Verbundprojekt
„Steuerung von Schulen des Zweiten Bildungswegs
(Schulen für Erwachsene) in Hessen“

Ergebnisse des Teilprojektes
„Zentrale Vergleichsarbeiten an den Schulen für Erwachsene“

Christoph Fuhrmann
Klaus Harney
Sascha Koch

Ruhr-Universität Bochum
Institut für Pädagogik
Lehrstuhl Berufs- und Wirtschaftspädagogik
Bochum, Juli 2005

Inhalt

Einleitung

Das Projekt „Steuerung von Schulen des Zweiten Bildungswegs (Schulen für Erwachsene) in Hessen“ begleitet die Einführung der Neuen Verwaltungssteuerung (NVS) an den Schulen für Erwachsene (SfE) in den Jahren 2003 bis 2006. Das Gesamtprojekt „Bildungssteuerung“ setzt sich aus verschiedenen Bausteinen zusammen, die jeweils für sich spezifische Fragestellungen im Gesamtkontext verfolgen (für weitere Informationen: www.rub.de/sfe-hessen). Der vorliegende Bericht schließt das Teilprojekt „Zentrale Vergleichsarbeiten an den Schulen für Erwachsene“ ab und resümiert die Anstrengungen auf einem Gebiet, das nicht nur für Hessen relevant ist: Zentrale Abschlussprüfungen.

Die hessische Landesregierung hat beschlossen, zentrale Abschlussprüfungen ab dem Jahr 2006 verbindlich einzuführen. Davon sind auch die Schulen für Erwachsene betroffen, die seit dem Schuljahr 2003/04 zentrale Vergleichsarbeiten durchführen, um die entsprechenden Verfahrensroutinen aufzubauen. Die zentralen Vergleichsarbeiten an den SfE haben somit einen doppelten Kontext. Zum einen erzeugen sie als Teil der Neuen Verwaltungssteuerung so genannte organisationsbezogene Kennziffern und sind somit wesentlicher Teil der angestrebten „Output-Steuerung“. Zum anderen sind sie aber auch Teil einer bundesweiten Entwicklung, die seit den Leistungsvergleichsstudien - wie z. B. TIMSS und PISA - die Frage der Qualitätssicherung und der bundesweiten Vergleichbarkeit von Schulabschlüssen in den Vordergrund bildungspolitischer Maßnahmen stellt. Zentrale Abschlussprüfungen und Lernstandserhebungen sind Teil einer Entwicklung, die bundesländerübergreifend stattfindet und Schulverwaltungen ebenso wie Lehrkräfte in den Schulen vor Ort in den nächsten Jahren intensiv beschäftigen wird.

Die zentralen Vergleichsarbeiten der Schulen für Erwachsene wurden von eigens dafür eingerichteten Arbeitsgruppen der an den SfE tätigen Lehrkräfte mit hohem Arbeitsaufwand erstellt. Die Aufgabe des Projektes sollte es sein, die Vergleichsarbeiten, d. h. insbesondere die Arbeitsaufgaben, in zweierlei Hinsicht zu untersuchen: Halten die einzelnen Arbeitsaufgaben dem Anspruch der Vergleichbarkeit stand und kann man mit ihnen einen Aufgabenpool einrichten, der die Erstellung zukünftiger zentraler und dezentraler Vergleichsarbeiten vereinfacht?

Dass die Neue Verwaltungssteuerung mit den von ihr angestrebten Zielen und Mitteln keineswegs unumstritten ist, davon bleibt ein Projekt wie das vorliegende natürlich nicht unberührt. Es bearbeitet nicht nur den Gegenstand, sondern ist auch in diskursive Auseinandersetzungen eingebunden. Im vorliegenden Fall hatte dies auch Folgen für das Projekt selbst: Die Frage der Umsetzbarkeit von Projektergebnissen in eine veränderte Professionspraxis wurde selbst zu einer eigenständigen Fragestellung. Um dies angemessen darzustellen, wurde die dem Bericht vorangestellte Zusammenfassung der Ergebnisse chronologisch geordnet. So wird deutlich, dass auch und gerade ein Projekt wie das hier beschriebene in soziale und mikropolitische Prozesse eingebunden ist. Solche Erfahrungen und Entwicklungen in das Projekt selber reflexiv einzubinden ist daher geboten. Nicht zuletzt scheint hier eine Thematik durch, die hier nicht bearbeitet werden konnte, die aber an Wichtigkeit nicht zu unterschätzen ist: Welche professionsbezogenen Veränderungen kommen auf die Lehrkräfte und Schulverwaltungen im Rahmen neuer Steuerungsverfahren zu?

Zwischenergebnisse des Projektes wurden im September 2004 und im Februar sowie März 2005 bei verschiedenen Veranstaltungen mit Lehrkräften und Schulleitern der SfE vorgestellt und diskutiert. Für die dort geäußerte Kritik, die Hinweise und Anregungen, welche soweit möglich in die weitere Auswertungsarbeit aufgenommen wurden, ist das Projektteam allen Beteiligten zum Dank verpflichtet.

Bochum, Juli 2005

2 Zusammenfassung der Ergebnisse

2.1 Projektbeginn (Frühjahr 2004)

2.1.1 Ausgangsfragestellungen:

Dem Projekt „Zentrale Vergleichsarbeiten an Schulen für Erwachsene“ lagen zwei Fragestellungen zugrunde:

1) Wie geeignet sind die bisher in den Schulen für Erwachsene eingesetzten zentralen Vergleichsarbeiten (d. h. die Arbeitsaufgaben und Bewertungskriterien) zur differenzierten und vergleichbaren Leistungsmessung?
2) Kann man die Arbeitsaufgaben der zentralen Vergleichsarbeiten nutzen, um anhand systematischer Aufgabendiagnostik einen Aufgabenpool zu erstellen?

Es geht also um eine Aufgabendiagnostik als Voraussetzung einer aussagekräftigen Studierendendiagnostik. Eine solche Aufgabendiagnostik ist unerlässlich, um Aussagen über die zentralen Vergleichsarbeiten hinsichtlich ihrer Vergleichbarkeit und ihres auf Studierendenleistungen bezogenen Differenzierungspotenzials treffen zu können. Als Vorgehensweise wurde daher eine statistische Analyse der Arbeitsaufgaben von zentralen Vergleichsarbeiten aller drei Hauptfächer ausgewählt:

- Englisch Vergleichsarbeit E2 (Dezember 2003; Abendgymnasien/ Hessenkollegs)
- Deutsch Vergleichsarbeit E2 (Dezember 2003; Abendgymnasien/ Hessenkollegs)
- Mathematik Vergleichsarbeit R4 (Dezember 2003; Abendrealschulen)

2.2 Analyse der Vergleichsarbeit im Fach Englisch

Im Anschluss an die Dateneingabe erfolgte eine deskriptive Auswertung einiger ausgewählter Grammatik- und Inhaltsaufgaben. Die deskriptive Analyse erbrachte erste Teilergebnisse, kam allerdings in ihrer Aussagekraft auch sehr schnell an ihre Grenzen. Eindeutige und brauchbare Aussagen über die „Schwierigkeit“ von Teilaufgaben sind z. B. aufgrund der Überschneidungen der Itemcharakteristiken nicht möglich. Die Anwendung eines wahrscheinlichkeitstheoretischen Verfahrens, das ebenfalls auf der genannten Idee der „Schwierigkeit“ von Aufgaben beruht, verspricht mehr Aussagekraft. Die nachträglich zu Projektschluss (Frühjahr 2005) durchgeführte wahrscheinlichkeitstheoretische Auswertung erbrachte dann etwas differenziertere Ergebnisse. Dabei zeigten die Grammatikaufgaben insbesondere das Problem, dass sie kaum eine Leistungsdifferenzierung zulassen, da ihr Schwierigkeitsgrad nur geringfügig differiert.

2.3 Analyse der Vergleichsarbeit im Fach Deutsch

Es erwies sich als unmöglich, die Ergebnisse der Aufgaben aus der Vergleichsarbeit im Fach Deutsch in eine statistisch auswertbare Form zu bringen. Die Bewertungsrichtlinien waren zwar sehr kleinschrittig, aber die Korrektur bzw. die Punktvergabe durch die Lehrkräfte ließ nur in Einzelfällen erkennen, welche Bewertungskriterien bei den jeweiligen Aufgaben zugrunde gelegt wurden. Im Regelfall ist nur eine einzige Note für mehrere/ alle Aufgaben vergeben worden, so dass eine differenzierte Aufgabenanalyse schon aus diesem Grund nicht möglich war. Die Analyse der Vergleichsarbeit Deutsch brachte somit zwar keinen Erkenntnisgewinn bezüg-

lich der Aufgabendiagnostik, aber sie war dennoch hilfreich für die viel grundlegendere Frage der professionsbezogenen Handlungsroutinen bei der Aufgabenkonstruktion und insbesondere bei den Bewertungsrichtlinien und der Bewertungspraxis. Letztere rückten in den Fokus des Projektes und führten folgerichtig zu einer Ergänzung der Projektfragestellung (s. u.).

2.4 Zwischenbilanz (Herbst 2004)

2.4.1 Bisherige Fragestellungen

1) Wie geeignet sind die bisher in den Schulen für Erwachsene eingesetzten zentralen Vergleichsarbeiten (d. h. die Arbeitsaufgaben und Bewertungskriterien) zur differenzierten und vergleichbaren Leistungsmessung?
2) Kann man die Arbeitsaufgaben der zentralen Vergleichsarbeiten nutzen, um anhand systematischer Aufgabendiagnostik einen Aufgabenpool zu erstellen?

2.4.2 Zwischenergebnisse aus der Analyse der Fächer Englisch/ Deutsch

Die Praxis der Erstellung von zentralen Vergleichsarbeiten sowie die Praxis der Bewertung lassen eine statistische Auswertung zwecks Aufgabendiagnostik nur eingeschränkt zu. Eine deskriptive Auswertung der Itemcharakteristiken ist nicht sehr aussagekräftig. Eine wahrscheinlichkeitstheoretische Auswertung (probabilistische Testtheorie) bietet sich als Alternative an. Eine Nutzung von solchen testtheoretischen Standards der Aufgabendiagnostik (z. B. Rasch-Modell) ist allerdings ebenso augrund der o. g. Aspekte nur eingeschränkt möglich.

Die Arbeitsgruppen zur Erstellung der Vergleichsarbeiten haben mit großem Aufwand Bewertungskriterien und Musterlösungen erarbeitet. Dabei haben sie auf die regulären und bewährten Verfahrensroutinen der Erstellung dezentraler Klassenarbeiten zurückgegriffen, d. h. eine individuelle, klassenbezogene Praxis wurde auf eine schulübergreifende Ebene übertragen. Der Anspruch der Vergleichbarkeit sollte durch eine Verringerung des Interpretationsspielraumes der einzelnen Lehrkraft geleistet werden, d. h. es wurden Musterlösungen und möglichst weitgehende Bewertungskriterien im Sinne eines maximal erreichbaren Konsenses formuliert. Dass die Aufgabenstellungen und die Operationalisierung der Bewertungskriterien teststatistischen Kriterien nur teilweise (i. S. v. zufällig) entsprechen, erklärt sich daraus, dass solche Kriterien nicht zu den professionsbezogenen Praxisroutinen gehören. Das heißt: Die Arbeitsgruppen haben versucht, eine möglichst weitgehende Vergleichbarkeit zu erzeugen. Da sie teststatistische Kriterien nicht kennen, können sie diese Kriterien folgerichtig auch nicht anwenden.

