W 701 3

PROBLEM
und
ENTSCHEIDUNG

Arbeiten zur Organisationspsychologie
aus der Abteilung für Angewandte Psychologie
des Psychologischen Instituts der Universität München
und der Fachgruppe Psychologie der Universität Augsburg

Heft 6

München 1971

Testtheoretische Aspekte akademischer Prüfungen

Hermann Brandstätter

Die Diskussion über die Lage der Universitäten, über mögliche Ziele und Wege der Erneuerung wird seit Jahren mit wachsender Intensität und Breitenwirkung geführt. Täglich werden Lehrende und Studierende mit den Schwierigkeiten konfrontiert, die sich aus der ständig steigenden Anzahl von Studenten, aus dem immer rascheren Fortschreiten spezialisierter Wissenschaften und der Unsicherheit in den Bildungszielen ergeben.

In einer solchen Situation ist es nicht zuletzt die Psychologie, die aufgefordert wird, ihren Beitrag zur Lösung der dringendsten Probleme zu leisten. Es ist offensichtlich auch eine psychologische Frage, ob die gewählten Bildungsziele unter den gegebenen Bedingungen überhaupt realisiert werden können, auf welchem Wege sie am besten anzugehen sind und wie geprüft werden kann, ob sie erreicht wurden.

Dem dritten Punkt, der Kontrolle des Ausbildungserfolges, kommt dabei insofern eine vorrangige Bedeutung in der Unterrichtsforschung zu, als die Analyse des Einflusses der Person- und Umweltvariablen auf das Lernen eine möglichst adäquate Erfassung des Lernerfolges voraussetzt.

Das Problem der Erfolgskontrolle des Studiums ist außerdem hochschulpolitisch von besonderer Aktualität: die herkömmliche Prüfungspraxis wird seit einiger Zeit von den Studenten, z.T. mit guten Gründen, heftig kritisiert und selbst von vielen Prüfern als sehr unzulänglich empfunden. Auch in den verschiedenen Kommissionen, die

an neuen Prüfungsordnungen arbeiten, sucht man nach zweckmäßigeren Formen der Prüfung, die den heutigen Ausbildungsbedingungen und -zielen besser gerecht werden. Die folgende testtheoretische Betrachtung akademischer Prüfungen soll zeigen, auf welchem Weg hier ein Fortschritt erreicht werden könnte.

Eine Prüfung ist, ähnlich wie ein psychologischer Test, eine absichtliche Anregung zu einem Verhalten, aus dem weitere Schlüsse über einen Menschen gezogen werden können. Es würde, so scheint mir, zu einer recht willkürlichen Abgrenzung führen, wollte man einen grundsätzlichen Unterschied zwischen einer herkömmlichen Schulleistungsprüfung und einem psychologischen Test herausstellen. Auch geschichtlich haben sich viele der strenger geregelten Tests aus den traditionellen Schul- und Berufsleistungsprüfungen entwickelt. Von improvisierten, methodisch ganz sorglos gestellten Fragen bis hin zu streng standardisierten und statistisch analysierten Aufgaben sind unter dem, was Prüfung oder Test heißt, alle Zwischenstufen anzutreffen.

Es ist für unsere Frage nützlich, Prüfungen danach einzuteilen, ob Aufgabenstellung, Antwortprotokollierung und Antwortbewertung (1) improvisiert oder (2) standardisiert sind. Bei mündlichen Prüfungen sind in der Regel sowohl Aufgabenstellung als auch die Protokollierung und Bewertung (Interpretation) der Antworten improvisiert. Bei Klausuren in Form von Aufsätzen, um ein weiteres Beispiel zu nennen, finden wir eine gewisse Standardisierung der Aufgabenstellung.

