicht besonders ausgebildeten Patienten sind leichte Verfügbarkeit, keine Kosten, kein Trainingsaufwand und Realitätsnähe. Nachteile schließen begrenzte Anzahl, unvorhersagbares Verhalten und Sprachbarrieren ein. Ferner wird von Patienten der Einsatz als SP möglicherweise als Belästigung empfunden.

2. Wenn tatsächliche Patienten verwandt werden, die jedoch für die Arbeit als SP ausgebildet werden, so bieten sie echte klinische Symptome, geringere Sprachbarrieren, leichte Verfügbarkeit und geringe Kostenintensität. Jedoch bleiben die Einsatzmöglichkeiten begrenzt, Belästigung mag empfunden werden, Verhalten ist variabel, klinische Symptome bleiben nicht immer stabil.

3. Setzt man kurz eingewiesene Nicht-Patienten ein, so ist ein breites Spektrum von Szenarien möglich, die SP sind leicht verfügbar, man bekommt bessere Rückmeldung über die Leistung der Studierenden und belästigt keine Patienten.

4. Gut ausgebildete Nicht-Patienten bieten breite Einsatzmöglichkeiten, Konsistenz im Verhalten und ausgzeichnete Rückkoppelung. Nachteile sind höhere Kosten und Probleme bei der Sicherstellung von adäquatem Personal und Training. [COLLINS. AA: 24ff; COHEN et al. AA: 274ff, 279ff].

Wer sind diese Standardized Patients? Folgende Charakteristika wurden an einer an mehreren Hochschulen in den USA durchgeführten Untersuchung gefunden [KACHUR et al. AA: 242ff]:

* mittleres Alter: 42.4 Jahre
* alleinlebend/verheiratet/geschieden: 24/29/8
* mittlere Kinderzahl: 1.3
* Bildung (nur High School/ College): 14/40
* SP-verwandte Tätigkeit: Schauspiel 49%, Lehre 74%
* mittlere Zahl der Krankenhausaufenthalte: 2.5
* häufigster Grund für Krankenhausaufenthalt: Entbindung
* letzter Krankenhausaufenthalt: durchschnittlich vor 5 Jahren
* durchschnittliche Zahl der Arztbesuche im letzten Jahr: 2.4
* häufigster Grund für Arztberuf: Routineuntersuchung

Der Bericht einer Referentin, die als SP und als SP-Trainerin arbeitet, begann etwa so: "Vor etwa sechs Jahren suchte ich einen interessanten Teilzeitjob und stieß auf ein Flugblatt, mit dem Interessenten für ein SP-Training gesucht wurden. Die SP sollten zur Beurteilung der Interaktion und Untersuchungstechnik von Medizinstudierenden eingesetzt

paganda. Die meisten SP beteiligen sich für eine lange Zeit an dem Programm. Das Programm wurde inzwischen schon exportiert und z.B. in China ein entsprechendes Programm mit aufgebaut worden [STANLEY. AA: 27ff].

In der Ausbildung werden z.B. auch Kleinkinder mit ihren Müttern als SP eingesetzt. Ziel ist die Verbesserung der Anamnese- und Interaktionsfähigkeit. Nach zwanzig-minütigen Gesprächen, die von einem Mentor auf Video aufgenommen werden, füllen Studierende einen Fragebogen aus, in dem es um das weitere Procedere mit den Patienten geht. Die Mütter füllen zwei Evaluationsbögen aus, einen mit Bezug zur Anamnese, einen mit Bezug zur Interaktionsfähigkeit. Außerdem wird von Studierendem und Mentor das Videoband analysiert. Sowohl Studierende als auch Mentoren profititieren von dieser Methode [COHEN et al. AA: 274ff].

Beim Einsatz von SP in Prüfungen wurde nur eine geringe Korrelation von MC-Prüfungsergebnissen gemessen [SWARTZ et al. AA: 261ff]. Der Einsatz von SP wird auch vom National Board of Medical Examinors (NBME), der weltgrößten medizinischen Prüfungsbehörde, getestet [KLAAS et al. AA: 58].

Seit November 1991 gibt es auch eine internationale Organisation (Advancement of Standardized Patients In Research and Education = ASPIRE), die an der Mount Sinai School of Medicine in New York angesiedelt ist und aus der auch die meisten Beiträge zu SP kamen. Ziele sind u. a. Hilfe bei der Entwicklung von SP-Programmen, Austausch

werden. Ich bewarb mich, wurde zum Vorstellungsgespräch eingeladen und schließlich in das Trainingsprogramm aufgenommen.