2.4.3 Schlussfolgerung

Bevor eine Aufgabendiagnostik erfolgen und der Aufbau eines Aufgabenpools erprobt werden kann, müssen teststatistische Kriterien in der *Erstellungs*praxis von zentralen Vergleichsarbeiten berücksichtigt werden. Es stellt sich also die Frage, ob entsprechende Verfahrensroutinen in die Professionspraxis eingebettet werden können. D. h.: Die *zweite* Fragestellung des Projektes muss um folgenden Aspekt ergänzt werden:

- Ist eine Implementierung von teststatistischen Anforderungen in die Professionspraxis der Erstellung von zentralen Vergleichsarbeiten möglich?

Die weitere Vorgehensweise des Projektes bestand daher aus drei Schritten:

- Die (notwendige) Umkodierung der Vergleichsarbeit Mathematik (R4), um den Erkenntniswert einer wahrscheinlichkeitstheoretischen Aufgabendiagnostik zu prüfen.
- Eine Analyse von Formaten und Musterlösungen der zentralen Vergleichsarbeiten um zu ermitteln, inwiefern eine teststatistische Auswertbarkeit der Arbeitsaufgaben prinzipiell möglich ist bzw. wie weitgehend diesbezügliche Änderungen gehen müssten.
- Eine Erstellung von Hinweisen zur Konstruktion von Aufgaben bzw. Bewertungskriterien für diejenigen Lehrkräfte, die mit der Erstellung der zentralen Vergleichsarbeiten betraut sind.

2.5 Analyse der Vergleichsarbeit im Fach Mathematik:

2.5.1 Vorgehensweise/ Ergebnisse

Im Sinne der ersten Fragestellung (s. u.) des Projektes wird eine Analyse der Arbeitsaufgaben und eine Umkodierung der Bepunktung der Vergleichsarbeit Mathematik (R4) vorgenommen, um den Erkenntniswert einer teststatistischen Auswertung nach dem Rasch-Modell zu prüfen. Eine analytische Zerlegung der Arbeitsaufgaben und eine Umkodierung der Bepunktung ist trotz sehr präziser Korrekturvorgaben notwendig, da verschiedene Voraussetzungen für eine Rasch-Auswertung bislang nicht erfüllt werden. So wurden unterschiedlich viele Punkte für zu vergleichende Aufgaben vergeben und es bestand eine interne Abhängigkeit verschiedener Teilaufgaben. Weiterhin wurden in einigen Aufgaben verschiedene Kompetenzen abgefragt, die für eine Aufgabendiagnostik differenziert werden müssen, da die zu analysierenden Aufgaben sich jeweils nur auf eine Kompetenz beziehen sollen.

Als Ergebnis lässt sich festhalten, dass es nur drei Leistungsstufen gibt, d. h. die Aufgaben weisen drei Schwierigkeitsgrade auf. Bezogen auf die sechsstufige Notenskala kann man von einer ggf. zu geringen Niveaudifferenzierung sprechen. Weiterhin zeigt sich, dass die leichtesten Aufgaben noch zu schwer für einige Studierende sind, d. h. das Leistungsspektrum der Studierenden wird nach unten hin nicht ausgelotet. Dies gilt auch in umgekehrter Richtung, denn die schwersten Aufgaben werden noch von zu vielen Studierenden gelöst, d. h. das Leistungsspektrum der Studierenden wird auch nach oben hin nicht vollständig erfasst.

2.6 Projektbilanz (Frühjahr 2005)

2.6.1 Ausgangsfragestellungen des Teilprojektes „Vergleichsarbeiten“

Wie bereits erwähnt, dienten ursprünglich zwei Fragestellungen als Ausgangspunkt für das Teilprojekt „Vergleichsarbeiten“, wobei die Zweite im Projektverlauf inhaltlich noch um einen weiteren Aspekt ergänzt worden ist:

1) Wie geeignet sind die bisher in den Schulen für Erwachsene eingesetzten Vergleichsarbeiten (Aufgaben) zur differenzierten und vergleichbaren Leistungsmessung?

2a) **Ursprünglich:** Kann man die Arbeitsaufgaben der Vergleichsarbeiten nutzen, um anhand systematischer Aufgabendiagnostik einen Aufgabenpool zu erstellen?

2b) **Ergänzend:** Ist eine Implementierung von teststatistischen Anforderungen in die Professionspraxis der Erstellung von zentralen Vergleichsarbeiten möglich?

2.6.2 Gesamtergebnisse der ersten Fragestellung:

1a) Für alle Fächer gilt generell: Die Arbeitsgruppen haben mit großem Arbeitsaufwand an den Vergleichsarbeiten gearbeitet, d. h. auch ausführliche Bewertungskriterien für die Arbeitsaufgaben erstellt (im Detail noch konkretisierbar).

1b) Aber: Die durchgängige bzw. nachvollziehbare *Anwendung* der vorgegebenen Bewertungskriterien differiert nach Fächern:

- Englisch: z. B. das Problem von Rechtschreibung bei Grammatikaufgaben
- Mathematik: z. B. Problem des Verhältnisses von Rechenweg und Rechenergebnis
- Deutsch: Es ist kaum identifizierbar, inwieweit Bewertungskriterien eingehalten werden.

Spezifizierte Aspekte des Unterrichtsfaches Deutsch:

- Die Konstruktion der Aufgaben erschwert eine differenzierte Zuordnung von Lösungen zu Kompetenzen. Diese ist allerdings keine grundsätzliche Frage, sondern eine Frage der Aufgabenstellung bzw. der Bewertungskriterien, z. B. der Operationalisierung von Lesekompetenz wie bei entsprechenden Aufgaben im Fach Englisch oder z. B. der präzise operationalisierten Bewertungskriterien für das Verhältnis von Form und Inhalt bei freien Texten.
- Im Fach Deutsch gibt es Traditionen der Aufgabenkonstruktion und der Bewertung, die sich nicht an einer kleinschrittigen Operationalisierung orientieren. Hinzu kommt, dass die Erstellung und Bewertung freier Textteile, die im Fach Deutsch dominieren, schwerer zu operationalisieren ist.
- Die fachspezifischen Prüfungsanforderungen (FAPAS) gehen auf die Problematik der Aufgabenkonstruktion nur unzureichend ein bzw. legen eine Interpretation nahe, die in *teststatistischer* Hinsicht zu einer problematischen Professionspraxis der Aufgabenerstellung führt.

1c) Eine probeweise Analyse der Vergleichsarbeit Mathematik der Abendrealschule nach dem Rasch-Verfahren zeigt:

- Das Leistungsspektrum der Studierenden wird weder nach oben noch nach unten hin ausgelotet.
- Eine noch stärkere Differenzierung der Arbeitsaufgaben nach Leistungsniveaus scheint möglich bzw. sinnvoll.

Quintessenz:

Die bisher in den Schulen für Erwachsene eingesetzten Vergleichsarbeiten (Aufgaben) ermöglichen *nur begrenzt* eine differenzierte und vergleichbare Leistungsmessung.

2.6.3 Gesamtergebnisse der zweiten Fragestellung:

2a) Professionspraxis und Diskussionsverlauf:

- Die Alltagspraxis der Klausurenerstellung und -bewertung entspricht nicht den formalen Anforderungen, die für eine systematische Aufgabendiagnostik notwendig wären (Aufgabenkonstruktion, Bewertungskriterien, Punktvergabe).
- Die Anwendung solcher formalen Voraussetzungen stößt auf deutliche Vorbehalte bei den Professionsangehörigen. Hier ist eine Trennung der Ebenen notwendig, d. h. es ging bei dem Projekt „Vergleichsarbeiten" nur um die Frage der Gestaltung zentraler Vergleichsarbeiten und nicht um die Frage einer generellen Anwendung teststatistischer Verfahren bei Klassenarbeiten.
- Die Anwendung von formalen Vorgaben der Erstellung von Aufgaben und Bewertungskriterien ist durch fachdidaktische Expertise leistbar, ohne diese einzuschränken.
- Zentrale Vergleichsarbeiten enthalten neben der Möglichkeit der vergleichenden Leistungsprüfung (Selektion der Studierenden) auch die Möglichkeit einer Förderung durch differenzierte Leistungsdiagnostik (gruppenbezogene Lernstandserhebung). Das Motiv einer Hilfestellung im Sinne einer differenzierteren Rückmeldung an die Lehrkraft wurde von vielen Lehrkräften bei den Diskussionen über Vergleichsarbeiten demjenigen der Selektion und Leistungskontrolle untergeordnet.

2b) Prinzipielle Anwendbarkeit einer Aufgabendiagnostik:

- Die Formate der Vergleichsarbeiten entsprechen den formalen Auswertungsanforderungen.
- Die Analyse von Musterlösungen der Vergleichsarbeiten Deutsch und Englisch (H2) zeigt, dass eine Umwandlung von bestehenden Entwürfen (Aufgabenkonstruktion, Bewertungskriterien, Punktvergabe) in auswertbare Formen *prinzipiell* möglich ist.

2c) Vorgehensweise Aufgabenpool:

- Die ursprüngliche Fragestellung zur Erstellung eines Aufgabenpools konnte nicht bearbeitet werden, da die untersuchten Vergleichsarbeiten nur unzureichend statistisch auswertbar waren. Die dazu notwendigen formalen Anforderungen waren nicht gegeben.
- Wahrscheinlichkeitstheoretische Testtheorie gehört nicht zum gängigen Professionswissen der Arbeitsgruppen, die die Vergleichsarbeiten erstellt haben.
- Die Erstellung eines Aufgabenpools über ein „induktives" Verfahren der Auswertung bestehender Vergleichsarbeiten, so wie sie z. Z. erstellt werden, würde einen umfangreichen Aufbau diesbezüglichen Professionswissens voraussetzen.

Quintessenz:

Die Arbeitsaufgaben der bisherigen Vergleichsarbeiten reichen nicht aus, einen Aufgabenpool zu erstellen.

Als effizienteste Lösung zur Erstellung eines Aufgabenpools bietet sich ein begleitetes Verfahren der Konstruktion und Eichung von Aufgaben an, d. h. eine Zusammenarbeit von Lehrkräften, Fachdidaktikern und Teststatistikern erscheint sinnvoll und notwendig.

3 Methodische Hinweise zur Aufgabendiagnostik

3.1 Grundannahmen der Leistungsmessung

Bei der Beurteilung der Leistungsfähigkeit von Studierenden wird angenommen, dass Studierende in unterschiedlichen Unterrichtsfächern unterschiedliche Leistungen aufweisen, z. B. kann ein Studierender in Mathematik gut, aber in Deutsch weniger gut sein. Die unterschiedlichen Fähigkeiten Studierender in unterschiedlichen Unterrichtsfächern können aus differierenden Eingangsvoraussetzungen, Interessen, Begabungen, Fähigkeiten oder anderen Rahmenbedingungen resultieren.

Ebenso wird angenommen, dass Studierende in Teilbereichen einzelner Unterrichtsfächer unterschiedliche Leistungen aufweisen können. Es ist z. B. vorstellbar, dass ein Studierender in Mathematik hervorragende algebraische Fähigkeiten aufweist, aber Schwierigkeiten bei geometrischen Aufgaben hat, oder dass ein Studierender im Unterrichtsfach Englisch die Grammatik beherrscht, aber große Vokabellücken hat.

Des Weiteren setzt die Leistungsbeurteilung von Studierenden auch Annahmen über den „Schwierigkeitsgrad“ von Aufgaben voraus, d. h. es existieren Aufgaben mit unterschiedlichen Schwierigkeitsgraden (leichte Aufgaben, schwere Aufgaben, ...). Auch hierzu ein veranschaulichendes Beispiel aus dem Bereich Mathematik: Das Auflösen einer Gleichung nach einer Unbekannten, in der nur die Operatoren Plus und Minus vorkommen, ist leichter als das Auflösen einer Gleichung nach einer Unbekannten, in der auch Potenzen und Logarithmen enthalten sind. Und auch im Unterrichtsfach Deutsch kann man annehmen, dass die Textanalyse einer Bildergeschichte von Plauen einfacher ist als die Interpretation des Eröffnungsmonologs in Faust I.