Die klassifikatorischen Begriffe "improvisiert - standardisiert" stellen die Zweiteilung eines Kontinuums dar; jeder der drei Schritte einer Prüfung kann mehr oder weniger improvisiert bzw. standardisiert sein. Es bedeutet eine Einschränkung des willkürlichen Improvisierens, wenn z.B. manche Prüfer aus einem Zettelkasten vorbereitete Fragen ziehen lassen oder Aufsätze anhand von vorher festgelegten Regeln bewerten.
Wie die Prüfungen lassen sich auch alle gebräuchlichen psychologischen Tests bzw. die Art ihrer Verwendung zwanglos auf ein Kontinuum zwischen "improvisiert" und"standardisiert" einreihen. Da sich die üblichen Universitätsprüfungen von psychologischen Tests in diesen formalen Aspekten nicht grundsätzlich, sondern nur graduell unterscheiden, liegt es nahe, die Brauchbarkeit von Prüfungen nach ähnlichen Kriterien zu bewerten wie die von psychologischen Tests.

Im Unterschied zu einem psychologischen Test wird aber von einer akademischen Prüfung nicht nur diagnostische Tauglichkeit, sondern auch didaktische Zweckmäßigkeit gefordert, doch soll dieses letzte Problem hier nicht weiter erörtert werden. Die diagnostische Brauchbarkeit einer Prüfung hängt ganz wesentlich davon ab, welche Schlüsse mit welcher Wahrscheinlichkeit aus dem konkreten Prüfungsergebnis eines Kandidaten gezogen werden können.
Man pflegt in der psychologischen Testtheorie drei verschiedene Arten von solchen Schlüssen zu unterscheiden: (1) den Schluß auf ein hypothetisches Konstrukt, z.B. Intelligenz oder Ängstlichkeit. (2) den Schluß auf künf-

tiges Verhalten, z.B. bei beruflichen Aufgaben. (3) den Schluß von den Leistungen in einer Stichprobe von Aufgaben auf die Leistungen in der Gesamtheit der Fragen und Problemsituationen, aus der die Prüfungsaufgaben ausgewählt wurden.
Mit dem wichtigen testtheoretischen Begriff der Validität ist nichts anderes als die Wahrscheinlichkeit des Zutreffens eines solchen Schlusses gemeint. Je nach Art des Schlusses spricht man von Konstrukt-, Prognose- oder Repräsentationsvalidität.
Bei Prüfungen kommt es vor allem darauf an, daß der Schluß von den Prüfungsleistungen auf die Gesamtheit der Leistungen, zu denen die Ausbildung befähigen sollte, möglichst gut abgesichert ist, d.h. daß die Repräsentationsvalidität möglichst hoch ist. Die beiden anderen Validitätsbegriffe, nämlich Konstrukt- und Prognosevalidität, sind hier von geringerer Bedeutung und können in diesem Rahmen nicht weiter erörtert werden.

Im Sinne der Statistik ist eine Stichprobe von Elementen aus einer Grundgesamtheit dann repräsentativ, wenn jedes Element der Grundgesamtheit die gleiche Chance hat, in die Stichprobe aufgenommen zu werden. Hier nun stellt sich das zentrale Problem, wie man eine Grundgesamtheit von Aufgaben definieren soll.

Es böten sich verschiedene Möglichkeiten an: man könnte z.B. versuchen, eine Grundgesamtheit von Prüfungsaufgaben dadurch herzustellen, daß man eine Reihe von Sachverständigen bittet, nach einem bestimmten Plan möglichst viele Aufgaben (Fragen) zu formulieren. Der Plan müßte

dabei eine Tabelle von Lehrzielen darstellen, die in den Zeilen die Kenntnisgegenstände, in den Spalten die Kenntnisform (z.B. Erinnern, Verstehen, Anwenden etc.) enthält. Damit nun verschiedene Experten äquivalente Aufgabensätze erstellen könnten, müßten die Ausbildungsziele, der Aufgabenplan und die Regeln für die Aufgabenkonstruktion eindeutig fixiert sein. Den Kandidaten könnte dann in der Prüfung eine Zufallsauswahl aus der so entstandenen Aufgabengesamtheit vorgelegt werden.

