Bernhard Jacobs, Fachrichtung Bildungswissenschaften der Universität des Saarlandes.
Email: b.jacobs@mx.uni-saarland.de
Version: 14.5. 2010

Bestanden/nicht bestanden im Vergleich zum traditionellen, mehrstufigen Benotungssystem

Abstract
Ziel dieses Betrags ist es, die Auswirkungen der Benotung "bestanden/nicht bestanden" im Vergleich zum traditionellen, mehrstufigen Benotungssystem im Hinblick auf schulische Leistungen und subjektive Variablen einzuschätzen. Bei einer lockeren Literatursuche fiel auf, dass die Thematik um das Jahr 1970 gewisse Forschungsaktivitäten anregte, die bald in der Versenkung verschwanden und erst in neuerer Zeit vornehmlich im Medizinstudium wieder aufgenommen wurden. Die frühen Studien, welche überwiegend die Note "ausreichend" als Kriterium des Bestehens definierten, stellten konsistent schlechtere Studienleistungen unter dem zweistufigen Notensystem fest. Einige neuere Studien fanden Belege für vergleichbare Studienleistungen unter beiden Notensystemen, setzten aber höhere Anforderungen an das Bestehen. Sie liefern auch Hinweise für eine geringere Stressbelastung und ein höheres Wohlbefinden bei einer Benotung "bestanden/nicht bestanden".

Schlagworte: Benotungssystem, Leistungsbewertung

Einleitung

Beim klassischen Pass/Fail Notensystem werden die Prüfungsleistungen aller Prüflinge in 2 Kategorien eingeteilt:

1. Prüfung bestanden (pass) oder
2. Prüfung nicht bestanden (fail).

Damit entspricht P/F einer kriteriumsorientierten Bewertung. Denn die Leistungen unter der Kategorie "bestanden" unterliegen ähnlich wie bei der Führerscheinprüfung in Deutschland keiner weiteren Differenzierung. Mit der Bewertung "bestanden" wird dem Prüfling attestiert, dass er die Anforderungen erfüllt hat.

Es existieren einige unechte Varianten gegenüber klassischen P/F, welche eine weitere Differenzierung nicht ganz aufgeben wollen, wie etwa "pass, marginal pass, fail", "honors, pass and fail", "honors, high pass, pass and fail". Derartige Formen unechten P/F's entsprechen aber im Prinzip dem konventionellen, mehrstufigen Benotungssystem, welches in Deutschland die Ziffern 1 bis 6 umfasst. Das traditionelle amerikanische Benotungssystem verwendet die 5 Buchstaben A, B, C, D sowie F für "durchgefallen". Es wird deshalb auch häufig letter-grading genannt und hier meist als A..F- oder konventionelles Notensystem bezeichnet. Ziel dieses Betrags ist es, auf der Basis empirischer Studien das P/F- im Vergleich zu einem A..F -System im Hinblick auf schulische Leistungen und subjektive Einschätzungen zu bewerten.

Zunächst stellt sich die Frage, wie hoch die Prüfungsleistung ausfallen muss, damit sie als bestanden gewertet werden soll. Hier sind mehrere Kriterien denkbar, die etwa bei einem Test von mastery pass (z.B. mindestens 80% korrekter Lösungen) bis "marginal pass, etwa 50% korrekte Lösungen" reichen könnten, gelegentlich gar nicht expliziert und leider selten hinreichend begründet werden. Die Wirkungen eines P/F-Systems auf Lernleistung und Stress hängen aber ganz entscheidend auch von diesem Leistungskriterium ab. In der Praxis wurde das P/F-System in der Regel nicht flächendeckend bzw. ausschließlich eingesetzt, sondern

meist von bestimmten Bedingungen abhängig gemacht und auf einige Seminare bzw. Semester beschränkt.

Motivierung für bzw. optimistische Erwartungen an ein P/F-System

Der Wechsel vom traditionell mehrstufigen Notenkonzept zum P/F-System soll den Notendruck mildern ohne das Leistungsstreben bzw. gute Leistungen ernsthaft zu beeinträchtigen. Die Minimalanforderung für das Bestehen muss daher eigentlich so gewählt sein, dass sie einen pädagogisch akzeptablen Standard erfüllt. Ziel des P/F-System müsste es sein, jeden Studierenden zu einem angemessenen Studiereinsatz zu bewegen, dabei aber ein überzogenes Anspruchsniveau, unnötigen Übereifer, übermäßiges Konkurrenzstreben oder nutzloses Überlernen weitgehend zu vermeiden. Dadurch ergäben sich auch etwas mehr Freiheiten für den Studierenden, eigenen Interessenschwerpunkten im Studium zumindest teilweise Rechnung zu tragen und nicht jede Kurswahl primär unter dem Aspekt der Notenoptimierung zu betrachten. Im Idealfall orientiert sich der Studierende an den fachlichen Anforderungen des Seminars, berücksichtigt dabei auch sein intrinsisches Fachinteresse und macht sich weniger Gedanken über die Leistungsbewertung, da er zu der Überzeugung gelangt, durch seinen Arbeitseinsatz das Kriterium relativ sicher zu erreichen. Ohne äußeren Druck versucht er darüber hinaus, ein seiner Leistungsfähigkeit entsprechendes, möglichst gutes Ergebnis zu erzielen. Da das P/F-System im Mittel die Bedrohung senkt, den erforderlichen Leistungsansprüchen nicht genügen zu können, erhöht sich die subjektive Lebensqualität der Studierenden, was sich in einem geringeren wahrgenommenen Stress bzw. einem besseren Wohlbefinden ausdrückt.

Ein weiterer Vorteil könnte sich in einem günstigeren sozialen Verhalten der Studierenden untereinander zeigen. Je nachdem, wie eine klassische Notengebung gehandhabt wird, erzeugt sie etwa bei strikter Anwendung nach Normalverteilung objektiv totale Konkurrenz, da der Einzelne nur auf Kosten des Mitstudierenden besser abschneiden kann. Selbst unabhängig von einer objektiv zutreffenden Konkurrenz des klassischen Notensystems, das sich in Deutschland ja eigentlich, rein formal betrachtet, auch an einer kriteriumsbezogenen Norm orientieren soll, antizipieren etliche Studierende ein Konkurrieren um die guten Noten. Derartige Bedingungen oder subjektive Befürchtungen entfallen unter P/F. Einige Forscher erwarten daher durch den Wechsel zu einem P/F- System eine bessere Gruppenkohäsion und mehr gegenseitige Hilfe der Studierenden untereinander.