Aus diesen Annahmen über die Leistungsfähigkeit von Studierenden und dem unterschiedlichen Schwierigkeitsgrad von Aufgaben lassen sich die folgenden Schlussfolgerungen ziehen:

- Leistungsfähige Studierende können schwere und leichte Aufgaben lösen, denn das zeichnet gute Studierende aus. Weniger gute Studierende können dagegen nur leichte Aufgaben lösen.
- Werden in einer Klausur Aufgaben unterschiedlichen Schwierigkeitsgrades verwendet (leichte, mittlere und schwere Aufgaben), so kann die Leistungsfähigkeit der Studierenden anhand der Anzahl der gelösten Aufgaben ermittelt werden, denn gute Studierende werden eine höhere Anzahl von Aufgaben gelöst haben - da sie leichte, mittlere und schwere Aufgaben zu lösen vermögen - als weniger gute Studierende, die eventuell nur leichte Aufgaben lösen können.

3.2 Deskriptive Aufgabendiagnostik

Werden die grundlegenden Annahmen der Leistungsmessung von Studierenden auf eine deskriptive Aufgabendiagnostik einer Vergleichsarbeit übertragen, so kann man Folgendes sagen:

- Aufgaben unterschiedlicher Schwierigkeit werden von Studierenden je nach Ausprägung ihrer Fähigkeiten gelöst.
- Leichte Aufgaben werden von allen Studierenden gelöst, die über gewisse Basisfähigkeiten verfügen (oder sie erraten die richtige Lösung).

- Schwere Aufgaben werden nur von Studierenden gelöst, die über weit reichende Fähigkeiten verfügen.

Werden diese Folgerungen in Noten (KMK-Punkten) ausgedrückt, so ergeben sich folgende Aussagen: Studierende, die über eine gute Gesamtnote (hohe KMK-Punktezahl) in der Vergleichsarbeit verfügen, werden schwere Aufgaben größtenteils gelöst haben. Studierende mit einem unteren KMK-Punkteniveau werden nur zu einem sehr geringen Anteil solche schweren Aufgaben lösen etc.

Somit erhält man eine Aussage über den Schwierigkeitsgrad einer Aufgabe, wenn man den prozentualen Anteil der Studierenden (bezogen auf die Gruppe derjenigen, die jeweils die gleiche KMK-Punktezahl als Gesamtnote erhalten haben) errechnet, welche die Aufgabe gelöst haben. In der folgenden idealtypischen Abbildung werden zur Verdeutlichung dieser Überlegung die prozentualen Anteile von Studierenden mit gleicher KMK-Punktezahl dargestellt, die eine Aufgabe gelöst haben und zwar für unterschiedlich schwere Aufgaben.

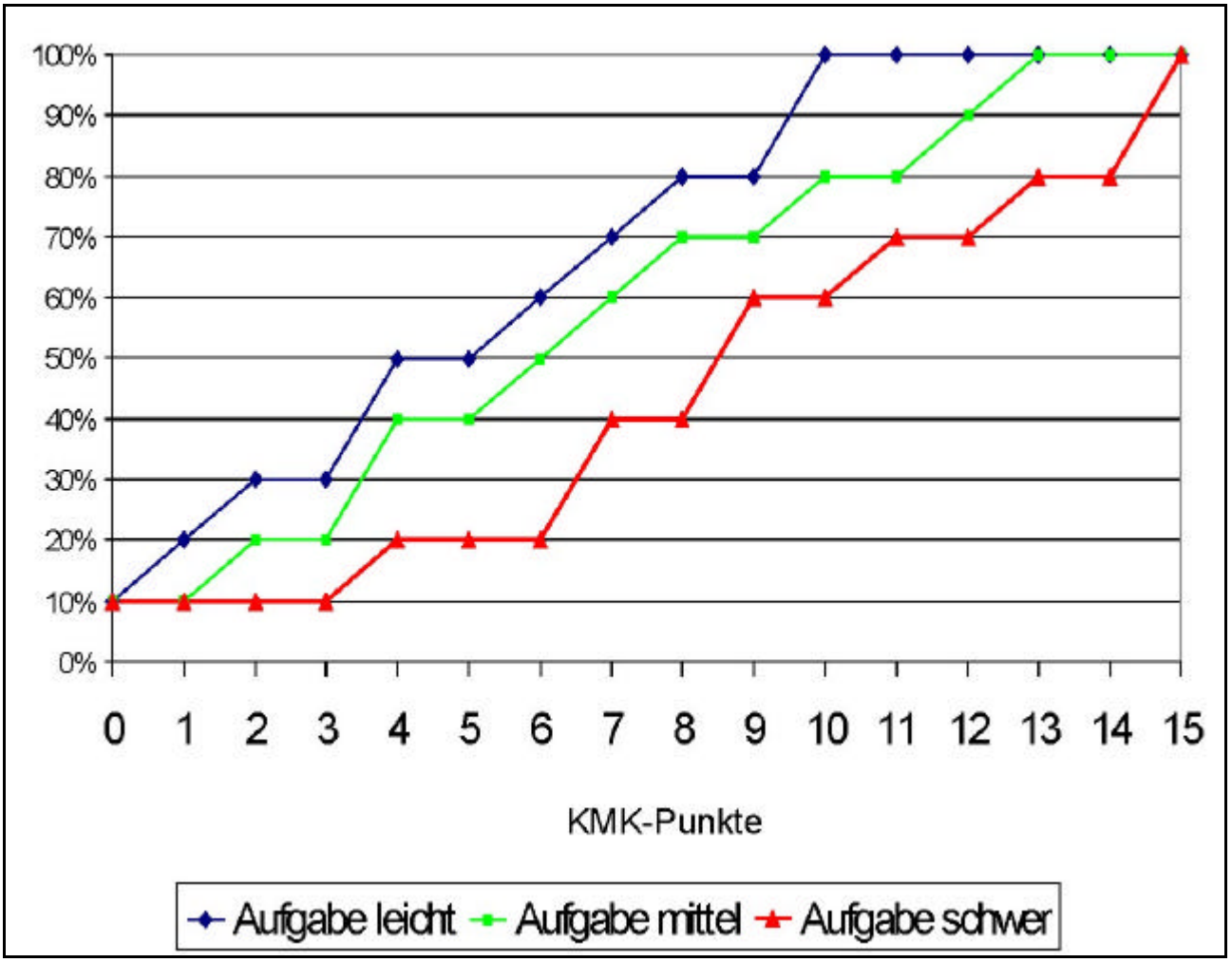


Abbildung 1: Fiktives Beispiel eines deskriptiven Verlaufs der Aufgabenschwierigkeiten von drei Aufgaben unterschiedlicher Schwierigkeit

Betrachtet man hier beispielsweise die Gruppe der Studierenden, die 6 KMK-Punkte erhalten haben, so haben von diesen Studierenden 20% die schwere Aufgabe gelöst, 50% die mittelschwere Aufgabe und 60% die leichte Aufgabe. Diese Unterschiedlichkeit ist für alle KMK-Punkte gegeben, d. h. es gibt keine Überschneidungen der Aufgabenschwierigkeiten.

Durch eine Überschneidung der Aufgabenschwierigkeiten würden sich die „Schwierigkeitsverhältnisse“ umkehren: eine Aufgabe, die für Studierende des unteren KMK-Punktespektrums leicht war, würde plötzlich für die Studierenden des oberen KMK-Punktespektrums schwer werden - und umgekehrt. Eine eindeutige Zuordnung von Aufgaben zu unterschiedlichen Schwierigkeiten für alle Studierenden wäre so nicht möglich.

Daher ist die obige Abbildung als Beispiel zu verstehen. Die Kurven überschneiden sich nicht, und eine eindeutige Zuordnung der Aufgaben zu unterschiedlichen Schwierigkeitsniveaus ist möglich, d. h. die leichte Aufgabe wird zu 100% schon von Studierenden gelöst, die insgesamt nur 10 KMK-Punkte erreicht haben. Die schwerste Aufgabe wird dagegen erst zu 100% von der Gruppe von Studierenden mit 15 KMK-Punkten gelöst. Analog im unteren Bereich der KMK-Punkteskala: Studierende mit einem einzigen KMK-Punkt lösen schon zu 20% die leichte Aufgabe, aber nur zu 10% die schwere oder mittelschwere Aufgabe. Man nennt diese Kurvenverläufe auch „Aufgabencharakteristik“ einer Aufgabe.

Anhand der Aufgabencharakteristik können außerdem zu leichte oder zu schwere (oder ggf. auch einfach nur missverständlich formulierte) Aufgaben identifiziert werden. Die folgende Abbildung zeigt die Verläufe der Aufgabencharakteristik für diese beiden Fälle, wiederum idealtypisch.

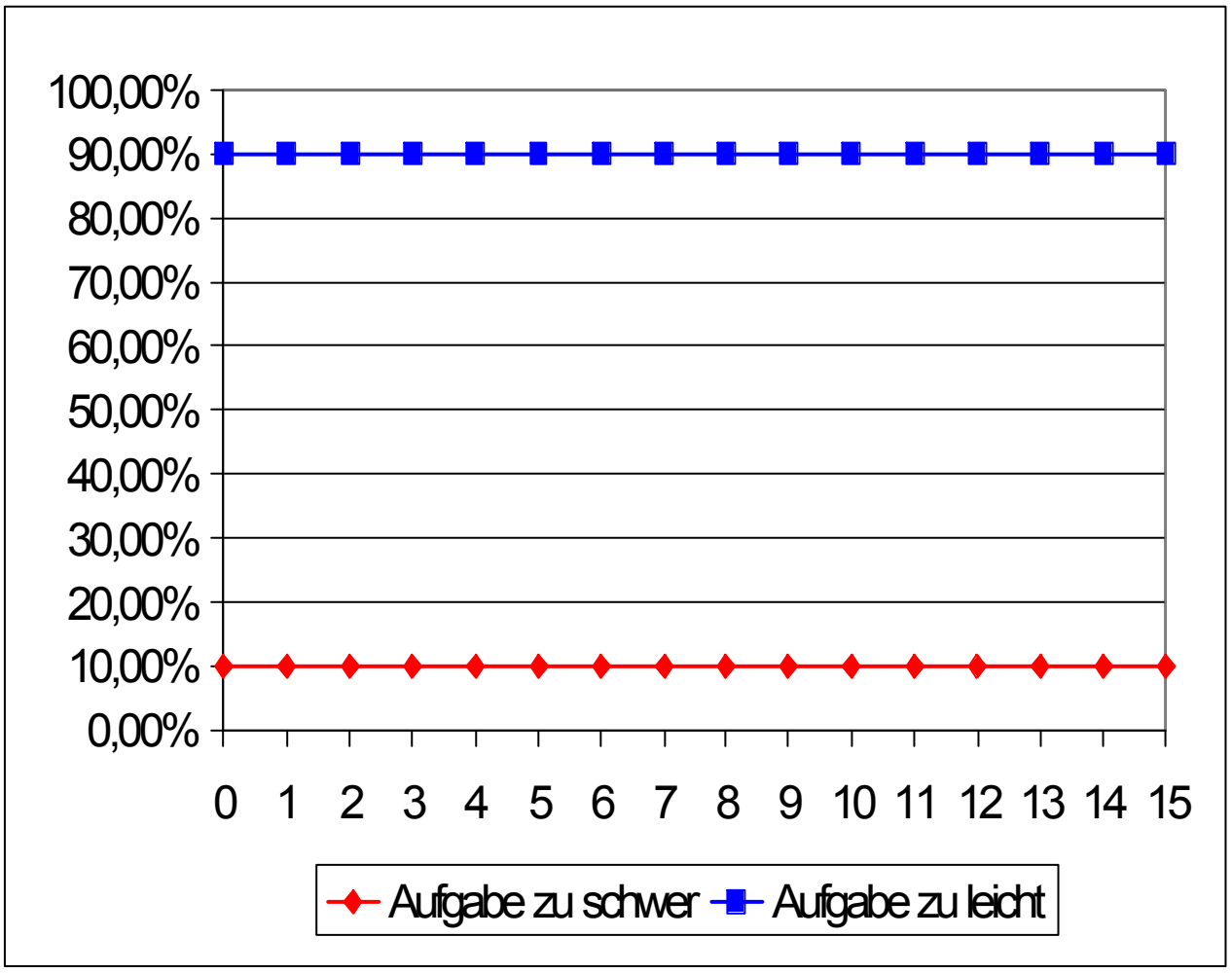


Abbildung 2: Beispiel eines Aufgabenschwierigkeitsverlaufs für eine zu leichte/ zu schwere Aufgabe

Man erkennt, dass eine der Aufgaben zu leicht für die Gruppe der Studierenden war, denn fast alle Studierenden (90%) mit jeweils gleicher KMK-Punkteanzahl konnten diese Aufgabe lösen

und zwar über das ganze Spektrum der möglichen KMK-Punkte. Mit Hilfe dieser Aufgabe können also keine Aussagen über Leistungsunterschiede zwischen den Studierenden getroffen werden.