Seitdem durchlebte ich schwere Episoden mit Palpitationen, lebte rücksichtslos als Kokainabhängige, trug mehrere uneheliche Kinder aus, bin regelmäßig Selbstmörderin und hatte während der letzten vier Jahre mehrmals pro Monat akute Appendizitis."

SP müssen ihren Fall wie eine zweite Identität kennen; darüber hinaus müssen SP sich sehr gut das Anamnese-Gespräch und die körperliche Untersuchung merken, um anschließend adäquat beurteilen zu können. Der Körper dient als Modell, zum Teil auch für invasivere Untersuchungen (rektale Untersuchung, Untersuchung des Beckens). Die Bezahlung hängt unter anderem auch von der Invasivität der Untersuchungen ab. Auch Gefahren sind mit der Arbeit verbunden: Als die Referentin einmal eine Patientenrolle als Prostituierte zu spielen hatte und den Klinikhof zu überqueren hatte, näherte man sich ihr zweimal ihrer Rolle entsprechend. Maximal können pro SP und Fall ca. 5 bis 10 Studierende verkraftet werden, nicht mehr als 10 Studierende in zwei Tagen. Abgesehen von der körperlichen Belastung verschwimmen bei zu hoher Zahl Fälle bzw. Anamnesegespräche und Erinnerungen an die einzelne körperliche Untersuchung.

SP ist kein Vollzeitberuf. Dementsprechend schwer ist es, junge SP für die Arbeit während des Tages zu bekommen. Die Rekrutierung erfolgt in erster Linie über Anzeigen auf dem Campus und über Mund-zu-Mund-Pro-

von Fällen, Entwicklung von Standards [SWARTZ. AA: 280ff].

Medical Record Audit: In einem weiteren Versuch notierten Studierende im letzten Studienjahr nach Begegnung mit einem SP in einer Kurve Untersuchungsergebnisse und weitere diagnostische bzw. therapeutische Pläne. Zusätzlich wurde jede Begegnung mit Videoband aufgezeichnet und anhand von Checklisten Kurven und Videobänder ausgewertet. Durchschnittlich wurden etwa 50% der Items auf der Checkliste bei Auswertung des Videobandes erreicht, dagegen nur 25% bei der Auswertung der Kurven. Weniges, was im Video nicht wahrgenommen wurde, war in der Kurve verzeichnet. Es wird angenommen, daß zum einen mit Medical Record Audit eine andere Konstellation von Fähigkeiten geprüft wird, andererseits die Methode kein Ersatz für direkte Beobachtung ist [CASE et al. AA: 471ff].

Multiple Choice Questions (MCQ): Im Hinblick auf MCQ-Prüfungen wird dem "Mustererkennen" von einer unbekannter Anzahl richtiger Antworten in einem Angebot von etwa zwanzig Antworten der Vorzug gegenüber konventionellen MC-Aufgaben (Auswahl von einer oder mehr Antworten aus Angebot von fünf Antworten) eingeräumt. Ein Einsatz dieses Fragetyps wird für die Verwendung im NBME geprüft. Der Zeitaufwand für Prüfung und Auswertung ist gering. Ein vergleichender Test bei Praktikanten vor und nach Abschluß des Praktikums zeigte, daß übliche MC-Fragen eine deutlich höhere Quote von falschen Antworten und vielen nicht sinnvollen Ergebnissen erbrachten. Demgegenüber ist bei den Musteraufgaben eine deutliche Verbesserung der Erkennung von wesentlichen Mustern als Funktion der zunehmenden klinischen Erfahrung nach Abschluß des Praktikums nachweisbar [CASE et al. 452ff, 459ff].

Didaktischer Unterricht für Wissenschaftler in Medizinischen Grundlagenfächern: Analysiert man Herangehensweisen von erfolgreichen Forschern und vergleicht diese mit der von erfolgreich Lehrenden, so ergeben sich deutliche Parallelen: Im Planungsstadium geht es um Ausbildung einer breiten Wissensbasis als Voraussetzung, Formulierung einer Hypothese durch den Forscher bzw. Einschätzung der Erfordernisse durch den Lehrer, Entwurf einer Strategie für Forschung bzw. Lehre. Anschließend wird das Labor bzw. der Unterrichtsraum betreten. Erfolg hängt dann ab von technischen bzw. didaktischen Fähigkeiten, Benutzung der geeigneten Taktik, Zugehen auf Forschung und Mitarbeiter bzw. Lehre und Studierende. Evaluation und erneute, darauf aufbauende Planungsphase für das nächste Experiment bzw. die nächste Unterrichtseinheit folgen. Dementsprechend können Forschungsqualitäten auch auf die Lehre gerichtet werden und so beidem, der erfolgreicheren Forschung und der Verbesserung der Lehre, dienen [HANSEN. ME: 104f].