Den bisherigen optimistisch getönten Ausführungen zu Folge müsste man bei einem empirischen Vergleich beider Notensysteme unter anderem folgende Ergebnisse erwarten.

- Studierende erzielen unter beiden Notensystemen vergleichbar gute Leistungsergebnisse im Seminar.
- Das jeweilige Notensystem hat keine Auswirkungen auf kumulative Studienleistungen, etwa nachfolgende Abschlussexamen.
- Unter einem P/F-System fällt die wahrgenommene Stressbelastung geringer aus.
- Unter einem P/F-System verbessert sich das Sozialverhalten in Richtung geringere Konkurrenz und höherer gegenseitiger Unterstützung.

Studien zum Vergleich zwischen P/F- und A..F- Notensystem

Leider gibt es nur sehr wenige neuere empirische Studien, welche die Auswirkungen eines P/F-Systems im Vergleich zur klassischen Notengebung untersuchten. Rohe et al. 2006 gehen kurz auf einige Untersuchungen ein und kommen hinsichtlich der Leistungsvariablen zu dem

Ergebnis "..research analyzing academic achievement and grading systems suggests little, if any, harm from the pass-fail system...". Diese Einschätzung kann aber offensichtlich nicht für die Studien der 70iger Jahre des vorherigen Jahrhunderts gelten, die deutlich schwächere Leistungsergebnisse unter P/F als unter dem üblichen Benotungssystem erbrachten. Im Folgenden werden zwei dieser Studien etwas genauer beschrieben und eine Gesamteinschätzung der frühen Studien versucht.

Frühe Studien, die auf schwächere Studienleistungen unter P/F-System hindeuten

von Wittich (1972)

Von Wittich (1972) untersuchte an ca. 900 Studierenden die Leistungsergebnisse in elementaren Fremdsprachenkursen einer Universität in Ohio. Ca. ein Drittel der Studierenden wurden hierbei unter dem Pass/Fail-System bewertet. Ziel der Untersuchung war es, die Kursendnoten unter beiden Benotungssystemen zu vergleichen. Nur Studierende mit wenigstens 60 Creditstunden waren überhaupt berechtigt, den Sprachkurs mit P/F- Benotung anzustreben, was unweigerlich zu Unterschieden im Alter (bzw. der Semesteranzahl) unter beiden Notensystemen führte. Da es sich um eine nichtexperimentelle Studie handelt, ergibt sich unweigerlich die Frage, ob beide 'Experimentalgruppen' hinsichtlich aller sonstigen, die Fremdsprachenleistung betreffenden Variablen hinreichend vergleichbar waren. Durch die Aufnahme einiger wichtiger verfügbarer Leistungsdaten (grade-point average, ACT score, und course load) in die Regressionsanalyse wurde dann der Versuch unternommen, die auf potentielle Leistungsfähigkeit zurück gehenden Effekte vom Treatmenteffekt isolieren zu können. An der Universität galt als Kriterium des Bestehens: "the 'pass' level is a D grade". Es wurde im Studentenhandbuch als "passing but unsatisfactory" bezeichnet.

Die Autorin stellt die Ergebnisse für beide Bedingungen in Form von Notenhäufigkeiten dar. Eine Gesamtsicht lässt insgesamt deutlich schwächere Endnoten für die P/F-Bewertung vermuten. Um einen griffigen zentralen Kennwert für beide Notensysteme zu erhalten, habe ich die Daten vom amerikanischen ins deutschen Notensystem (A=1, B=2, ...F=5) konvertiert und anschließend Mittelwerte berechnet. Danach erzielten die Studierenden unter P/F einen Mittelwert von 2.91 und unter A..F von 2.24. Ähnlich deutliche Unterschiede zeigten sich im intraindividuellen Vergleich über verschiedene Kurse hinweg (Noten unter P/F und A..F bei denselben Studierenden).

Tabelle 1: Prozentuale Häufigkeitsverteilung der Noten unter Pass/Fail und A..F Bewertungssystem nach von Wippich 1972

	P/F	A..F
A	5,9	29,2
B	20,7	35,1
C	50,1	23,9
D	22,6	6,3
F	0,7	5,5

In Tabelle 1 erkennt man unschwer die klare Häufung der guten Noten unter A..F und die der mittelmäßigen Noten unter P/F. Ähnliche Ergebnisse zeigten sich übrigens bei Giometti (1976). Das Ausmaß des Effektes ist relativ hoch einzuschätzen, was man an der Korrelation zwischen dem Notensystem und der Endnote von r= .29 erkennen kann. Die Ergebnisse deuten darauf hin, ein P/F-System mit einem Pass-kriterium der Note D ermutige insbeson-

dere gute Studierende dazu, ihre Studierleistungen im entsprechenden Kurs etwas herunter zu fahren. Obwohl beide Treatmentgruppen sehr ähnliche Werte in den sonstigen Leistungsvariablen (z.B. GPA und ACT-Tests) aufwiesen, bleibt die Aussagekraft des Studie schwer abzuschätzen, weil die Zuordnung der Studierenden zu den Bedingungen auf Selbstselektion beruhte und Studierende P/F-Kurse nicht "nach Zufall" auswählen.