Die andere Aufgabe ist zu schwierig, denn hier hat fast keiner der Studierenden (10%) aus der Gruppe mit gleicher KMK-Punktezahl die Aufgabe lösen können, und wiederum ist dies unabhängig davon, wie viele KMK-Punkte die Studierenden haben. Allerdings besteht hier auch die Möglichkeit, dass diese Aufgabe nicht zu schwierig war, sondern, dass z. B. niemand die Aufgabenstellung verstanden hat, da die Aufgabe vielleicht missverständlich formuliert wurde.

3.3 Probabilistische Aufgabendiagnostik

Diese o. g. Überlegungen zu Fähigkeiten und Aufgabenschwierigkeiten sind jedoch nicht deterministisch zu verstehen, sondern sind mit Unsicherheiten behaftet. Es ist durchaus möglich, dass einige schwere Aufgaben durch leistungsschwache Studierende erraten werden bzw. auch einige gute Studierende an schwierigen und leichten Aufgaben scheitern. Diese Unsicherheit kann durch die Wahrscheinlichkeit der Lösung einer Aufgabe beschrieben werden. Als Frage könnte man diesen Sachverhalt folgendermaßen formulieren: Wie wahrscheinlich ist es, dass ein leistungsfähiger Studierender eine leichte Aufgabe löst bzw. ein weniger leistungsfähiger Studierender eine schwere Aufgabe etc.?

Diese Unsicherheiten bearbeitet die probabilistische (wahrscheinlichkeitstheoretische) Testtheorie. Im Folgenden sollen deren Grundannahmen am Beispiel des so genannten Rasch-Modells kurz erläutert werden.

Die dem einfachen Rasch-Modell zugrundeliegende Aufgabencharakteristik („Itemcharakteristik“) hat die folgende grafische Verlaufsform:

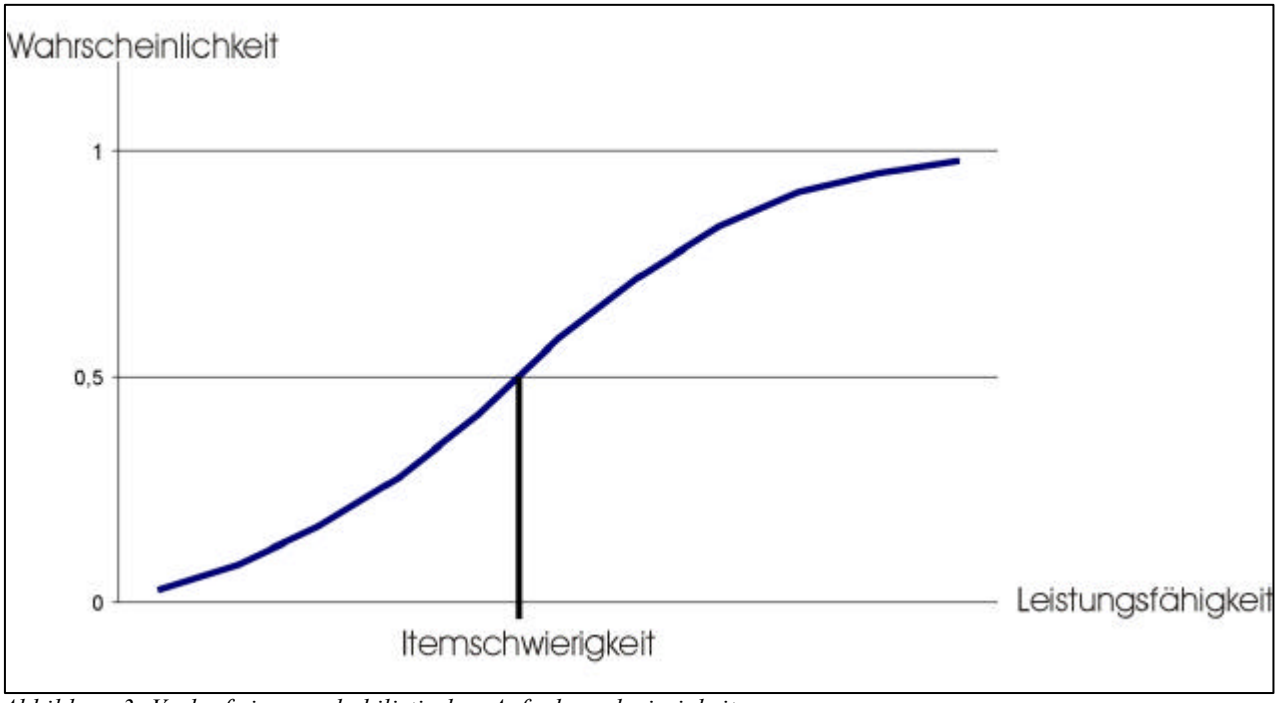


Abbildung 3: Verlauf einer probabilistischen Aufgabenschwierigkeit

Die Kurve stellt die Wahrscheinlichkeit der Lösung einer Aufgabe in Abhängigkeit von der Leistungsfähigkeit der Studierenden dar. Die Koordinate auf der Achse der Leistungsfähigkeit,

an der die Wahrscheinlichkeit der Lösung der Aufgabe 50% beträgt, wird als „Itemschwierigkeit“, „Aufgabenschwierigkeit“ oder „Itemlocation“ bezeichnet. Ein Studierender, der exakt diese Leistungsfähigkeit besitzt, löst die durch diese Aufgabencharakteristik beschriebene Aufgabe mit einer Wahrscheinlichkeit von 50%. Studierende mit einer höheren Leistungsfähigkeit lösen die Aufgabe entsprechend dem Verlauf der Aufgabencharakteristik mit höherer Wahrscheinlichkeit, Studierende mit geringerer Leistungsfähigkeit als der Itemschwierigkeit mit entsprechend geringerer Wahrscheinlichkeit.

Das Rasch-Modell ist ein mathematisches Modell, das diese Aufgabenschwierigkeiten unter der Annahme bestimmt, dass sich die angetroffenen richtigen Lösungen entsprechend der Zuordnung der Aufgaben zu Schwierigkeitsgraden normalverteilen, d. h. dass z. B. eine Aufgabe mittleren Schwierigkeitsgrades von den meisten derjenigen Studierenden gelöst wird, die über eine mittlere Leistungsfähigkeit zur Lösung dieser Aufgabe verfügen.

Dabei werden unterschiedliche Rasch-Modelle unterschieden. Das einfache, dichotome Rasch-Modell kann nur Daten/ Variablen verarbeiten, die in der Form „richtig oder falsch“ bzw. mit der Eigenschaft „vorhanden oder nicht vorhanden“ erfasst worden sind. Mit dem mehrkategoriellen Rasch-Modell können, wie der Name bereits ankündigt, mehrkategorielle Variablen verarbeitet werden, wobei die Variablen allerdings nicht mehr als fünf Kategorien aufweisen sollten, z. B. in der Form einer mehrstufigen Bewertungsskala von „Anforderung gar nicht erfüllt“ bis „Anforderung voll erfüllt“. Diese Bedingungen an die Korrekturschemata von Aufgaben müssen erfüllt sein, um eines der genannten Rasch-Modelle anwenden zu können.

Eine weitere Voraussetzung zur Anwendung des Rasch-Modells ist, dass die verwendeten Aufgaben jeweils für sich nur eine Fähigkeit überprüfen dürfen. Anhand der Aufgabe Nr. 4 des Grammatikteils der Vergleichsarbeit Englisch (s. u.) soll die Berücksichtigung dieser Voraussetzungen veranschaulicht werden.

Zunächst scheint die Korrektur dieser Aufgabe die Anforderung an die Korrektur für das einfache Rasch-Modell zu erfüllen, d. h. es gibt einen (einzigen) Punkt für die Verwendung der richtigen Form und keinen Punkt für die Verwendung der falschen Form. Somit könnte man das einfache Rasch-Modell anwenden. Doch gleichzeitig gab es auch Punktabzüge für Fehler bei der Rechtschreibung, daher müsste das mehrkategorielle Rasch-Modell angewendet werden. Diese Punktabzüge wurden jedoch nicht einheitlich von allen Lehrkräften durchgeführt. Somit lässt sich die nur teilweise mehrkategoriell ausgewertete Unteraufgabe nicht mehr durch das mehrkategorielle Rasch-Modell analysieren. Es bleibt nur die Möglichkeit, eine Auswertung mit dem einfachen Rasch-Modell durchzuführen, das nur die richtige Verwendung der Grammatik untersucht.

Durch die zusätzliche Bewertung der Rechtschreibung wird auch die Voraussetzung verletzt, dass eine Aufgabe jeweils nur eine Fähigkeit überprüfen darf. Es handelt sich also weniger um ein Problem, welches auf das Rasch-Modell angewendet werden kann, sondern vielmehr um das grundsätzliche Problem, in welcher Weise die Bewertung der Rechtschreibung die Bewertung der Grammatikfähigkeit beeinflusst.

Die Vergabepraxis der Punkte müsste also dahingehend geändert werden, dass die korrekt verwendete grammatikalische Form mit einem Punkt bewertet wird - unabhängig von der verwendeten Rechtschreibung. Das Modell zwingt somit zu einer Präzisierung und Objektivierung der Bewertungspraxis, indem entweder möglichst kleinschrittige Bewertungen eingeführt werden, die keinen oder nur geringen Spielraum der Bewertung zulassen, oder alternativ weitgefasste Bewertungspraxen begründet werden. Das Bewertungsproblem des Beispiels könnte z. B. gelöst werden, indem entweder die Rechtschreibung nicht bewertet wird, oder aber die Bewertung der Rechtschreibung unabhängig davon erfolgt, ob die grammatikalische Anforderung erfüllt ist.