Haltung von Lehrenden gegenüber der Lehre: Etwa 150 von 186 angeschriebenen Lehrenden an einer britischen medizinischen Fakultät beantworteten per Post verschickte Fragebögen zu ihrer Haltung gegenüber der

Lehre. 29% schätzten ihre eigene Lehrqualität als überdurchschnittlich, hingegen nur 6% als unterdurchschnittlich ein. 49% würden mehr unterrichten, wenn es mehr Gelegenheit dafür gäbe. Für die Aussage "Lehrende werden geboren und nicht gemacht" gab es genauso viel Zustimmung wie Ablehnung. "Lehre ist genauso wichtig wie Forschung" meinten 76% (6% waren anderer Meinung, der Rest unentschieden). 45% meinten, daß für sie "Teacher training" hilfreich wäre; 20% hielten dies für überflüssig [FINUCANE et al. ME: 105].

Lernen, mit schwierigen Arzt-Patient-Gesprächssituationen zurechtzukommen: Es wurden auf der Basis von Gesprächssituationen, die Allgemeinmedizinern Schwierigkeiten gemacht haben, Fälle ausgewählt und entsprechende Situationen für SP simuliert. Das Programm wurde von Studierenden und Assistenten gut aufgenommen [BIRAN et al. ME: 105]. Eine andere Studie zeigte, daß Fragen nach Drogenkonsum und Sexualverhalten immer noch nicht integraler Bestandteil der meisten Aanamnesen durch Allgemeinärzte sind. Daraus ergeben sich Anfragen an die ärztliche Ausbildung im Hinblick auf HIV-Risikoeinschätzung und -prävention. [MAHEUX et al. ME: 105]

Stört die Anwesenheit von Studierenden in einer Praxis die Kommunikation zwischen Arzt und Patient? Diese Fragestellung machte ein Medizinstudent zum Ausgangspunkt seiner Untersuchung, die mit dem Upjohn-Jahrespreis der ASME ausgezeichnet wurde. Dazu befragte er Patienten direkt nachdem sie aus dem Sprechzimmer einer Poliklinik kamen. Viele konnten sich nicht einmal daran erinnern, ob ein Student dabei war. Im wesentlichen zeigte sich kein negativer Effekt durch die Anwesenheit von Studierenden. Allerdings legt die Untersuchung die Vermutung nahe, daß die Beteiligung der Studierenden nicht wesentlich über eine reine Beobachtung hinausging [RAHIM. unveröffentlicht].

Einsatz von Computern zur Evaluation: Zu diesem Thema gab es nichts Neues; allgemein läßt sich sagen, daß die bestehenden Programme auf diesem Gebiet jeweils nur auf die lokalen Bedürfnisse zugeschnitten sind und sich kaum übertragen lassen. Ähnlich sieht es mit den lernunterstützenden Programmen aus, da standardisierte Erklärungen bzw. Therapieempfehlungen oft nicht mit den ansonsten gelehrten bzw. praktizierten übereinstimmen. Dies hat dazu geführt, daß sich Computer-Assisted Clinical Learning entgegen allen Vorhersagen (noch) nicht hat durchsetzen können [VAN DER LEE. AA: 684ff].

3. Zusammenfassender Eindruck und Ausblick

Generell sollte bei der Einführung von Prüfungen mit angemessener Umsicht vorgegangen werden. Bislang kann in Deutschland mit wissenschaftlicher Planung, Durchführung und Auswertung von universitären Prüfungen nur in Ausnahmefällen gerechnet werden. Die Auswahl von Prüfungsverfahren sollte sich an der ihnen zukommenden Funktion (z. B. Qualitätssicherung im Hinblick auf die Berufspraxis) unter Berücksichtigung der Kosten-Nutzen-Relation orientieren. Tendenziell fiel

uns in den Diskussionen eine Verlagerung der Konzentration von der Form auf den Inhalt von Prüfungen auf. Vorausgesetzt, der gleiche Inhalt wird abgeprüft, sind eher größere Variationen innerhalb einer Methode (also Unterschiede zwischen einzelnen Prüflingen und den Prüfungsgebieten) als Variationen zwischen Methoden (also Unterschiede der Prüfungsergebnisse eines Prüflings, der mit verschiedenen Verfahren geprüft wird, wie z.B. mit MCQ, OSCE, Standardisierten Patienten, etc.) zu erwarten. Die Evaluation ist abhängiger von dem, was geprüft wird, als von dem, wie geprüft wird.

