Die Problematik der objektiven Beurteilung von Prüfungen

Von Dr. Kurt Hügi, Vizedirektor SPI

Wie sichert sich der Ausbilder die Aufmerksamkeit der Teilnehmenden in einem Kurs? Beispielsweise indem er den Inhalt des Unterrichts als prüfungsrelevant bezeichnet. Zugegeben, dies ist nicht die hohe Schule der Methodik und Didaktik. Trotzdem haben Prüfungen den Nebeneffekt, dass sie mindestens kurzfristig als extrinsischer Lernmotivator wirken. Empirische Untersuchungen belegen aber auch immer wieder, dass die Objektivität von Examen sehr relativ ist. Mit der Androhung einer Prüfung wird der Teilnehmer also einerseits zu Leistung angespornt, mit der durchgeführten Prüfung wird man andererseits dieser nicht immer gerecht.

Genau wie sich die Ausbildungsformen in den letzten Jahren verändert haben und heute stärker projekt-, handlungs- und kompetenzorientiert sind, müssen sich auch die Prüfungsformen anpassen (Paradies et al 2005). Praktische Prüfungen und Diplomarbeiten sind zwar schwieriger zu beurteilen als schriftliche Wissensüberprüfungen, sie nähern sich aber mindestens diesem kompetenzorientierten Ansatz. Sie stellen zudem höhere Anforderungen an die Experten. Und diese sind deshalb angehalten, ihr Beurteilungsverhalten bezüglich Objektivität und deren Beeinflussungsmöglichkeiten regelmässig zu reflektieren.

Eine klassische empirische Untersuchung legte 92 Lehrern den gleichen Schüleraufsatz zur Beurteilung vor (K. Ingenkamp et al 2005). Die Gesamtnoten streuten in dieser Untersuchung von 6.0 bis 3.0 mit einem Mittelwert von 4.9. Verständlich ist das Ausmass der Streuung bei Kriterien wie Stil und Inhalt. Dort ist nicht zu erwarten, dass zwischen allen Lehrern ein Fachkonsens bestand, was stilistisch und inhaltlich als gut zu bezeichnen ist.

Sogar wenn dieser gegeben ist, existiert jedoch der wahre Wert der Beurteilung eines Schüleraufsatzes nicht. Die Objektivität wird zwar erhöht, wenn zwei Lehrer den gleichen Aufsatz beurteilen und sich in einem anschliessenden Beurteilungsgespräch einigen. Auch wenn beide Experten zur gleichen Beurteilung kommen, ist dies aber noch kein Beweis, dass dort der wahre Wert liegt (kollektiver Irrtum). Das Risiko, dass die Experten Fehlbeurteilungen vornehmen und diesen angeblich wahren Wert verfehlen, besteht latent.

Erstaunlicher war bei der Untersuchung mit den Schüleraufsätzen aber, dass auch bei einem klaren Kriterium wie der Rechtschreibung die Streuung von Note 6.0 bis 3.0 ging mit einem Mittelwert bei 4.6. Es ist zu vermuten, dass viele Experten einen guten Inhalt und Stil mit einer guten Rechtschreibung assoziierten und umgekehrt.

Wurden den Lehrern Vorinformationen zum Autor des Schüleraufsatzes gegeben, beeinflusste dies die Beurteilung noch einmal stark. Positive Informationen zum Schüler (z.B. zum soziokulturellen Status der Eltern) führten dazu, dass der gleiche Text im Durchschnitt um 0.7 Noten besser beurteilt wurde.

Das Phänomen der grossen Streuungen zeigt sich in zahlreichen Studien übrigens nicht nur in geistes- und sozialwissenschaftlichen Fächern, sondern auch in exakten Wissenschaften, wie z.B. der Mathematik.

Erhöhung der Objektivität durch Konzentration auf Wissensfragen

Völlige Objektivität würde heissen, dass die Beurteilung unabhängig von der Person des Experten ist oder dass mehrere Experten unabhängig voneinander zum gleichen Ergebnis kommen. Dies ist wenig realistisch, wie die oben zitierten empirischen Untersuchungen belegen.

Versuche, die Subjektivität der Experten durch den Computer zu ersetzen, beispielsweise mit geschlossenen Multiple-Choice-Fragen, befriedigen in der Regel nicht. Solche Testverfahren konzentrieren sich auf reines Faktenwissen. Ein solcher Test erlaubt aber nur die Folgerung, dass der Kandidat die Fakten kennt und bei entsprechender Abfrage reproduzieren kann. Es geht also eher um eine Lernzielüberprüfung als einen Kompetenznachweis. Letztlich geht es aber darum, Kompetenzen zu erwerben und diese lebenslang zu vertiefen und zu erweitern.