3.4 Kriterien für Aufgaben und Bewertung

Um eine statistische Auswertung von Aufgaben zu gewährleisten, müssen Aufgaben und Bewertungskriterien bestimmte Kriterien erfüllen. Diese Kriterien werden hier kurz aufgelistet (für eine ausführlichere Darstellung siehe Anhang):

- Die Anwendung des Rasch-Modells, ob einfach oder mehrkategoriell, setzt voraus, dass die verwendeten Aufgaben die bei den Studierenden zu messende Fähigkeit tatsächlich erfassen, z. B. die Verwendung von „if"-Sätzen, das Erstellen einer Einleitung oder das Lösen von geometrischen Aufgaben.
- Die Bewertung der Aufgaben muss unabhängig vom Bewerter (Signierobjektivität) und vom Bewerteten sein, d.h. die Bewertung selbst ist einheitlich und es gibt keinen oder nur einen möglichst geringen Interpretationsspielraum bei der Bewertung. Aus dieser Forderung leitet sich die Kleinschrittigkeit der Korrektur (nicht der Aufgabenstellung) ab. (Diese Forderung ist nicht nur an die Auswertung von Vergleichsarbeiten mittels des Rasch-Modells geknüpft. Jede Bewertung von Leistung sollte im höchsten Maß objektiv, d. h. unabhängig von Bewerter oder Bewertetem und nachvollziehbar im Sinne einer Kleinschrittigkeit sein.)
- Die für die Überprüfung einer Fähigkeit herangezogenen Aufgaben müssen alle die gleiche Bewertungsskala aufweisen, z. B. „richtig oder falsch" oder "richtig, fast richtig, halb richtig, falsch" etc.
- Eine weitere Forderung an die Aufgaben ist, dass sie untereinander unabhängig voneinander sind, d. h. ob eine Aufgabe gelöst wird, hängt nur von der Person und der Aufgabe selbst ab, nicht davon, welche Aufgaben bisher gelöst wurden (logische Unabhängigkeit). Andernfalls können sich logisch Widersprüche durch richtig geratene Aufgaben ergeben. (Zum Beispiel: Eine leichte Aufgabe wurde nicht gelöst, aber die davon abhängige mittlere Aufgabe wurde richtig erraten.)
- Sollen aufgabenübergreifende Anforderungen untereinander verglichen werden, so müssen die für die Korrektur der Aufgaben geltenden Bedingungen auch für die zu vergleichenden Anforderungen erfüllt sein, z. B. einheitliche Benotung von aufgabenübergreifenden Anforderungen der Vergleichsarbeiten wie Grammatik, Inhalt, Ausdruck oder Fehlerindex.
- Eine Gewichtung von Aufgaben oder Anforderungen ist möglich, darf aber erst in einem letzten Schritt z. B. der Endnotenbestimmung erfolgen.

4 Analyse der Vergleichsarbeit Englisch (E2)

Die Auswertung der Vergleichsarbeit in Englisch (E2) gliedert sich in folgende vier Schritte:

1. Für die Vergleichsarbeit Englisch E2 wurden durch die Arbeitsgruppe vier Teilnoten definiert: Ausdruck, Inhalt, Fehlerindex und Grammatikteil. Daher erfolgte zunächst anhand einer Faktorenanalyse eine Überprüfung der Klausurteile dahingehend, ob es eine den Teilnoten der Vergleichsarbeit zugrunde liegende gemeinsame latente Variable gibt. Diese Auswertung begibt sich noch nicht auf die Ebene der einzelnen Aufgaben und ist als eine grundlegende „Übersichtsanalyse“ gedacht, bevor einzelne Aufgaben analysiert werden.
2. In einem zweiten Schritt erfolgt eine Vorstellung der deskriptiven Überprüfung von ausgewählten Aufgaben anhand von deskriptiven Aufgabencharakteristiken.
3. Im Weiteren werden einige Erfahrungen aufgeführt, die bei der Dateneingabe gemacht wurden.
4. Diese drei Auswertungsschritte wurden bereits im September 2004 bei einem Workshop für Vertreter der Englisch-Fachkonferenzen präsentiert. Als ein wesentliches Ergebnis stellte sich heraus, dass eine deskriptive Auswertung für sich allein leider nicht zu eindeutigen und brauchbaren Aussagen führt. Nachdem auch die Vergleichsarbeiten in Deutsch und Mathematik einer Analyse unterzogen worden waren, wurden die Arbeitsaufgaben der Vergleichsarbeit Englisch noch einer weiteren Analyse nach dem Rasch-Modell unterzogen.

4.1 Faktorenanalyse der vier Teilnoten (Inhalt, Ausdruck, Fehlerindex, Grammatik)

Die Überprüfung, ob den Teilnoten der Vergleichsarbeit Englisch E2 eine gemeinsame latente Variable zugrunde liegt, erfolgte mit Hilfe einer Faktorenanalyse. Das Verfahren der Faktorenanalyse konstruiert aus gegebenen empirischen Daten so genannte Faktoren, d. h. gemeinsame latente Variablen/ Eigenschaften, auf welche die gemessenen empirischen Daten inhaltlich zurückgeführt werden können. Gleichzeitig liefert die Faktorenanalyse ein Gütekriterium dafür, inwieweit diese konstruierten Variablen tatsächlich die empirischen Variablen erklären, d. h. inhaltlich dazugehören.

In die Berechnung zur Konstruktion der latenten Eigenschaften wurden die Note für den Ausdruck, für den Inhalt, die Note für den Fehlerindex und die Note des Grammatikteils einbezogen. Dabei geben die in der Abbildung ausgewiesenen Werte den Zusammenhang der Teilnote zu der konstruierten latenten Eigenschaft an. Diese Werte können nur in einem Intervall von -1 bis +1 liegen, wobei die Werte -1 und +1 einen eindeutigen Zusammenhang zwischen der empirischen und der latenten Variable kennzeichnen.

Als Ergebnis ist festzuhalten: Das Verfahren der Faktorenanalyse liefert nur eine einzige latente Variable (Faktor), welche die Note für den Ausdruck, für den Inhalt, die Note für den Fehlerindex und die Note des Grammatikteils erklärt. Da die errechneten Werte für den Zusammenhang von Teilnoten und latenter Variable nahe dem Maximum von Eins liegen, hat die latente Variable einen hohen Anteil am Zustandekommen der Einzelnoten. Es gibt also eine latente Variable, die das Zustandekommen der Teilnoten erklärt. Diese eine latente Variable/ Faktor wurde von uns als „Englischkompetenz“ bezeichnet. Die Teilbereiche der Vergleichsarbeit Englisch E2 wurden demnach von der Arbeitsgruppe so konstruiert, dass die zu überprüfende Leistungsfähigkeit in Englisch mittels der Teilnoten sehr gut dargestellt werden kann.

Komponentenmatrix[a]

	Komponente
	1
T1_IN_NO	,867
T1_F_NOT	,896
T1_AUSDR	,945
T2_NOTE	,814

Extraktionsmethode: Hauptkomponentenanalyse.

a. 1 Komponenten extrahiert

Abbildung 4: Ergebnis der Faktorenanalyse über die Teilnoten Inhalt, Fehlerindex, Ausdruck und Grammatik (Vergleichsarbeit - Englisch)

Allerdings zeigt die Faktorenanalyse auch, dass die Leistung der Studierenden im Grammatikteil (T2_NOTE) durch die latente Variable deutlich weniger erklärt werden kann, als die anderen Teilbereiche. (Bei einer hundertprozentigen Erklärung des Teilbereichs durch die latente Variable würde sich der Extraktionswert „1“ ergeben.) Dies wird auch deutlich, wenn die Kommunalitäten betrachtet werden, d. h. der Anteil der Varianz der latenten Variablen (Faktor), der durch die einzelnen empirischen Variablen, in diesem Fall die Teilnoten, aufgeklärt wird.

Kommunalitäten

	Anfänglich	Extraktion
T1_IN_NO	1,000	,751
T1_F_NOT	1,000	,802
T1_AUSDR	1,000	,893
T2_NOTE	1,000	,663

Extraktionsmethode: Hauptkomponentenanalyse.

Abbildung 5: Anteil der durch den Faktor aufgeklärten Varianz (Vergleichsarbeit - Englisch)

Der Anteil gemeinsamer Varianz (quadrierte Streuung) des Faktors „Englischkompetenz“ beträgt bzgl. der Teilnote „Grammatik“ nur 66,3%, wohingegen z. B. der Anteil gemeinsamer Varianz der Teilnote „Ausdruck“ mit dem Faktor „Englischkompetenz“ bei 90% liegt.

Es lag daher nahe, den Zusammenhang der Teilnote „Grammatik“ in Abhängigkeit der anderen drei Teilnoten gesondert zu untersuchen. Diese Überprüfung erfolgte mit Hilfe einer multiplen Regression: Die Überprüfung des Zusammenhangs zwischen den Teilnoten Inhalt, Ausdruck, Fehlerindex als unabhängige Variablen und der Grammatiknote als abhängiger Variable mit Hilfe einer multiplen Regression stärkt das Ergebnis der Faktorenanalyse. Ein Anteil von nur

48,2% der Grammatiknote wird durch die drei anderen Teilnoten bestimmt. Der Grammatikteil nimmt also eine Sonderstellung innerhalb der Teilnoten ein.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,694[a]	,482	,477	2,927

a. Einflußvariablen : (Konstante), T1_AUSDR, T1_IN_NO, T1 F NOT

Abbildung 6: Einfluss der Teilnoten Ausdruck, Inhalt und Fehlerindex auf die Teilnote Grammatik (Vergleichsarbeit - Englisch)

4.2 Deskriptive Analyse

4.2.1 Deskriptive Analyse Aufgabe 1 (Teil I – Textarbeit)

I. Working with the text:

1. Tell whether these statements are right or wrong.

	right	wrong
1. Lily grew up in the United States.		
2. Lily thinks that American children are sometimes disrespectful to their parents.		
3. Lily's friends wanted her to move into an apartment with them.		
4. Lily's parents encouraged her to move out.		
5. Lily was torn between independence and obedience.		
6. Lily decided to move out.		
7. Lily now regrets the decision she made.		
8. Lily lost some friends as a result of her decision.		

to encourage – ermutigen

to be torn between – hin- und hergerissen sein zwischen

obedience – Gehorsam

to regret – bedauern **8p**

Abbildung 7: Aufgabe 1, Teil I - Textarbeit (Vergleichsarbeit - Englisch)

Mit dieser Aufgabe wird die Lesekompetenz der Studierenden überprüft, indem sie die Richtigkeit einer Aussage erkennen sollen, die sich auf einen Text der Klausur bezieht. Für das Erkennen der Richtigkeit oder Unrichtigkeit der Aussage gab es jeweils einen Punkt. Insgesamt waren somit 8 Punkte erzielbar. Die folgende Grafik stellt den prozentualen Anteil der Studierenden dar, welche die Aufgabe vollständig gelöst haben und zwar gruppiert nach der jeweiligen KMK-Punktezahl, welche die Studierenden als Gesamtnote der Vergleichsarbeit erzielten.