Gold, Reilly, Silberman & Lehr (1971)

Die Untersuchung von Gold et. al. 1971 verfolgte einen interessanten experimentellen Ansatz und erfasste auch die Auswirkungen des Notensystems auf die Ergebnisse nachfolgender Semester. Studienanfänger wurden nach ihren Scholastic Aptitude Testwerten (SAT) parallelisiert und dann (vermutlich nach Zufall) nachfolgenden Benotungsbedingungen zugeteilt.

```
1.) EG1: alle Kurse mit P/F-System
2.) EG2: ein Kurs mit P/F, die restlichen Kurse mit A..F
3.) KG : nur A..F
```

Junior-students absolvierten nur die Bedingung 2 oder 3. Noch bevor die Studierenden Kenntnis davon erhielten, welcher Bedingung sie zugeordnet worden waren, sollten sie verbindliche Angaben dazu machen, ob sie für alle Kurse bzw. für welchen Kurs sie das P/F System wählen wollten. Ca. 80 % aller Studierenden wollten mindestens einen P/F Kurs belegen. Das weitere Vorgehen wird leider nicht hinreichend genau beschrieben, dürfte aber vermutlich wie folgt abgelaufen sein: Die Wahl eines P/F-Systems war zunächst freiwillig. Tatsächlich konnte es jedoch nur derjenige Studierende erhalten, der zuvor für diese Bedingung ausgewählt wurde. Als jeweilige Kontrollgruppe dienten dann diejenigen Studierenden, welche ebenfalls ein P/F System anstrebten, es aber wegen der vorheriger Kontrollgruppenzuweisung nicht realisieren durften. Auf diese Weise wurde die Vergleichbarkeit der Experimentalgruppen hinsichtlich wichtiger Störfaktoren der internen Validität hergestellt.

Um das Pass-Kriterium zu erfüllen, musste mindestens die Note D (im Deutschen Notensystem 4) erzielt werden. Die jeweiligen Dozenten ermittelten für alle Kursteilnehmer die üblichen Noten, die dann unter der P/F Bedingung lediglich als bestanden/nicht bestanden gewertet wurden. Als Leistungskriterium für die Bedingung EG1 ["alle Kurse unter P/F-System"] diente der GPA (=Durchschnitt aus den konventionellen Noten aller Kurse eines Semesters). Hierbei erzielten die Studierenden unter EG1-Bedingungung mit der Gesamtnote C- signifikant schwächere Durchschnittsnoten als die KG mit der Gesamtnote C+. Das bedeutet unter anderem., dass diejenigen Studierenden, welche eine vollständige P/F-Bewertung anstrebten, sie aber nicht erhielten, bessere Leistungen erzielten, weil man ihnen diese Wahl verweigerte und sie stattdessen dem konventionellen Notensystem unterwarf. Besonders stark waren die Leistungseinbußen unter P/F-System bei Studierenden mit hoher Leistungsfähigkeit (oberes Drittel im SAT), was unter rein ökonomischen Nutzenerwägungen des gewährten Leistungsnachweises verständlich ist. Beim Vergleich von EG2 gegenüber der KG ergaben sich nur bei Zusammenlegung aller verfügbaren Studenten schwächere Leistungen unter der Bedingung P/F. D.h. EG2-Studierende, die nur in einem Kurs nach P/F- bewertet wurden, erzielten in diesem Kurs schwächere Noten als die jeweilige Kontrollgruppe, in allen anderen Kursen hingegen nicht. Zudem erreichten die Studierenden in ihren konventionell bewerteten Kursen insgesamt signifikant bessere Ergebnisse als in ihrem P/F bewerteten Kurs, was allerdings auch für die Kontrollgruppe im Hinblick auf den Vergleich "P/F angestrebt vs. sonstige Kurse" zutraf und letztlich darauf hindeutet, Studierende würden bei freier Wahl des Benotungssystems vornehmlich solche Seminare für eine P/F-Wertung auswählen, bei denen sie aus welchen Gründen auch immer schlechter abschneiden. Mit dieser Strategie lässt sich eine etwas bessere Semesterdurchschnittsnote erzielen.

Nachdem ab dem zweiten Semester das konventionelle Notensystem flächendeckend wieder eingeführt wurde, sollten die Erhebungen des GPA im zweiten und dritten Semester Auskunft über die langfristige Wirkung des umfassenden P/F-Systems der EG1 geben. Hierbei schnitt die EG1 im zweiten Semester noch signifikant schlechter ab als die KG. Erst im dritten Semester glichen sich die Leistungen beider Gruppen wieder an. Insgesamt belegt die Untersuchung auf methodisch hohem Niveau relativ konsistente Leistungseinbußen durch die Einführung eines P/F-Notensystems, die sich bei Ausdehnung auf alle Seminare eines Semesters noch etliche Zeit negativ auf das kumulative Lernen auswirken könnten.

Weitere Hinweise und Gesamteinschätzung der früheren Studien

Weitere Studien (z.B. Karlins, Kaplan &. Stuart 1969, Stallings & Schmock 1971, Reiner & Jung 1972, Otto 1972, Giometti 1976, Bain et al. 1972, siehe auch den Review von Davidovicz (1972) belegen sehr konsistent schwächere Leistungsergebnisse unter einer P/F-Benotung, wobei als P/F-Cut-off stets die Note D zur Anwendung kam. Weber (1974) berichtet von Erfahrungen mit dem P/F-System in etlichen Universitäten. Hierbei ergaben sich meist deutlich geringere Leistungen in P/F-Kursen, die stellenweise eine ganze Notenstufe umfassten. Nach Karlins, Kaplan &. Stuart (1969) erzielten Studierende der Princeton-Universität in ihren P/F-Kursen im Durchschnitt eine Note, die ca. 0.8 Notenstufen schlechter ausfiel als ihre Durchschnittsnote unter dem mehrstufigen Bewertungssystem. Dieser Unterschied ist praktisch sehr bedeutsam, entspricht etwa einer Effektstärke von d = 1 und ist wegen der hohen Probandenzahl von mehr als 2000 auch ziemlich zuverlässig einzustufen. Giometti 1976 stellte ähnlich große Leistungsdifferenzen fest.