Gegen das Prüfen von reinem Wissen spricht auch, dass dessen Halbwertszeit (Dauer bis 50% des Wissens nicht mehr gültig ist) relativ kurz ist. Von

Verwendete Literatur:

Arbeitsgruppe Hochschuldidaktische Weiterbildung an der Albert-Ludwigs-Universität Freiburg i. Br., 2000, *Besser Lehren, Heft 10: mündliche Hochschulprüfungen,* Weinheim, Beltz

K. Ingenkamp, U. Lissmann, 2005, *Lehrbuch der pädagogischen Diagnostik,* Weinheim, Basel

Paradies Liane, Wester Franz, Greving Johannes, 2005, *Leistungsmessung und -bewertung,* Cornelsen, Berlin

Wottreng Stephan, 2002, *Handbuch Handlungskompetenz, Einführung in die Selbst-, Sozial- und Methodenkompetenzen,* Sauerländer, Aarau

technologischen Kenntnissen ist bereits nach einem Jahr die Hälfte nicht mehr aktuell. Bei Schul- und Hochschulwissen beträgt die Halbwertszeit heute etwa 10 Jahre. In diesem Zusammenhang ist zu berücksichtigen, dass die Kandidaten das Gelernte, welches sie nach der Prüfung nicht regelmässig brauchen, bereits nach drei Tagen wieder zu 50% vergessen haben (Wottreng 2002). Deshalb rechtfertigt sich auch aus ökonomischen Gründen die Frage, welches Gewicht das Wissen im Rahmen einer Abschlussprüfung erhalten soll.

Scheinobjektivität durch stabile Urteilstendenzen

Da Experten immer mit dem Risiko konfrontiert sind, eine Leistung zu gut oder zu schlecht zu beurteilen, entsteht die Gefahr der stabilen Urteilstendenzen. Einigen sie sich, unabhängig von der Leistung des Kandidaten, auf die Note 4.5, so ist dies schliesslich weder gut noch schlecht und wird kaum eine Beschwerde provozieren. Die Beurteilenden nehmen sich so aus dem Schussfeld der Kritik und gelten weder als zu wohlwollend noch als scharfe Hunde.

Wenn aber mehrere Experten so denken, streuen die Ergebnisse der Kandidaten nicht mehr genügend. Man spricht dabei von sogenannten Verteilungsfehlern (Arbeitsgruppe Hochschuldidaktische Weiterbildung, 2000), das heisst, die Notenskala der Leistungsbeurteilung wird nicht ausgenutzt. Damit fehlt die Differenziertheit als wichtiges Qualitätskriterium, welches Aufschluss über die Stärken und Schwächen der einzelnen Kandidaten gibt. Die fehlende Spreizung kann natürlich auch damit zusammenhängen, dass die gestellten Aufgaben zu einfach waren.

Die Prüfungskommission der eidgenössischen Berufsprüfung Polizist/Polizistin hat deshalb im Rahmen eines Workshops Qualitätskriterien zur Streuung der Noten erarbeitet (siehe Rahmen). Mit diesen empirischen Werten sollen in den nächsten Jahren Erfahrungen gesammelt werden. Dazu werden die Notenbilder aller Prüfungssessionen statistisch ausgewertet. Es soll dadurch eine vertiefte Fachdiskussion über die weitere Qualitätsentwicklung dieser Prüfungen in Gang kommen.

Kombination von schriftlichen, mündlichen und praktischen Prüfungen

Prüfungsmodelle, wie z.B. die Berufsprüfung Polizist/Polizistin, welche schriftliche, mündliche und praktische Elemente vereinen, begehen in diesem Spannungsfeld einen sinnvollen Mittelweg. In schriftlichen Prüfungen lässt sich lernzielorientiertes Wissen überprüfen und objektiv beurteilen. Dies auch darum, weil alle Kandidaten gleichzeitig die gleichen Fragen vorgelegt bekommen und diese auch von den gleichen Experten beurteilt werden können. Trotz dieser recht hohen Objektivität zeigen die Erfahrungen, dass sich Beschwerden meistens auf schriftliche Prüfungen beziehen. Dies dürfte damit zusammenhängen, dass die Faktenlage (Fragen, Antworten, Beurteilung) hier am transparentesten ist. Und wenn für eine erfolgreiche Prüfung eine halbe Note fehlt, starten viele Kandidaten den Versuch, die dazu nötigen Punkte durch eine Beschwerde zu holen. Aussicht auf Erfolg hat diese üblicherweise nur, wenn formale Verstösse gegen die Prüfungsordnung nachgewiesen werden können.