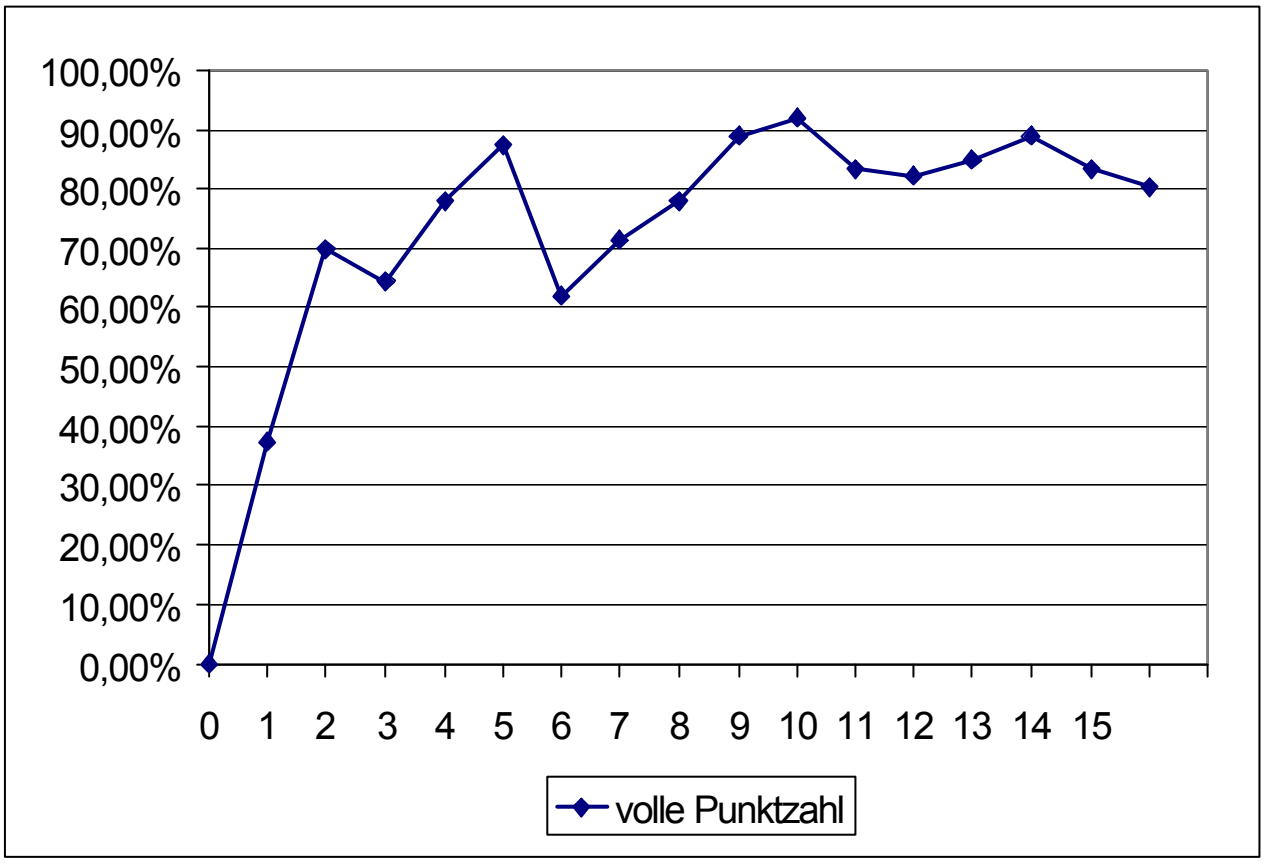


Abbildung 8: Prozentualer Anteil vollständiger Lösungen von Aufgabe 1 des Teil I- Textarbeit für die Studierenden mit gleicher KMK-Punktzahl (Vergleichsarbeit - Englisch)

Die Aufgabencharakteristik der Aufgabe 1 aus Teil I der Vergleichsarbeit zeigt, dass schon 70% der Studierenden in der Gruppe, die insgesamt 2 KMK-Punkte erzielt, diese Aufgabe vollständig lösen. Alle Studierende mit höherer KMK-Punktezahl in der Gesamtnote weisen ähnlich hohe Werte auf. Diese Aufgabe ist also nach den oben angestellten Überlegungen als „zu leicht“ einzustufen. Sie trägt nicht zur Leistungsdifferenzierung der Studierenden bei. Mit Hilfe dieser Aufgabe sind keine Rückschlüsse auf individuelle Defizite der Studierenden möglich.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,219[a]	,048	,046	3,351

a. Einflußvariablen : (Konstante), T1_1_SUM

Abbildung 9: Einfluss der Aufgabe 1 von Teil I - Textarbeit auf die Gesamtnote (Vergleichsarbeit - Englisch)

Dass diese Aufgabe keinen Beitrag zur Leistungsdifferenzierung leistet, wird durch die Ergebnisse der linearen Regression gestützt, die für den Zusammenhang von Gesamtnote/ KMK-Punkte und Lösung der Aufgabe 1 Teil 1 durchgeführt wurde. Da der Wert für R-Quadrat fast Null beträgt, trägt die Aufgabe 1 Teil 1 nichts zur Aufklärung des Zustandekommens der Gesamtnote bei. Im Rahmen des Workshops für die Englisch-Lehrkräfte im September 2004 wurde in der Diskussion dieses Ergebnisses darauf hingewiesen, dass es sich um eine typische „Warm-up-Aufgabe" handele, deren Sinn es sei, im Sinne einer Motivationssteigerung von möglichst vielen Studierenden gelöst zu werden.

4.2.2 Deskriptive Analyse Aufgabe 4 (Teil II - Grammatik)

4. Adjective or Adverb – choose the right form:

Lily comes form an _____ (extreme) conservative family.

She was taught to be very _____ (respectful) to her parents.

In American culture, most children are encouraged to live an _____ (independent) life.

Family matters are discussed _____ (open) in front of the children.

4p

Abbildung 10: Aufgabe 4, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Diese Aufgabe dient zur Überprüfung der Fähigkeit der richtigen Verwendung von Adjektiven oder Adverbien. Für jede richtige Verwendung von Adjektiv oder Adverb sollte ein Punkt vergeben werden. Da die Lehrkräfte auch die korrekte Rechtschreibung beachteten, wurden auch halbe Punkte vergeben. Die folgende Abbildung stellt daher den prozentualen Anteil derjenigen Studierenden, dar, die für den jeweiligen Lückentext die volle Punktzahl erhielten. Die Studierenden wurden dafür wieder in Gruppen gemäß ihrer Gesamtnote aufgeteilt. (Kodierung: t24a = extreme, t24b = respectful, t24c = independent, t24d = open).

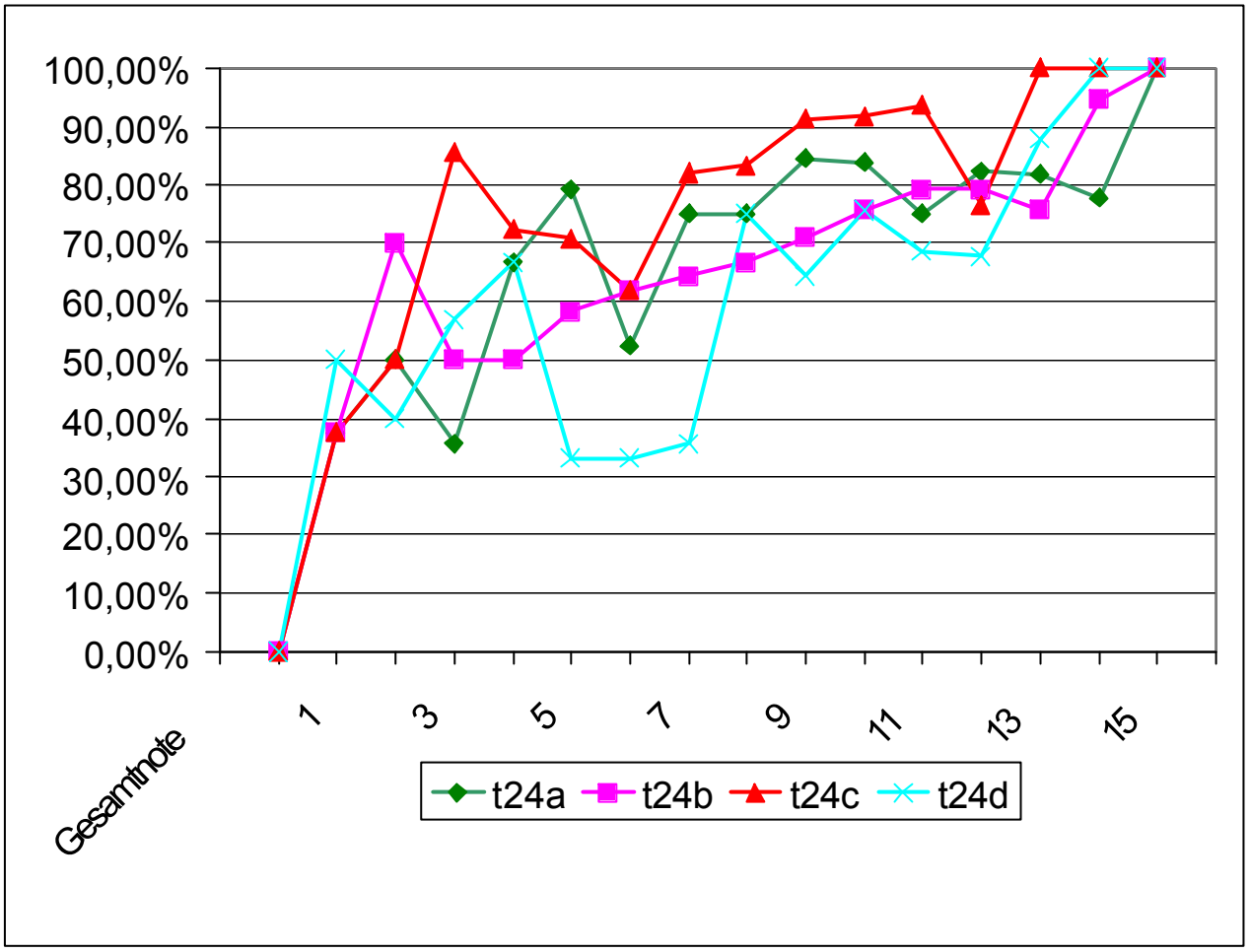


Abbildung 11: Aufgabencharakteristiken für Aufgabe 4, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Ersichtlich ist, dass sich die Teilaufgaben aufgrund der Überschneidungen der Aufgabencharakteristik keinen eindeutigen Schwierigkeitsniveaus zuordnen lassen. Bei Betrachtung des Randbereichs von 13-15 KMK-Punkten könnte Teilaufgabe a) als die Schwierigste und Teilaufgabe c) als die Leichteste interpretiert werden.

Betrachtet man die prozentuale Lösungshäufigkeit der Teilaufgaben aller Studierenden, unabhängig davon welche KMK-Punkte sie jeweils erzielten, so ergibt sich ein anderes Bild. Die prozentuale Gesamtverteilung für die Teilaufgaben ergibt:

- Aufgabenteil a) wird von 74,28 % aller Studierenden gelöst
- Aufgabenteil b) wird von 70,34 % aller Studierenden gelöst
- Aufgabenteil c) wird von 70,34 % aller Studierenden gelöst
- Aufgabenteil d) wird von 64,57 % aller Studierenden gelöst

Danach wäre Aufgabenteil d) am schwierigsten und Aufgabenteil a) am leichtesten. Die nach Gesamtnote sortierte Betrachtung der Aufgabencharakteristik liefert also eine andere Information als die Gesamtbetrachtung der Studierenden. Hier zeigen sich bereits die Grenzen einer deskriptiven Aufgabecharakteristik.

4.2.3 Deskriptive Analyse Aufgabe 3 (Teil II - Grammatik)

3. What do the friends say? What does Lily say? /(if-clauses Types I + II)

1. If your parents were Americans, they _____ (let) you live independently.
2. If you really like me, you _____ (accept) that I live with my parents.
3. You will come and live with us if you _____ (be) a true friend.
4. If my parents _____ (be) Americans, they would be less conservative.
5. You will be one of us if you _____ (not listen) to your parents.
6. If I was raised in the US, I _____ (not respect) my parents.

6p

Abbildung 12: Aufgabe 3, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Diese Aufgabe dient zur Überprüfung der Fähigkeit der richtigen Verwendung von „if"-Sätzen. Auch hier gab es für das richtige Ausfüllen des Lückentextes einen Punkt und auch hier gab es für Fehler bei der Rechtschreibung Abzüge von einem halben Punkt. Die Aufgabencharakteristiken sind daher analog zu Aufgabe 4 (s. o.) erstellt worden.