Offen bleibt aber trotzdem noch die Frage, ob die Aufgaben, die zu einem bestimmten Feld der Plantabelle konstruiert wurden, tatsächlich das erfassen, was nach Plan erfaßt werden soll. Diese Frage läßt sich nur dann beantworten, wenn der Plan genaue Angaben darüber enthält, welche Merkmale die Leistungen aufweisen müssen, die in den einzelnen Feldern der Plantabelle (Lehrzielmatrix) verlangt werden. Eine Aufgabe wird nur dann akzeptiert, wenn sie Leistungen mit den im Plan verlangten Merkmalen provoziert.

Man könnte sich auf den Standpunkt stellen, daß jede Aufgabe, welche die für die betreffende Aufgabenklasse geforderten Merkmale enthält, gleichwertig ist. Der Nachweis der Repräsentationsgültigkeit beschränkte sich dann darauf zu zeigen, daß die Aufgaben unter den Klassenbegriff subsummiert werden können. Auf diese Weise wäre, so scheint es, das Problem der Repräsentationsgültigkeit am einfachsten gelöst.

Diese Lösung ist jedoch psychologisch nicht befriedigend. Die Aussage, daß eine bestimmte Aufgabe zu einer bestimm-

ten Klasse von Aufgaben gehört, läßt nämlich weitere Merkmale zu, in denen sie sich von anderen Aufgaben der Aufgabenklasse unterscheidet. Diese weiteren Merkmale können - mehr oder weniger gut erkennbar - für den Prüfungszweck ebenfalls relevant sein.

Man kommt also letztlich, auch wenn man die Aufgabenklassen ziemlich eng definiert, nicht um eine Art von Zufallsauswahl innerhalb der Klassen herum, weil man (1) nicht alle relevanten Merkmale kennt, (2) nicht alle vermutlich relevanten Merkmale planmäßig in der Aufgabenkonstruktion berücksichtigen kann.

Um bei der Konstruktion der zu einer bestimmten Klasse gehörigen Aufgaben Einseitigkeiten zu vermeiden, empfiehlt es sich, die Aufgaben in allen Merkmalen, die nicht zu den klassenbildenden Merkmalen gehören, möglichst zu variieren. Wenn die Aufgaben außerdem von verschiedenen Autoren stammen, ist die Gefahr von Einseitigkeiten noch geringer. Die üblichen Tendenzen zur Bevorzugung von leicht konstruierbaren Wissensfragen oder von Fragen, die dem Autor gerade geläufig sind, können weitgehend durch spezifizierte Aufgabenpläne vermieden werden.

Bis jetzt war nur von der Repräsentativität der Aufgaben die Rede. Ein weiteres Problem ist die Repräsentativität der Umstände.

Die gleichen Aufgaben können unter ganz verschiedenen Umständen gestellt werden. Je nach Umständen werden auch die Leistungen verschieden ausfallen. Ein gereizter Prüfer

erhält vielleicht auf die gleichen Fragen ganz andere Antworten als ein ausgeglichen freundlicher Prüfer. In der bedrohlich erlebten Prüfungssituation kommen bei den gleichen Aufgaben oft ganz andere Leistungen zustande als in einer entspannten, harmlosen Situation. Bestimmte Umstände bewirken also systematische Fehler in der Leistungsmessung.

Es wäre nun unrealistisch zu fordern, jede Aufgabe mit einer repräsentativen Stichprobe von Umständen zu kombinieren, etwa in der Annahme, daß sich dann die mit den wechselnden Umständen verbundenen Fehler in der Leistungsmessung ausgleichen. Vernünftig ist es dagegen, die Aufgaben unter solchen Umgebungsbedingungen zu stellen, die auch im beruflichen Leben am häufigsten anzutreffen sind.