© SPI

Empirische Parameter gegen stabile Urteilstendenzen

(Basis ist immer eine repräsentative Anzahl von mindestens 40 Kandidaten):

- Positionsnoten, welche in ganzen und halben Noten ausgedrückt werden, spreizen über mindestens fünf Halbnoten (z.B. von 3.5 bis 5.5).
- Die statistisch berechnete Streuung ist bei jeder Positionsnote grösser als 0.5.
- Maximal 60% der Noten eines Prüfungsfaches verteilen sich auf die zwei Positionen mit den grössten Häufigkeiten (z.B. 40% der Kandidaten haben eine 4.5, 20% eine 5.0).
- Die Note 6.0 kommt in jedem Fach vor.
- Der Mittelwert von Positions- und Fächernoten hat eine Vier vor dem Komma.

Wichtig ist, dass Prüfungen, welche knapp ungenügend sind, in den Notensitzungen der Kreiskommissionen und der Prüfungskommission besprochen und im Sinne der Verhältnismässigkeit nachbeurteilt werden. Dabei gibt es letztlich immer die Möglichkeit, den einen oder anderen Punkt mehr oder weniger zu geben, sonst würde man eine falsche Objektivität vortäuschen.

Natürlich können auch in schriftlichen Prüfungen transfer- und anwendungsorientierte offene Fragen oder Fallbeispiele behandelt werden. Deren Korrektur ist aber sehr aufwändig, insbesondere wenn durch die formale Antwortstruktur nicht ein relativ enger Rahmen vorgegeben wird.

Mündliche Prüfungen sind subjektiver, aber aussagekräftiger

Mündliche Prüfungen lassen sich hingegen weniger standardisieren. Die Kandidaten sind aus organisatorischen und prüfungstechnischen Gründen mit unterschiedlichen Fragen und Experten konfrontiert. Der eine Kandidat hat dabei Glück, der andere Pech. Wer kennt das nicht aus eigenen Prüfungserfahrungen! Je mehr voneinander unabhängige Themen in mündlichen Examen angesprochen werden, desto stärker reduziert sich dieser Effekt.

Mündliche Prüfungen haben eine höhere Aussagekraft als schriftliche. Der Experte kann nach einer eher wissensorientierten Einstiegsfrage auch Reproduktions- und Transferfragen stellen und solange nachfragen, bis er sicher ist, dass der Kandidat die Inhalte verstanden hat und allenfalls sogar anwenden könnte.

Der Experte muss, genau wie es beim Schiessen geschieht, zuerst das Trefferbild kennen, bevor er die Visierung verstellt

Praktische Prüfungen und Diplomarbeiten sind kompetenzorientiert

Praktische Prüfungen kommen dem Berufsalltag schon sehr nahe. Die Beurteilung für den Experten wird aber anspruchsvoller, da ein praktischer Ablauf nicht einfach die Folge von einzelnen, messerscharf zu beurteilenden Teilschritten ist. Er braucht viel Praxiserfahrung, solide Beurteilungskompetenzen und den Blick fürs Ganze, um dem Kandidaten gerecht zu werden. Erschwerend kommt hinzu, dass bei praktischen Prüfungen bei der Polizei jeweils zwei Kandidaten im Team arbeiten, aber jeder einzeln beurteilt werden muss. Neben den fachlichen Aspekten werden hier also explizit auch die sozialen und methodischen Kompetenzen geprüft. Dieses Verfahren nähert sich den berufsrelevanten Handlungskompetenzen sehr stark.

Seminar- und Diplomarbeiten sind auch zu den praktischen Prüfungen zu zählen. Hier geht es darum, einen gesamten Projektablauf von der Fragestellung über das methodische Vorgehen, die Aussagekraft der Ergebnisse bis hin zur redaktionellen Verarbeitung in einem Bericht zu beurteilen. Dazu sollen klare Beurteilungskriterien bestehen, und diese sind den Kandidaten auch zu kommunizieren. So wird auf beiden Seiten die grösstmögliche Transparenz zur Leistungserwartung geschaffen. Bei der Beurteilung von Diplomarbeiten der höheren Fachprüfung Polizist/Polizistin arbeiten die Experten beispielsweise mit einem Beurteilungsblatt, welches zu 13 formalen und inhaltlichen Kriterien ein Beurteilungsraster vorgibt.