Gelegentlich wurden subjektive Einschätzungen der Studierenden erhoben, die sich jedoch meist auf Itemniveau bewegten. Karlins et al. 1969 gewähren einen sehr interessanten Einblick in die studentischen Bewertungen zum P/F-Systems. Danach investierten Studierende eigenen Angaben zufolge erkennbar weniger Lerneinsatz für die P/F-Kurse, fühlen sich aber auch weniger unter Stress gesetzt. Sie gaben an, die eingesparte Lernzeit für die benoteten Seminare einzusetzen. Insgesamt favorisierten Studierende und die meisten Lehrenden eher das P/F-System. Manchmal wurde der Frage nachgegangen, ob die schwächeren Leistungen in den P/F-Seminaren durch bessere Leistungen in den sonstigen Kursen kompensiert worden seien. Ein derart optimistisches Szenario ließ sich aber nicht bestätigen. Desgleichen habe ich keine überzeugenden Befunde entdecken können, die durch eine P/F-Wahl eine deutliche Veränderung der Kurswahlen nach sich gezogen hätten, was der Ansicht widerspricht, Studierende würden durch das traditionelle Notensystem dazu genötigt, bestimmte (z.B. die leichten oder nur aufs Hauptfach bezogene) Kurse zu bevorzugen. Hierbei gilt auch zu bedenken, dass Studienpläne meist nur eine sehr begrenzte Wahlfreiheit gewähren. In der Studie von Gold et al.1969 gaben Studierende an, wegen des P/F-Systems keine andere Kurswahl vorgenommen zu haben. Davidovicz (1972) kommt in seinem Review zu der Einschätzung, Studierende suchten P/F-kurse eher aus, um sich die Arbeit im Semester zu erleichtern. Bei freier P/F-Kursoption wählen Studierende vornehmlich unbeliebte, aber notwendige Seminare, in denen sie eher schwache Noten erwarten und auch erzielten. Die Studierenden tun dies häufig, um ihren Notendurchschnitt möglichst hoch zu halten - weil P/F-Kurse meistens nicht in den GPA eingehen - und weniger, um einen anspruchsvollen Kurs zu belegen. Dieses Wahlverhalten gefährdet im Übrigen die interne Validität nicht experimenteller Studien beim einfachen Vergleich beider Notensysteme mit vorgegebenen Gruppen (=Selbstselektion der Probanden zu ihren Bedingungen). Deshalb sind auch die relativ großen Leistungsunterschiede zwischen den Notensystemen in solchen Fällen nicht ausschließlich der Bewertungsmethode anzulasten, da Kurswahl und Notensystem zu ungunsten von P/F konfúndiert zu sein scheinen.

Insgesamt komme ich auf der Basis der empirischen Studien zu der eindeutigen Gesamteinschätzung, man müsse im Mittel mit signifikanten und zum Teil recht deutlichen Leistungsverlusten rechnen, wenn ein marginales P/F-System eingeführt wird, welches für das Bestehen die Note "ausreichend" zu Grunde legt. Vielleicht liefert die Klarheit und Konsistenz der Befunde eine Erklärung dafür, warum die Forschung zu dieser Thematik in den nachfolgenden Jahren praktisch zum Erliegen kam.

Neuere Studien zur Wirkung eines Pass-Fail-Notensystems

Neuere Studien zum Vergleich des traditionellen mit dem P/F- Notensystem habe ich ausschließlich im Bereich des Medizinstudiums gefunden. Etliche Universitäten in Amerika sind dazu übergegangen, ein P/F System in den ersten Semestern des Medizinstudium einzuführen Diesen Wechsel des Notensystems haben einige Forscher dazu genutzt, die Auswirkungen auf die Studierleistungen und Befindlichkeiten der Studierenden zu überprüfen. Hierbei fällt auf, dass der Cut-off für das Bestehen mit ca. 70% korrekter Lösungen in der Regel deutlich höher liegt als in den früheren Studien. Auch die durchschnittlichen Examensleistungen von ca. 85% korrekter Lösungen oder mehr erreichen dort ein mir schwer nachvollziehbares, exzellentes Niveau. Im Vergleich zu den frühen Studien legten die Forscher neben der Leistungsvariablen besonderes Gewicht auf die erhofften Änderungen im subjektiven Erleben und setzten entsprechende Messverfahren ein. Einige Studien werden etwas ausführlicher dargestellt und abschließend eine Gesamtbewertung angeschlossen.

Bloodgood, Short, Jackson & Martindale (2009)

Bloodgood, Short, Jackson & Martindale (2009) untersuchten die Auswirkungen eines Pass/Fail Benotungssystems im Medizinstudium auf die akademische Leistung und das Wohlbefinden der Studierenden. Dazu nutzen die Autoren eine Art Kohortendesign, indem sie zwei sukzessiv aufeinander folgende Studienjahrgänge mit unterschiedlichen Bedingungen gegeneinander testeten. Während im Vorgängerjahrgang noch die klassische Notengebung (A..F: A,B,C,D,F) zur Anwendung kam, wurde das Benotungssystem im nächsten Studienjahrgang auf Pass/Fail umgestellt. Persönlichen Angaben von Bloodgood in einer Email vom 6.4.2010 zufolge, wurden alle Kurse dazu ermutigt, einen pass-fail cut-off von 70% korrekter Lösungen anzuwenden, was mir als recht strenges Passkriterium anmutet. Ca. 10 % aller Studierenden erreichten das erforderliche Kriterium auch nach möglichen Prüfungswiederholungen überhaupt nicht. Die P/F Bedingung kam allerdings nicht in Reinform zum Einsatz, weil zusätzlich eine Honor-Variante eingeführt wurde, die ab dem vierten Semester den 20% besten Studierenden verliehen wurde. Jeder Studiengang umfasste ca. 140 Studierende. Tabelle 2 soll das Vorgehen verdeutlichen.

Tabelle 2: Versuchplanformalisierung der Studie von Bloodgood et al. 2009

		S1		S2		S3		S4	
	N	A..F	O	A..F	O	A..F	O	A..F	O
+ 1J	N	P/F	O	P/F	O	P/F	O	P/F	O
								+ honors	

Zwei Studienjahrgänge mit unterschiedlichen Benotungssystemen wurden, um ein Jahr versetzt (+1J), 4 Semester (S1..S4) lang untersucht, wobei jeweils Messungen (O) erhoben wurden. Die Achillesferse eines solchen Versuchsplans liegt in der Unklarheit bezüglich der Vergleichbarkeit beider Gruppen hinsichtlich aller Treatment relevanter Variablen (N=nonequivalent groups). Da aber wichtige Leistungsvariablen wie der undergraduate grade point average oder vorhandene Testergebnisse in Biologie und Physik keine signifikanten

Unterschiede ergaben, können zumindest einige für den Leistungsvergleich wichtige Störfaktoren definitiv ausgeschlossen werden.