So wurde in Maastricht eine hohe Korrelation sowohl von Ergebnissen in Wissensprüfungen mit performance-based Prüfungen (wie z. B. OSCE) und der Selbsteinschätzung als auch zwischen MCQ und schriftlichen "open-ended" Fragen festgestellt [JANSEN et al. AA: 176ff; SCHUHWIRTH et al. AA: 486]. Werden hingegen keine signifikanten Korrelationen erreicht, wie dies in den USA zwischen den Ergebnissen von Tests mit Standardisierten Patienten mit Collegenoten, NBME I und II, und Noten in Blockpraktika festgestellt wurde, muß davon ausgegangen werden, daß eine Anzahl von Prüfungen nicht sehr valide sind und sich insbesondere als Aufnahmeprüfungen nicht eignen [SWANSON et al. AA: 465ff].

Trotzdem scheinen bestimmte Prüfungsformen für die Prüfung bestimmter Inhalte besonders geeignet zu sein. Daneben ist zu berücksichtigen, daß Studierende lernen, was geprüft wird (steuernde Funktion von Prüfungen). Dadurch kontrolliert die Prüfung oft das Curriculum statt umgekehrt [BROWN. AA: 3ff; HARDEN. AA: 9ff].

Folgendes Vorgehen wurde für die Einführung von Prüfungen bewährt [BRAILOVSKY et al. AA: 476f; ALLEN et al. AA: 478ff]:

1. Bestimmen der Funktion der geplanten Prüfungen (z. B. Auswahl von Bewerbenden, Überwachung und Steuerung des Lernprozesses, Evaluierung des Programms, Qualitätssicherung);
2. Kosten-Nutzen-Analyse der Methoden (SPs und OSCEs sind meist teurer als MCQs und mündliche Prüfungen);
3. Methodenauswahl ("Don`t get married to a method");
4. Analyse von klinischen Problemen z. B. anhand von Aufnahme-Diagnosen -> Entwicklung von Schlüsselwissen -> Entwicklung von Prüfungsaufgaben.

Bei der Auswahl und Entwicklung von Prüfungsverfahren sollte die Zielgruppe im Auge behalten werden: D.h., möchte man in erster Linie etwas über die Mehrheit der Studierenden oder über die marginalen fünf Prozent, die fast alles oder fast nichts erreichen, erfahren?

Eine weitere auf der Konferenz erkennbare Tendenz ist die Fortentwicklung vom Vergleich Studierender miteinander hin zu einem eher "diagnostischen Ansatz": Was hat der bzw. die Einzelne gelernt, welcher Lernstil ist überwiegend, welche Lernschwächen bedürfen der Korrektur?

Aber auch Lücken in der bisherigen Forschung wurden erkennbar: Hinsichtlich der

Bestimmung von Standards ist bislang wenig geschehen, obwohl aber allgemein als Hilfe für Auswahl und Einsatz von Prüfungsverfahren für wünschenswert gehalten. Standards trennen Bestehende und Durchfallende nach Ausbildungskriterien. Unterschieden werden die Entwicklung von absoluten Standards (ausgedrückt durch den Testinhalt, im Hinblick auf Kompetenz interessanter) und relative Standards (ausgedrückt durch das relative Abschneiden der Testperson, z. B. 95% pass/ 5% fail-Struktur, für Auswahlverfahren interessanter). Dichotome Standards (1/0) sind valider, kontinuierliche Standards (1,2,3, ...,10) reliabler [NORCINI. AA: 32ff].

Alle vorgestellten und diskutierten Prüfungsverfahren beschränken sich auf das Testen von Curriculums-Ausschnitten. Bislang gibt es wenig Beispiele für eine gelungene Evaluation ganzer Ausbildungsprogramme. Dies gilt als eine der Aufgaben für die Zukunft.

In seinem Schlußwort warnte Ian Hart [AA: 17] allerdings auch vor der Illusion, daß es jemals von allen akzeptierte Prüfungsstandards geben wird. Trotzdem sollte die Suche nach ihnen nicht aufgegen werden. Die nächsten Schritte werden im nächsten Jahr in Toronto präsentiert werden - dann hoffentlich schon mit deutschen Beiträgen!

Dr. med. Reinhard Busse, M.S.P.
Abteilung Rheumatologie
Medizinische Hochschule Hannover
30623 Hannover

Christoph Schmidt
Planungsgruppe Reformstudiengang Medizin
Universitätsklinikum Rudolf Virchow
Spandauer Damm 130
14050 Berlin