Fazit

Wenn sich Experten regelmässig mit den Gütekriterien der Objektivität auseinandersetzen und sich auch der subjektiven Seiten und Gefahren des Prüfens bewusst sind, ist dies der beste Garant für die kontinuierliche Qualitätsverbesserung der Prüfungen. Wichtig ist dabei auch die Reflexion des eigenen Beurteilungsverhaltens (Hoch-, Zentral- oder Tiefzensierer) in Notensitzungen, Briefings und Weiterbildungen. Genau wie beim Schiessen muss man zuerst das Trefferbild kennen, bevor die Visierung verstellt wird.

Prüfungen wirken also letztlich nicht nur als billiger Lernmotivator. Sofern sie als einheitliche Qualifikationsverfahren in einer ganzen Branche durchgeführt werden, garantieren sie unbestrittenerweise auch einen hohen Ausbildungsstandard. Voraussetzung für solche Prüfungen ist nämlich, dass sich die Branche bezüglich Inhalten und Anforderungsniveaus einigt und dies in einem Reglement und einem Rahmenlehrplan fixiert. Von diesem gemeinsamen Ausgangspunkt aus, kann sich ein Berufsbild regelmässig neuen Herausforderungen anpassen und qualitativ weiterentwickeln, was in schnelllebigen Zeiten unabdingbar ist.

A propos de l'objectivité dans l'évaluation des examens

Par Kurt Hügi (Traduction ISP)

Comment s'assurer, en tant que formateur, l'attention des participants dans un cours ? En déclarant, par exemple, que tout contenu traité en classe fera partie de l'examen. Il faut avouer qu'au niveau didactique et méthodologique ce comportement n'est pas à privilégier. Toutefois, il est vrai que les examens fonctionnent, à court terme, comme motivateur extrinsèque, bien que leur objectivité reste très relative. D'un côté, la pression d'un examen a comme effet de pousser les candidats à fournir de bonnes prestations, de l'autre, ces dernières ne sont pas toujours appréciées à leur juste valeur. Il est donc absolument nécessaire que, de temps en temps, les experts remettent en question leur manière d'évaluer les examens, sachant que le respect de l'objectivité est un facteur difficile à gérer. Le risque de se tromper de note en évaluant un test est constamment présent, car même si deux ou plusieurs experts trouvent un consensus, cela ne signifie pas automatiquement qu'ils aient attribué la bonne appréciation à l'épreuve qu'ils ont jugée (erreur collective). Pour se rapprocher le plus possible de la vraie valeur d'un travail, il faudrait se limiter à contrôler le savoir factuel uniquement. Or, des examens ne portant que sur du cognitif iraient à l'encontre d'un apprentissage orienté vers les compétences. Cette méthode ne prend que marginalement ce type de savoir en compte mais elle contrôle les acquis plutôt par des cas d'école, des questions de transfert ou des vérifications transverses.

Pour échapper au danger d'être trop sévères ou, a contrario, trop indulgents, les experts ont tendance à attribuer la note 4.5 avec comme conséquence un manque d'écart entre les notes, alors que c'est justement leur variété qui constitue un critère de qualité important. Ces réflexions ont mené la Commission d'examen de l'Examen professionnel fédéral de Policier/Policière à élaborer des critères de qualité, afin de garantir une bonne répartition des notes. Les résultats des sessions d'examen seront analysés et serviront de base pour approfondir la discussion sur le développement ultérieur de ces examens qui réunissent déjà maintenant des éléments écrits, oraux et pratiques permettant de valider les connaissances des candidats à plusieurs niveaux.

Tout ce qui relève de l'interrogation écrite se prête plus facilement à tester le savoir factuel, alors qu'à l'oral l'expert peut vérifier si le candidat est capable de répondre à des questions ouvertes et de transfert. S'il est vrai que les examens oraux sont moins standardisés et plus subjectifs, ils offrent en revanche à l'examinateur la possibilité d'évaluer le candidat de manière plus complète. Les examens pratiques, finalement, sont plutôt complexes, se rapprochent beaucoup du travail quotidien et exigent une grande expérience de la part des experts qui doivent disposer de compétences d'évaluation solides et garder une vue d'ensemble pour juger correctement les participants. Les travaux de séminaire et de diplôme peuvent également être considérés comme des examens pratiques. Ils révèlent les compétences des candidats dans les détails (maîtrise du processus dans la conduite du projet du début à la fin et fil rouge clair). Pour la correction de ces travaux, les experts recourent à une grille d'évaluation donnée.

S'il est prêt à intégrer les critères qui favorisent l'objectivité et qu'il est conscient des risques de subjectivité liés à sa fonction, l'expert peut contribuer à améliorer la qualité des examens. Comme dans le tir, il faut d'abord identifier la cible avant d'ajuster le viseur !

© ISP