In jedem Semester wurden subjektive Daten zum Wohlbefinden (Dupuy Schedule) erhoben, die insgesamt, sowie differenziert nach etlichen Facetten (Well-being, Depression, Angst, usw.) ausgewertet wurden. Im Verlauf der untersuchten Semester sowie später fielen etliche Leistungsdaten an. Von besonderer Relevanz erscheinen die Kurs- bzw. Seminarergebnisse unter beiden Benotungssystemen, da diese eine direkte Auswirkung der Benotungsmethode auf die aktuellen Studierleistungen in den Seminaren reflektieren. Obwohl unter Pass/Fail die tatsächlich erbrachte Leistung nur im Hinblick auf das Kriterium "bestanden" relevant erscheint, hatten die Dozenten stets auch den Prozentsatz der korrekten Lösungen ermittelt, der einen tatsächlichen Leistungsvergleich mit der 5-sufigen Notenskala ermöglicht. Für jeden Studierenden wurde aus allen Kursen der 4 Semester ein umfassender Durchschnittswert des Prozentsatzes der korrekten Lösungen aus Quiz, Examen, Labortests und sonstigen Prüfungsleistungen ermittelt. Die Mittelwerte dieser Gesamtkursergebnisse fielen mit 87,1 und 87,5 unter beiden Benotungssystemen fast identisch aus. Insgesamt ergaben sich hoch vergleichbare Ergebnisse objektiver akademischer Leistungen in nachfolgenden Testgebieten, die teilweise erst in der Folgezeit erhoben wurden:

- course performance in the first two years of the curriculum,
- United States Medical Licensing Examination (Step 1) scores,
- clerkship grades,
- United States Medical Licensing Examination (Step 2) Clinical Knowledge scores
- residency placement

Bei den subjektiven Daten sind teilweise erhebliche Ausfälle zu konstatieren, die dann weniger gravierend einzustufen sind, wenn sie nicht auf differenziellen Ausfallquoten basieren, sondern eher zufällig zustande kommen, was ich einmal annehmen möchte. Der Gesamtscore der General Well-Being Schedule (Dupuy Schedule), verstanden als ein umfassendes Maß für das Wohlbefinden, erbrachte in den ersten 3 Semestern stets signifikant günstigere Werte für die Benotung Pass/Fail, die vom Ausmaß her zwischen einer Effektstärke von d= 0.13 bis d= 0.6 liegen. Dieser Vorteil einer Pass/Fail-Notengebung in der Wahrnehmung des subjektiven Wohlbefindens ließ sich auch für die meisten Subtests der Dupay-Schedule, insbesondere Angst, Depression, positive well-being und Vitalität nachweisen. Die Studierenden unter Pass/Fail-Bedingung gaben darüber hinaus in den ersten 3 Semestern signifikant bessere Bewertungen für die Qualität der medizinischen Ausbildung ab und waren mit ihrem persönlichen Leben in den letzten Monaten zufriedener. Im vierten Semester jedoch fanden sich keinerlei Unterschiede zwischen beiden Benotungssystemen in der gesamten subjektiven Bewertung mehr, was die Autoren auf das im 4. Semester wirksame Honor-System und das anstehende United States Medical Licensing Examination zurückführten. Die Honor-Variante berücksichtigte alle Kursleistungen in den ersten 4 Semestern und teilte offenbar schon recht früh die Studierenden in zwei virtuelle Gruppen auf: diejenigen, die entsprechende Ambitionen hegten und solche, die mit einem klassischen P/F-System zufrieden waren. Wie eine zusätzliche Erhebung ergab, fühlten sich die Honor-Anstreber mehr unter Stress gesetzt. Es bleibt zu vermuten, das subjektive Wohlbefinden gestalte sich unter striktem P/F noch günstiger. Zugleich bleibt dann nicht ganz auszuschließen, die Leistungsergebnisse könnten so auch etwas schwächer ausfallen.

Ohne die irritierende Honor-Bedingung würde die Schlussfolgerung aus dieser Studie ziemlich eindeutig lauten: P/F in den ersten beiden Studienjahren erhöht die Lebensqualität ohne Abstriche an der Studierleistung nach sich zu ziehen. Sollten sich derartige Ergebnisse bestä-

tigen lassen, dann wären sie als fundiertes Argument für die Einführung des P/F-Notensystems zu werten.

Rohe et al.. (2006)

Rohe et. al. 2006 interessierten sich vornehmlich für die subjektiven Auswirkungen eines P/F-Systems und erwarteten günstigere Emotionen hinsichtlich des wahrgenommenen Stresses und der Prüfungsängstlichkeit sowie eine verbesserte Gruppenkohäsion der Studierenden. Leider wird das P/F-System nicht hinreichend beschrieben. Die Autoren bezeichnen Ihr System nicht als striktes, sondern als pass marginal -pass-fail system, welches im Endeffekt aber ein P/F darstelle. Leider finde ich auch keine Angaben zum Pass-Kriterium, was die Interpretation der Befunde fast unmöglich macht. Marginal pass deutet eher auf die Note D hin, normales pass offenbar auf eine bessere Note, die man nach einem marginal pass durch eine Prüfungswiederholung erzielen konnte. Letztlich entnehme ich dem Artikel, durch die Einführung des neuen Systems sei auf jeden Fall die klassische Notengebung abgeschafft worden und die erbrachte Leistung könne nicht höher als Pass bewertet werden. Der Versuchsplan basiert ebenfalls auf einem Vergleich zweier Zeit versetzter Studienjahrgänge von jeweils ca. 40 Medizinstudierenden und ist in Tabelle 3 etwas genauer beschrieben.