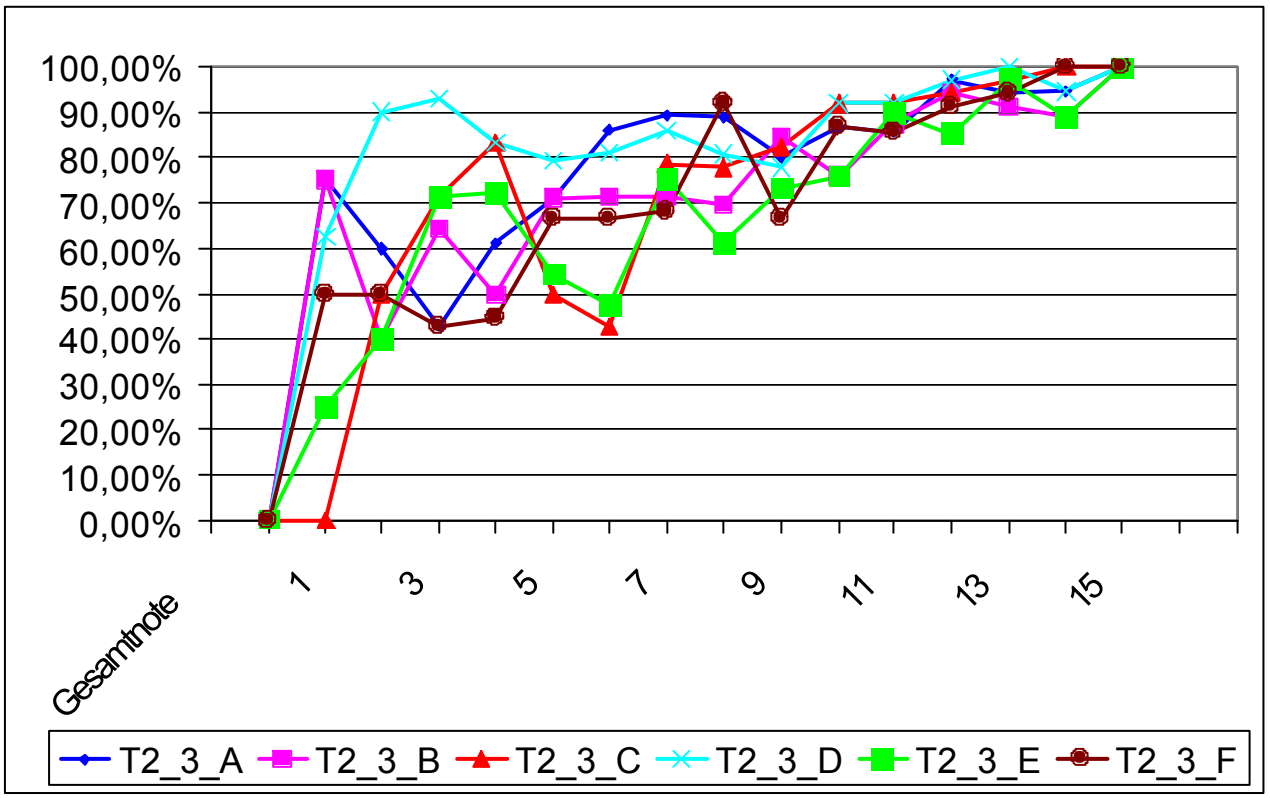


Abbildung 13: Aufgabencharakteristiken für Aufgabe 3, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Auch hier lässt sich aufgrund der häufigen Überschneidungen der Aufgabencharakteristiken keine eindeutige Zuordnung zu Schwierigkeitsniveaus vornehmen. Auffällig ist allein Teilaufgabe d). Sie wird von Studierenden in dem KMK-Punkteintervall von drei bis sechs Punkten

überdurchschnittlich häufig gelöst, was den Schluss nahe legen könnte, dass es sich um eine leichte Teilaufgabe handelt.

4.2.4 Deskriptive Analyse Aufgabe 2 (Teil II - Grammatik)

2. Transform the following sentences into the passive:
(example: We respected our elders ⇨ Our elders were respected)

a. Sometimes they treat their parents with disrespect.

b. They asked me to move in with them.

c. Most of my friends have accepted my decision.

d. I will find a solution.

e. American parents considered it acceptable.

f. My friends could not understand this.

12p

Abbildung 14: Aufgabe 2, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Diese Teilaufgabe überprüft die Fähigkeit der richtigen Anwendung des Passivs im Englischen. Für die korrekte Transformation der Sätze ins Passiv gab es jeweils 2 Punkte. Da der Spielraum für die Vergabe von einem halben bis zu anderthalb Punkten nicht nachvollziehbar war, erfolgte die Auswertung der Aufgabencharakteristik mittels der prozentualen „Nichtlösung" der Teilaufgaben je KMK-Punktegruppe, d. h. die Aufgabencharakteristiken geben den prozentualen Anteil innerhalb der Gruppe von Studierenden einer KMK-Punktezahl an, die eine Aufgabe nicht gelöst haben.

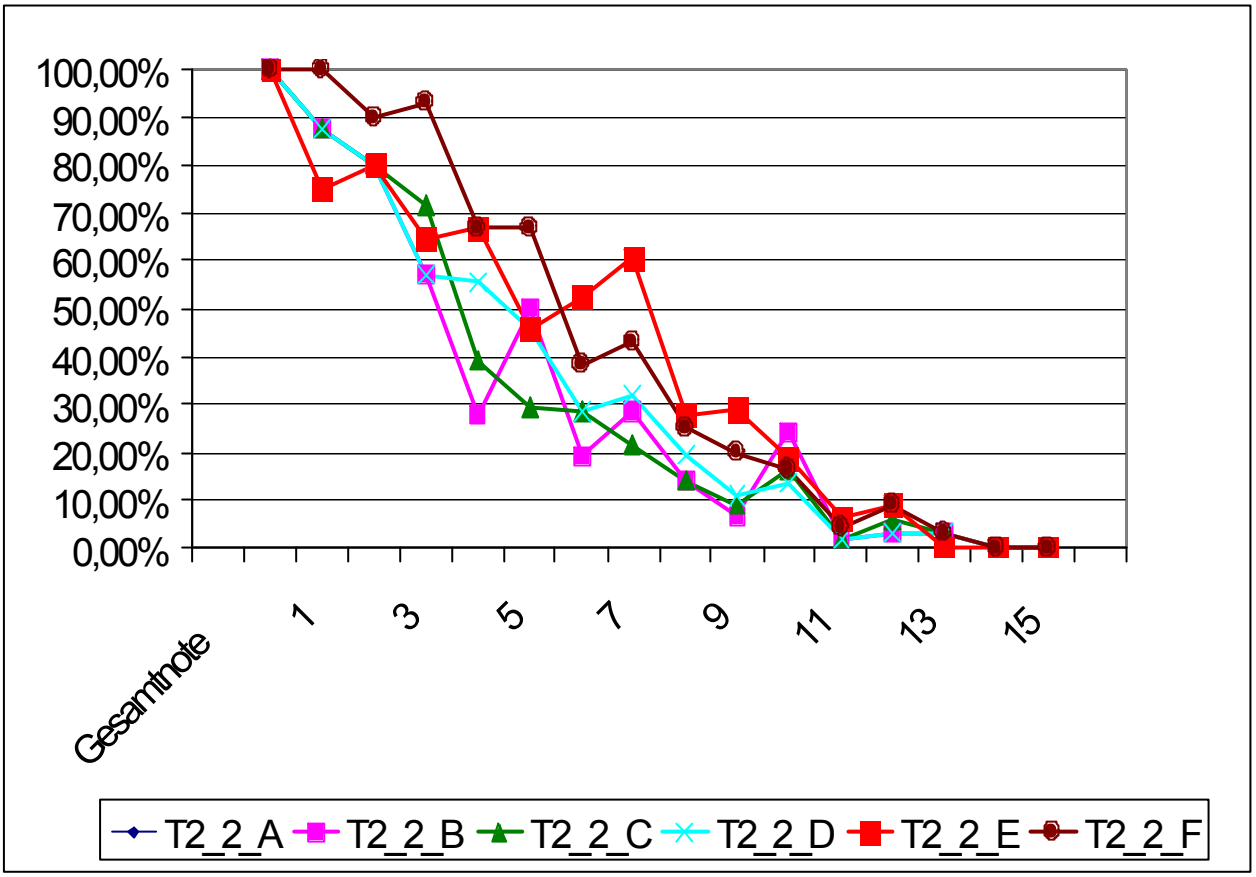


Abbildung 15: Aufgabencharakteristiken für Aufgabe 2, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Auch hier lässt sich aufgrund der Überschneidungen der Aufgabencharakteristiken nur eine Teilaufgabe eventuell einem Niveau zuordnen: Teilaufgabe f). Sieht man von den Überschneidungen der Aufgabencharakteristik mit Teilaufgabe e) ab, so ist Aufgabenteil f) am schwersten zu lösen gewesen.

4.2.5 Deskriptive Analyse Aufgabe 1 (Teil II - Grammatik)

1. Explain in your own words (Look at the context!):

a. they kept telling me (line 12)

_____

b. I felt caught between my friends and my parents. (lines 12/13)

_____

c. "in with the crowd" (line 14)

_____

d. "that it did not matter" (lines 20/21)

_____

8p

Abbildung 16: Aufgabe 1, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Für diese Aufgabe wurde die gleiche Kodierung wie für Aufgabe 2 des Grammatikteils (s. o.) vorgenommen.

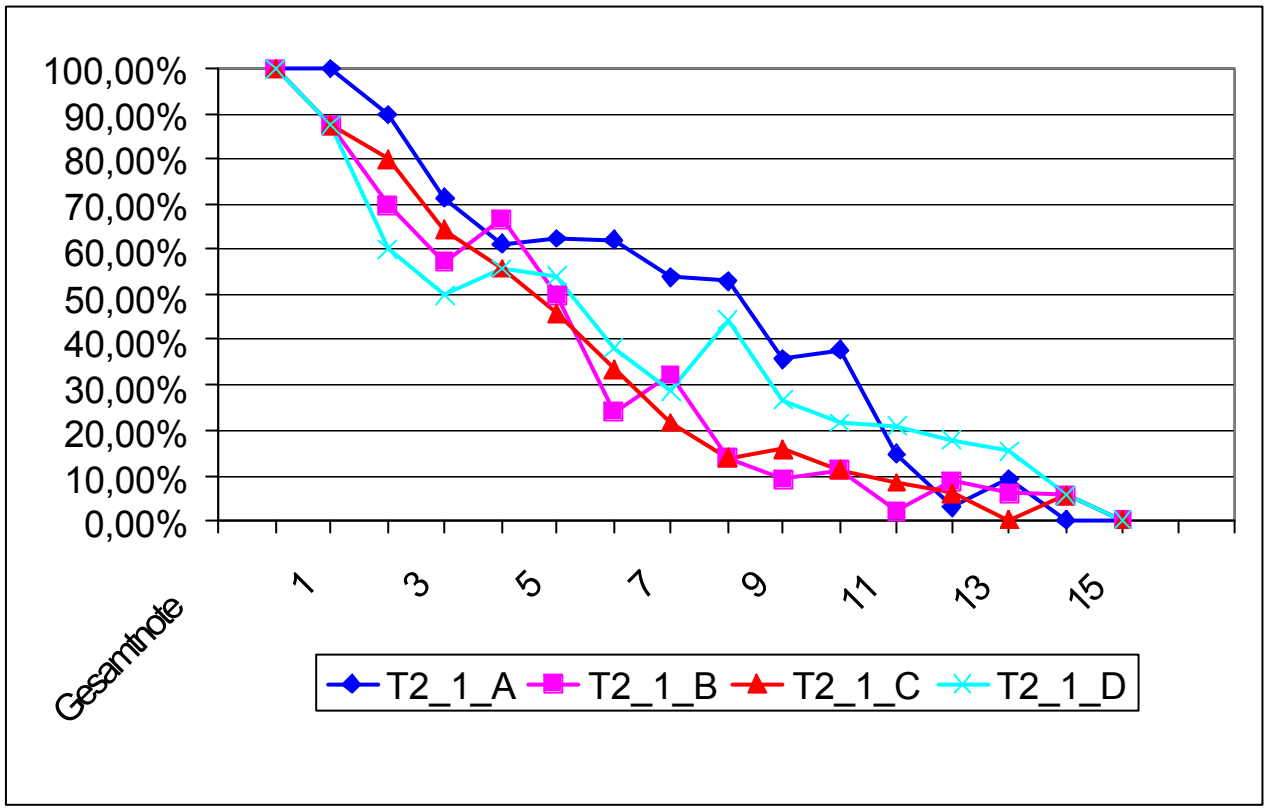


Abbildung 17: Aufgabencharakteristiken für Aufgabe 1, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Als Ergebnis lässt sich für diese Aufgabe feststellen, dass Teilaufgabe a) am schwierigsten zu lösen war, obwohl für diejenigen Studierenden, die 12 oder mehr KMK-Punkte erreichten, der Aufgabenteil d) die größere Hürde darstellte.