Hier wird nun das Dilemma jeder geplanten und geregelten Prüfung deutlich: einerseits ist die Prüfungssituation bestimmt nicht typisch für die Umstände, unter denen im Beruf Probleme zu lösen sind; andererseits ist es schwer vorstellbar, wie man zu einem bestimmten Termin prüfen soll, ohne bei vielen eine außerordentliche und leistungshemmende Belastung, bei anderen einen ungewöhnlichen, leistungsfördernden Ansporn hervorzurufen. Die verschiedenen Untersuchungen zur Prüfungsangst und Risikobereitschaft zeigen - bei gleicher Leistungsfähigkeit in entspannter Situation - erhebliche individuelle Unterschiede in der Fähigkeit, den Prüfungsbelastungen standzuhalten. Dies soll als Beispiel für systematische Fehler genügen.

Die Validität einer Prüfung wird aber nicht nur durch systematische, sondern auch durch zufällige Fehler eingeschränkt. Ein Aufgabensatz kann nur dann valide Ergebnisse bringen, wenn die Aufgabenlösungen reliabel, d.h. möglichst frei von Zufallsfehlern sind.

In der "klassischen" Testtheorie wird angenommen, daß sich jeder Meßwert aus einem sogenannten wahren Wert und einem zufälligen Fehlerwert zusammensetzt. Der Fehlerwert kann sich wiederum aus verschiedenen Komponenten ergeben. Man unterscheidet gewöhnlich drei Gruppen von Fehlern, für die Zufälligkeit gefordert wird:

(1) zufälliger Wechsel in den Anregungsbedingungen (Aufgabenauswahl und Umstände)

(2) zufälliger Wechsel im Zustand des Pb (Schwankungen der intellektuellen Leistungsfähigkeit und der Motivation)

(3) zufälliger Wechsel (intra- und interindividuell) im Zustand der Beurteiler (Schwankungen bzw. Unterschiede in den Beurteilungsaspekten, im Adaptationsniveau, in den Absichten und Motiven).

Wir befassen uns jetzt nicht mit der Frage, ob die Veränderungen in den Anregungsbedingungen, im Probanden und im Beurteiler, für die im Modell Zufälligkeit vorausgesetzt wird, wirklich zufällig, oder ob sie - wenigstens zum Teil - nicht doch regelhaft sind. Auch wenn die Vor-

aussetzungen des Modells bezüglich der Zufälligkeit der Fehler nur annähernd erfüllt sind - und dies soll für die weiteren Erörterungen angenommen werden - kann man damit praktisch arbeiten. Wenn eine Reihe von Kandidaten eine Prüfung über den gleichen Gegenstand mit anderen Fragen, zu einer anderen Zeit und bewertet von einem anderen Beurteiler wiederholt, müßten sich für jeden Kandidaten wieder die gleichen Noten ergeben, wenn die Prüfung voll zuverlässig wäre. Je größer die Abweichungen sind, desto geringer ist auch die Validität, sei es daß diese als Konstrukt-, Prognose- oder Repräsentationsvalidität aufgefaßt wird. Weiß man, wie zuverlässig eine Prüfung ist, so kann man angeben, wie hoch ihre Validität maximal sein kann - immer vorausgesetzt, daß die Annahmen des Modells zutreffen.

Im Hinblick auf die eben ganz grob skizzierten Gütekriterien einer Prüfung müssen Prüfungsgespräche und Aufsätze, die beiden häufigsten Arten akademischer Prüfungen, als sehr fragwürdig erscheinen.

Wie ein Überblicksreferat von COX (1967) zeigt, ist die Übereinstimmung in der Benotung von Aufsätzen und Prüfungsarbeiten aus verschiedensten Gebieten im allgemeinen recht unbefriedigend. Charakteristisch sind Korrelationen von $r = 0{,}50$. Das bedeutet, daß höchstens 50 % der Notenvarianz gültige Varianz sind. Da aber zu den Fehlern infolge mangelnder Objektivität der Bewertung noch die Fehler kommen, die sich mit der zufälligen Wahl des betreffenden Themas und mit dem zufällig angetroffenen Zustand des Kandidaten ergeben, ist die gültige Va-