Tabelle 3: Versuchsplan bei Rohe et al. 2006

		Jahr 1		Jahr 2	
	N	A..F	O	A..F	O
+ 1J	N	P/F	O	A..F	O

Während ein Studienjahrgang durchgehend dem üblichen gestuften Benotungsystem (A..F) unterlag, galt für die Studierenden des nachfolgenden Jahrgangs im ersten Studienjahr das P/F-System, das im zweiten Studienjahrgang allerdings wieder durch das mehrstufige Notensystem ersetzt wurde. Am Ende jedes Studienjahres fanden die Erhebungen zu den subjektiven Einschätzungen statt. Als Leistungskriterium diente das erste United States Medical Licensing Examination (USMLE) nach dem zweiten Studienjahr. Für die Vergleichbarkeit der kognitiven Leistungen beider Studienjahrgänge zu Beginn des Studiums sprechen sehr ähnliche Testwerte im grade point average sowie im Medical College Admission Test. Da sich die späteren Leistungen beider Gruppen im ersten USMLE auch nicht signifikant unterscheiden, erscheint die Interpretation gerechtfertigt, das P/F-System ziehe keine schwächeren Ergebnisse in nachfolgenden Examensleistungen nach sich. Allerdings fehlen Angaben zu den Examensleistungen der entsprechenden Kurse während der Intervention. Sowohl im ersten, wie im zweiten Studienjahr erzielten die Studierenden unter P/F aber signifikant günstigere Ergebnisse in der Perceived Stress Scale, dem Profile of Mood States und der Perceived Cohesion Scale (PCS), jedoch nicht in der Test Attitude Inventory (TAI von Spielberger). Während mir die vorteilhafteren Emotionen unter P/F im ersten Jahr theoriekonform erscheinen, hätte man im zweiten Jahr mit einer Verschlechterung der subjektiven Befindlichkeiten auf das Niveau der ersten Treatmentgruppe rechnen müssen, weil das mehrstufige Notensystem ja wieder eingeführt wurde. Da zu Beginn des Studiums keine subjektiven Daten erhoben wurden, bleibt nicht ausgeschlossen, beide Gruppen hätten sich schon von Anfang an in diesen Werten unterschieden. Es bedarf hier schon einiger theoretischer Verrenkungen, die Stabilität der Emotionen trotz Einführung des üblichen Notensystems hinreichend zu erklären, z.B.: "Die Studierenden sind im ersten Jahr derart gefestigt worden, dass objektiv zunehmende Belastung den Stress nicht mehr weiter erhöht."

White & Fantone (2009)

White und Fantone 2009 verglichen ein mehrstufiges Benotungssystem (Honors, High Pass, Pass, Fail) - dort discriminating grading genannt, hier auf grades verkürzt-, mit einem klassischen P/F System. Der Versuchsplan in Tabelle 4 soll das Vorgehen näher verdeutlichen

Tabelle 4: Versuchsplan bei White und Fantone 2009

		Jahr 1	Jahr 2
	N	P/F	grades
+1J	N	P/F	P/F

Alle Studierenden absolvierten zunächst das erste Studienjahr unter einem sehr anspruchsvollen P/F System (Pass >=75%). Üblicherweise folgte dann im zweiten Studienjahr das mehrstufige Benotungssystem, welches jedoch später wieder zugunsten eines P/F-Systems rückgängig gemacht wurde. Dadurch konnte ein Studienjahrgang unter mehrstufiger Benotung mit dem nachfolgenden Studienjahrgang unter P/F, jetzt mit Passkriterium 70%, verglichen werden. Der mit Daten unterlegte Vergleich bezieht sich für beide Bedingungen jeweils auf das zweite Studienjahr. Jeder Studienjahrgang umfasste ca. 170 Studierende. Zunächst belegen fast identische Testwerte im GPA und einem medizinischen Test (Medical College Admission Test.) sehr ähnliche Ausgangsbedingungen hinsichtlich der relevanten Leistungsfähigkeit beider Studienjahrgänge.

Als unmittelbare Leistungsvariablen fungierten alle Kursergebnisse im zweiten Studienjahr. Diese ließen nur unwesentliche Unterschiede zwischen beiden Bedingungen erkennen (keine Unterschiede in 8 Kursen, signifikant bessere Ergebnisse für grades in zwei, für P/f in einem Kurs). Alle Kursmittelwerte fielen im Übrigen mit geringfügigen Schwankungen um 90% recht hoch aus. Beide Jahrgänge erzielten darüber hinaus hoch vergleichbare Leistungen in den nachfolgenden United States Medical Licensing Examination (Step 1 und 2) sowie vergleichbar gute Angebote beim residency placement. Die Leistungsergebnisse belegen insgesamt, die Beibehaltung des des P/F-Systems habe keine negativen Auswirkungen auf die Seminarleistungen, nachfolgende Examensleistungen oder Ausbildungschancen ausgeübt.

Zur Zufriedenheit mit dem Bewertungssystem lagen Daten für beide Studienjahrgänge zu beiden Jahren vor. Dabei hatten die Studierenden auf einer Ratingskala ihre Zufriedenheit mit demjenigen Notensystem einzuschätzen, dem sie jeweils unterworfen waren. Alle möglichen Vergleiche sprechen eindeutig für höhere Zufriedenheit mit dem P/F-System, die vom Ausmaß des Effektes sehr hoher praktischer Bedeutsamkeit entsprechen. Erhebungen beim ausschließlichen Studiengang unter P/F-System ergaben weitgehende Zustimmung hinsichtlich der Wahrnehmung eines erhöhten Spielraums für weitere Aktivitäten im akademischen, sozialen oder persönlichen Bereich. Es bleibt aber schwer abzuschätzen, ob und wie weit dieser empfundene Spielraum gegeben war und auch tatsächlich genutzt wurde. Leider kamen keine professionellen Messverfahren zum wahrgenommen Stress oder der subjektiven Lebenszufriedenheit zum Einsatz, die - wären sie kontinuierlich in beiden Studienjahrgängen erhoben worden - eine bessere Einschätzung erlaubt hätten, ob ein P/F-System die subjektive Lebensqualität nachweislich verbessert. Insgesamt belegt die Studie jedoch ziemlich überzeugend, das P/F-System mit einem Passkriterium von 70% könne die Zufriedenheit mit der praktizierten Leistungsbewertung deutlich verbessern, ohne dabei die akademischen Leistungen zu schmälern.

Gesamteinschätzung neuere Studien zum Medizinstudium

Auch einige sonstige, mir zugängliche bzw. von den oben genannten Autoren zitierte Studien aus dem Bereich des Medizinstudiums scheinen die vergleichbaren Leistungseffekte beider Notensysteme weitgehend zu bestätigen. So fanden Robins et al. (1995) keine Leistungsunterschiede im Midterm- und Abschlussexamen zwischen Medizinstudenten unter einem P/F- und einem mehrstufigen Benotungssystem. Um bei P/F System ein "bestanden" zu erzielen, waren allerdings 75% korrekte Lösungen notwendig. Außerdem mussten die Studierende mit P/F Benotungssystem wöchentliche Quiz über sich ergehen lassen, die insgesamt 30% der Seminarwertung umfassten. Die Examensleistungen lagen alle signifikant über der Cut-off Grenze und die Autoren sahen keine Anhaltspunkte dafür, die Studierenden hätten gerade so viel gelernt, um die Pass-hürde nehmen zu können. Das wäre allerdings bei einem 75% Kriterium auch recht riskant. Zwar werden gelegentlich auch negative Resultate für P/F berichtet, wobei mir jedoch die Methoden recht zweifelhaft vorkommen, etwa Vergleiche zwischen Universitäten mit verschiedenen Notensystemen, Korrelationen zwischen Noten und weiteren Kriterien usw. Insgesamt deuten die Ergebnisse darauf hin, bei einem anspruchsvollen Cut-off-Kriterium von ca. 70% korrekter Lösungen würden gegenüber dem traditionellen Notensystem vergleichbare Leistungen erzielt und das subjektive Wohlbefinden der Studierenden eher gefördert.

Abschließende Bewertung

Die Befunde deuten insgesamt in die Richtung, die Auswirkung eines P/F-Systems auf die Studienleistung hänge entscheidend vom Pass-Kriterium ab. Allerdings habe ich keine Studie gefunden, welche verschiedene Pass- Cut-off -Werte im direkten Vergleich gegeneinander analysiert und so eine bessere Bewertungsgrundlage ermöglicht hätte. Jedoch sprechen theoretisch plausible Überlegungen relativ klar für bessere Mittelwertsleistungen mit wachsendem Kriteriumsanspruch an das Bestehen. Dadurch dürften vor allem die fähigen Studierenden eher dazu bewegt werden, mehr Lerneinsatz zu zeigen, da das Seminar für diese nicht mehr im Schongang zu absolvieren ist. Schließlich könnte man das Pass-Kriterium derart adjustieren, dass im Mittel in etwa gleiche Leistungsanforderungen wie im traditionellen Notensystem herauskommen müssten. Dann wäre zwar der klassische Konkurrenzdruck ausgeschaltet, allerdings schwer zu begründen, warum die subjektive Belastung sinken sollte, weil ein hohes Kriterium ebenfalls einen Leistungsdruck ausübt. Eine andere Möglichkeit, gute Studierende selbst bei relativ geringem Pass-Kriterium besser zu motivieren, bestünde darin, den Prozentsatz korrekter Lösungen rückzumelden, um so das reine Leistungsmotiv anzuregen.

Hinreichende Befunde belegen recht konsistent deutliche Leistungseinbußen durch den Wechsel vom traditionellen Notensystem zu einem P/F-System mit Pass-Kriterium von "ausreichend". Damit liefern diese Befunde im Übrigen eine deutliche empirische Bestätigung für die leistungssteigernde Wirkung der traditionellen Notengebung und lassen quasi denjenigen Leistungsunterschied einschätzen, der gegenüber einer schulischen Leistungsmessung (z.B. Quiz, Examen) ohne jede Benotung mindestens erzielt werden müsste. Es bleibt schwer einzuschätzen, wie nachhaltig der Leistungsabfall durch die P/F Kurse bestehen bleibt, da diese Frage zu selten untersucht wurde. Gold et. al. 1971 fanden zwar zunächst verbleibende Leistungsschwächen, die aber nach Einführung der traditionellen Benotung nach einem Studienjahr wieder aufgeholt wurden. Etliche optimistische Erwartungen, die mit der Einführung eines P/F-Systems verbunden waren, wie ein annähernd vergleichbarer Lerneinsatz wie bei üblicher Benotung, die häufigere Wahl fachferner oder anspruchsvoller Kurse aus echtem

Interesse, eine Steigerung der intrinsischen Motivation usw., sind entweder nicht adäquat untersucht worden oder haben sich meistens als unhaltbar erwiesen. Die verfügbare, bzw. die vom Studierenden insgesamt für das Lernen vorgesehene Lernzeit wird günstigstenfalls von den P/F-Kursen auf die benoteten Kurse verlagert, um den Gesamtschnitt zu verbessern. Dieser mögliche, verstärkte Lerneinsatz reichte aber nicht aus, um in den benoteten Kursen auch deutlich bessere Ergebnisse zu erzielen.

Neuere Studien, welche sich ausschließlich auf das Medizinstudium beziehen, stellten vergleichbare Leistungsergebnisse unter mehrstufigem- und P/F-System fest und darüber hinaus auch keinerlei Einbußen bei Examen in der Folgezeit. Hierbei erforderte das Passkriterium stets mindestens 70% korrekter Lösungen, was übertragen auf das traditionelle Notensystem wenigstens der Note "befriedigend" entspricht. Bei diesen Studien konnten für die P/F-Systeme auch häufiger positivere Einschätzungen im Sinne eines geringeren wahrgenommen Stresses bzw. eines besseren Wohlbefindens gesichert werden. Die Ergebnisse zur Prüfungsängstlichkeit fallen nicht eindeutig aus. Hinsichtlich der aktuellen Angst vor Prüfungen fehlen entsprechende Daten. Auf der Basis meiner Forschungen zur aktuellen Prüfungsangst hege ich erhebliche Zweifel, durch ein P/F-System die aktuelle Angst vor einer Prüfung signifikant senken zu können (Jacobs 2011, S.16). Diese Erkenntnis steht aber nicht im Widerspruch zur Annahme, während des Seminars könne die subjektiv erlebte Belastung geringer ausfallen. Es konnte nur eine Studie gefunden werden, welche gewisse Aussagen über die Veränderung des sozialen Klimas erlaubt. Sie bestätigte eine verbesserte Gruppenkohäsion durch Einführung des P/F-Systems. Obwohl auch die neueren Studien teilweise einige methodische Mängel beinhalten und weitere Untersuchungen dringend erforderlich sind, vermitteln sie insgesamt eine gewisse Hoffnung, es könne durchaus gelingen, durch ein verändertes Notensystem eine höhere Lebensqualität zu erzielen, ohne dafür Leistungseinbußen in Kauf nehmen zu müssen.