rianz meist noch wesentlich geringer.
Noch ungünstiger dürfte es um die bisher noch kaum untersuchte Reliabilität und Validität der Ergebnisse mündlicher Prüfungen bestellt sein. Es erscheint daher als vernünftig und notwendig, an den Universitäten Prüfungen in Form von standardisierten Kenntnis- und Verständnistests einzuführen. Nach den bis jetzt gemachten Erfahrungen bringen die zuverlässigsten Ergebnisse Serien von Auswahlaufgaben: bei diesen Aufgaben ist aus mehreren Lösungsvorschlägen die beste Lösung vom Kandidaten herauszufinden.
Der Rangplatz eines Kandidaten wird von der Anzahl der richtig gelösten Aufgaben abgeleitet. Die Voraussetzungen und das Verfahren einer solchen Transformation können hier nicht weiter dargestellt werden.
Auswahlaufgaben werden oft mit der Begründung abgelehnt, daß sie einseitig das Abfragen von Faktenwissen fördern. Wenn aber ein gut durchdachter Aufgabenplan vorliegt und wenn der problemadäquate Aufgabentyp gewählt wird, gelingt es bei einiger Sorgfalt und Erfahrung sehr wohl, mit Auswahlaufgaben auch den Erfolg komplexerer Lern- und Denkprozesse zu prüfen. Man kann außerdem auch Aufgaben verwenden, bei denen die Lösung nicht aus vorgegebenen Alternativen ausgewählt, sondern frei produziert wird. Anregungen dazu findet man in den Testserien zur Prüfung kreativer Leistungen (vgl. TORRANCE, 1968).

Obwohl Aufsätze und mündliche Prüfungen als Meßverfahren wenig geeignet sind, erweisen sie sich u.U. als Anreiz zum gewünschten Lernverhalten als sehr nützlich. Sie haben nämlich den Vorteil, daß sie stärker als Auswahlaufgaben zu einer integrierenden und persönlich akzentuierenden Auseinandersetzung mit einem größeren Gebiet anregen. Selbst wenn man die Einseitigkeit der Auswahlaufgaben durch zusätzliche Verwendung von Produktionsaufgaben ausgleicht, wäre es nicht ratsam, ganz auf Prüfungsgespräche und Aufsätze zu verzichten. Improvisiert mündlich oder schriftlich zu wissenschaftlichen Problemen Stellung zu nehmen, ist immerhin eine Aufgabe, die in vielen akademischen Berufen alltäglich ist. Prüfungsgespräche und Aufsätze geben einen Eindruck von der Geschicklichkeit des Kandidaten in diesen wichtigen Leistungen. Sie sind für diesen Zweck nicht durch andere Prüfungsformen ersetzbar. Allerdings sollten auch dafür Verfahrensregeln entwickelt und angewendet werden, die eine zuverlässigere Beurteilung ermöglichen.

Die wichtige Frage, wie sich Prüfungen auf das Studium auswirken, wurde nicht behandelt, obwohl die didaktische Funktion der Prüfung vielleicht noch wichtiger ist als die diagnostische.

Es sei abschließend nur darauf hingewiesen, daß der Einfluß der Prüfung auf das Lernen ständig kontrolliert werden sollte. Man könnte etwa zu jedem Prüfungstermin die Kandidaten anonym befragen, wie sie sich auf die Prüfung vorbereitet und welchen Eindruck sie von der Prüfung gewonnen haben; dann wüßte man ungefähr auch, was den Kommilitonen erzählt wird und wie sich diese auf die künftigen Prüfungen einstellen werden. Mit routinemäßig gesammelten Informationen dieser Art müßte es gelingen, Prüfungen noch besser auf die Ausbildungsziele abzustimmen und so die gewünschte Art des Studiums zu fördern.

Literaturhinweise

COX, R. Examination and higher education: Survey of the Literature; Universities Quarterly 1966/67, 21, 292 - 340

TORRANCE, E.P. Neue Itemarten zur Erfassung kreativer Denkfähigkeiten in: Ingenkamp & T. Marsolek (Hgb.) Möglichkeiten und Grenzen der Testanwendung in der Schule, Weinheim, 1968