Sowohl in den früheren, wie in den späteren Studien wurde das P/F-System von Studierenden als die bessere oder zumindest wünschenswertere Bewertungsmethode favorisiert. Bei der Bewerbung um Arbeitsstellen sind Studierende mit P/F-System mitunter benachteiligt, weil ihre Leistungen gegenüber herkömmlicher Notengebung eine geringere Differenzierung erkennen lassen. Dieses Problem ist bei flächendeckender Einführung eines bestimmten Notensystems natürlich nicht mehr relevant.

Die angemessene Bewertung eines Notensystems sollte sich an mehreren Variablen orientieren und darf die Studierleistung nicht zum einzigen Maßstab erheben. Leistung spricht nicht für sich selbst, sondern muss auch unter dem Aspekt betrachtet werden, unter welchen Bedingungen sie zustande kommt. Es entbehrt zwar hinreichender empirischer Evidenz, anzunehmen, mehr Freiheit durch ein weniger strenges Notensystem erhöhe das Interesse und die so erzeugte intrinsische Motivation entfache eine Lernfreude, welche quasi automatisch Lernen und Leistung beflügelte. Studieren ist anstrengend und nicht selten mühsam. Es entspricht somit einem intelligenten ökonomischen Verhalten, bei sehr starker Beanspruchung die Lernprioritäten auf die konventionell benoteten Seminare zu konzentrieren sowie auch eine angemessene Zeit für die Erholung einzuplanen. Die Studierleistung in einem Seminar lässt sich zwar ganz einfach und ziemlich sicher durch entsprechenden Notendruck steigern (Jacobs 2009). Es stellt sich aber die Frage, welchen Preis an Stress, Lebensqualität, Wohlbefinden und sozialem Klima man dafür zahlen will. Hierbei bleibt vornehmlich zu überlegen, ob einige marginale P/F-benotete Kurse langfristig gravierende Leistungslücken hinterließen, welche die Bildungsqualität entscheidend schwächten. Dies erscheint mir selbst bei gänzlich unbenoteten Kursen eher unwahrscheinlich. Es gilt daher, beim Umfang konventionell benoteter Leistungen das rechte Maß walten zu lassen und die Gesamtanforderungen eines Studiums zumindest erträglich zu gestalten.

Literatur

Bain, P. T.; Hales, L. W. & Rand, L. P (1973). n Investigation of Some Assumptions and Characteristics of the Pass-Fail Grading System Journal of Educational Research, 67, (3), 134-136.

Bloodgood, R. A., Short, J. G., Jackson, J. M. & Martindale J. R. (2009). A Change to Pass/Fail Grading in the First Two Years at One Medical School Results in Improved Psychological Well-Being. Academic Medicine 84 (5). 655-662.

Davidovicz, H. M. (1972). Pass-Fail Grading - A Review. Abstract and Reviews of Research in Higher Education, 17, 1-11.ERIC Full Text

Giometti, T. (1976). One Experience with the Pass-Fail Grading System Hispania, 59,(2) 302-307

Gold, R. M., Reilly, A. , Silberman R. & Lehr R. (1971). Academic Achievement Declines under Pass-Fail Grading. The Journal of Experimental Education, 39, (3), 17-21

Jacobs, B. (2009).Leistungssteigerung durch Notendruck? - Die Wirkung der Benotung auf die Studierleistungen in einem Seminar.
URN: urn:nbn:de:bsz:291-psydok-25299
URL: http://psydok.sulb.uni-saarland.de/volltexte/2009/2529/

Jacobs, B. (2011). Musterlösungen durcharbeiten als Alternative zu Testen mit Feedback - Eine Replikationsstudie.
URN: urn:nbn:de:bsz:291-psydok-27127
URL: http://psydok.sulb.uni-saarland.de/volltexte/2011/2712/
[Quelle nach dem 14.5. 2010 eingefügt]

Karlins, M, Kaplan, M. &. Stuart, W. (1969).Academic Attitudes and Performance as a Function of Differential Grading Systems: An Evaluation of Princeton's Pass-Fail System. The Journal of Experimental Education, 37 (3) 38-50

Otto, D. J. (1972). A Study of the Pass/Fail Grading System. University of Alberta (Canada). ERIC #: ED077472
http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED077472

Reiner, J. R. & Lorne B. Jung, L.B. (1972). Enrollment patterns and academic performance as a function of registration under a pass-fail grading system . Interchange, 3(1), 53-62

Robins, L S., Fantone, J. C., Oh, M. S., Alexander, G. L., Shlafer, M., Davis, W K. (1995) The effect of pass/fail grading and weekly quizzes on first-year students' performances and satisfaction.Academic Medicine, 70(4), 327-329.

Rohe, D. E., Barrier; P. A., Clark, M.M. Cook, D. A., Vickers, K.S. & Decker, P. A. (2006). The Benefits of Pass-Fail Grading on Stress, Mood, and Group Cohesion. Mayo in Medical Students. Clinic Proceedings. 81 (11), 1443-1448.

http://www.mayoclinicproceedings.com/content/81/11/1443.full

Stallings, W.M. & Smock, H.R. (1971). The Pass-Fail Grading option at a state University: A five Semester Evaluation. Journal of Educational Measurement, 8 (3), 153-160)

von Wittich, B (1972) The Impact of the Pass-Fail System upon Achievement of College Students Journal of Higher Education, Vol. 43(6), Jun, 1972. pp. 499-508.

Weber, C. A. (1974). Pass/Fail: Does It Work? NASSP Bulletin 58, 104 bul.sagepub.com/cgi/reprint/58/381/104.pdf [15.4.2010]

White, C. B. & Fantone, J. C. (2009). Pass–fail grading: laying the foundation for self-regulated learning. Advances in Health Sciences Education, published online, Springer

---