4.3 Wahrscheinlichkeitstheoretische Analyse

4.3.1 Aufgabe 1 (Teil I – Textarbeit)

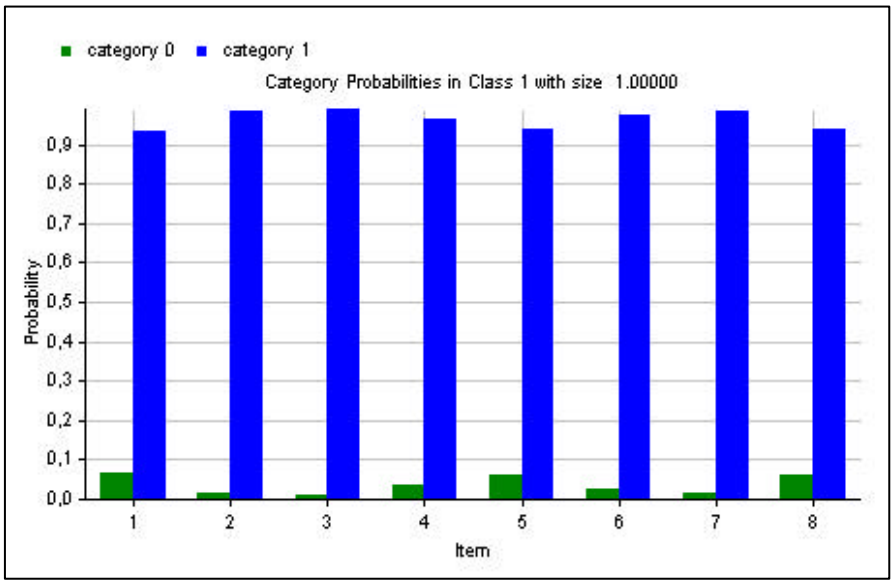


Abbildung 18: Lösungswahrscheinlichkeit der Teilaufgaben von Aufgabe 1, Teil I- Textarbeit für alle Studierenden (Vergleichsarbeit - Englisch)

Die Grafik gibt die Wahrscheinlichkeit für alle Studierenden an, dass die jeweilige Teilaufgabe gelöst wird. Zu erkennen ist, dass jede der Teilaufgaben eine Lösungswahrscheinlichkeit von über 92% hat. Unsere deskriptive Beschreibung der Aufgabe als zu leicht wird also bestätigt. Mit dieser Aufgabe kann die Leistungsfähigkeit der Studierenden nicht differenziert werden.

4.3.2 Aufgabe 4 (Teil II - Grammatik)

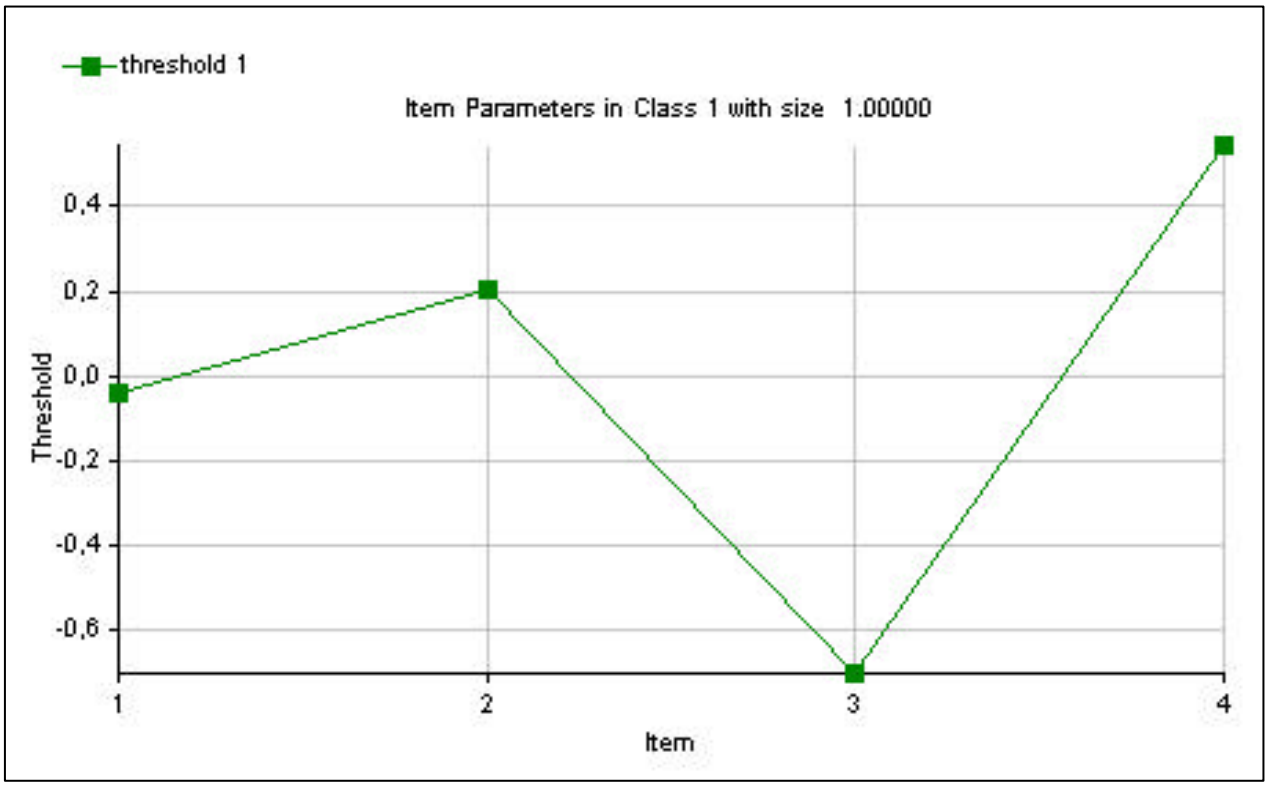


Abbildung 19: Aufgabenschwierigkeiten von Aufgabe 4, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Die Kodierung dieser Aufgabe erfolgt analog der deskriptiven Auswertung (s. o.). Die Grafik veranschaulicht die Aufgabenschwierigkeiten von Aufgabe 4 des Grammatikteils. Das Rasch-Modell bestätigt unsere deskriptive Analyse, dass Unteraufgabe c) am leichtesten ist. Die Analyse ermittelt aber Unteraufgabe d) als die Schwierigste. Schaut man sich aufgrund dieses Ergebnisses noch einmal die deskriptiven Aufgabencharakteristiken an, so wird auch deutlich, warum die Analyse zu diesem Ergebnis kommt: Unteraufgabe d) wird im Bereich von 6 bis 8 KMK-Punkten von unerwartet wenig Studierenden gelöst und der weitere Verlauf der Aufgabencharakteristik liegt grundsätzlich unter den anderen Verläufen. Die Interpretation, dass Unteraufgabe a) am schwierigsten zu lösen sei, beruhte auf der ausschließlichen Betrachtung des oberen Bereichs der KMK-Punkte. Die Ergebnisse der deskriptiven Analyse wurden somit nur zum Teil bestätigt.

4.3.3 Aufgabe 3 (Teil II - Grammatik)

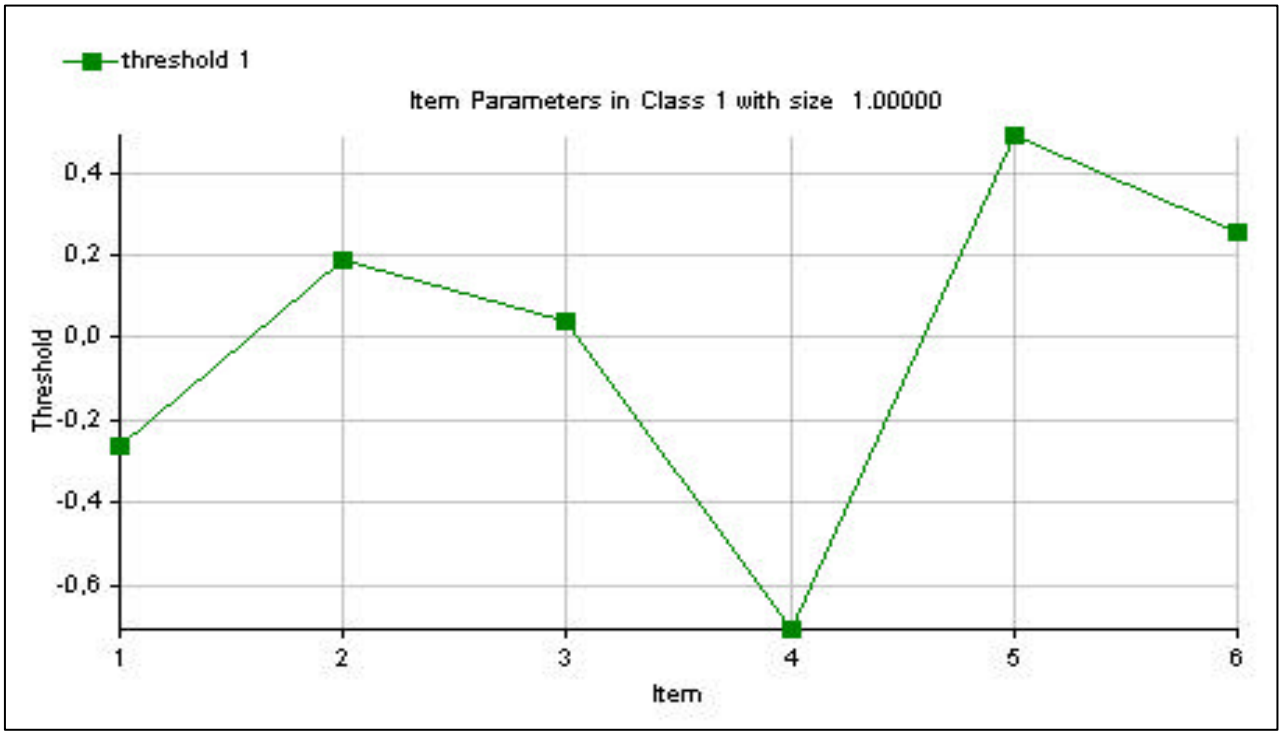


Abbildung 20: Aufgabenschwierigkeiten von Aufgabe 3, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Auch hier bestätigt die Rasch-Analyse unsere Vermutung, dass Unteraufgabe d) am leichtesten ist. Die Analyse liefert uns außerdem Aussagen über die Aufgabenschwierigkeiten für die anderen Unteraufgaben. Dies war mit der deskriptiven Analyse aufgrund der vielen Überschneidungen nicht möglich.

4.3.4 Aufgaben 1 und 2 (Teil II - Grammatik)

Für die ersten beiden Aufgaben des Grammatikteils der Vergleichsarbeit Englisch E2 konnten für jede Teilaufgabe maximal zwei Punkte vergeben werden. Es wurden jedoch nicht nur ganze, sondern auch halbe Punkte vergeben. Da diese Bewertungspraxis nicht einheitlich war, musste eine Umkodierung vorgenommen werden, die wir folgendermaßen gestalteten:

0 für eine Bewertung mit 0 Punkten,

1 für die Bewertung mit 0,5 Punkten und einem Punkt,

2 für die Bewertung mit 1,5 oder zwei Punkten.

Durch die so erfolgte Umkodierung erhält man für die beiden ersten Aufgaben des Grammatikteils Teilaufgaben mit jeweils drei Kategorien.

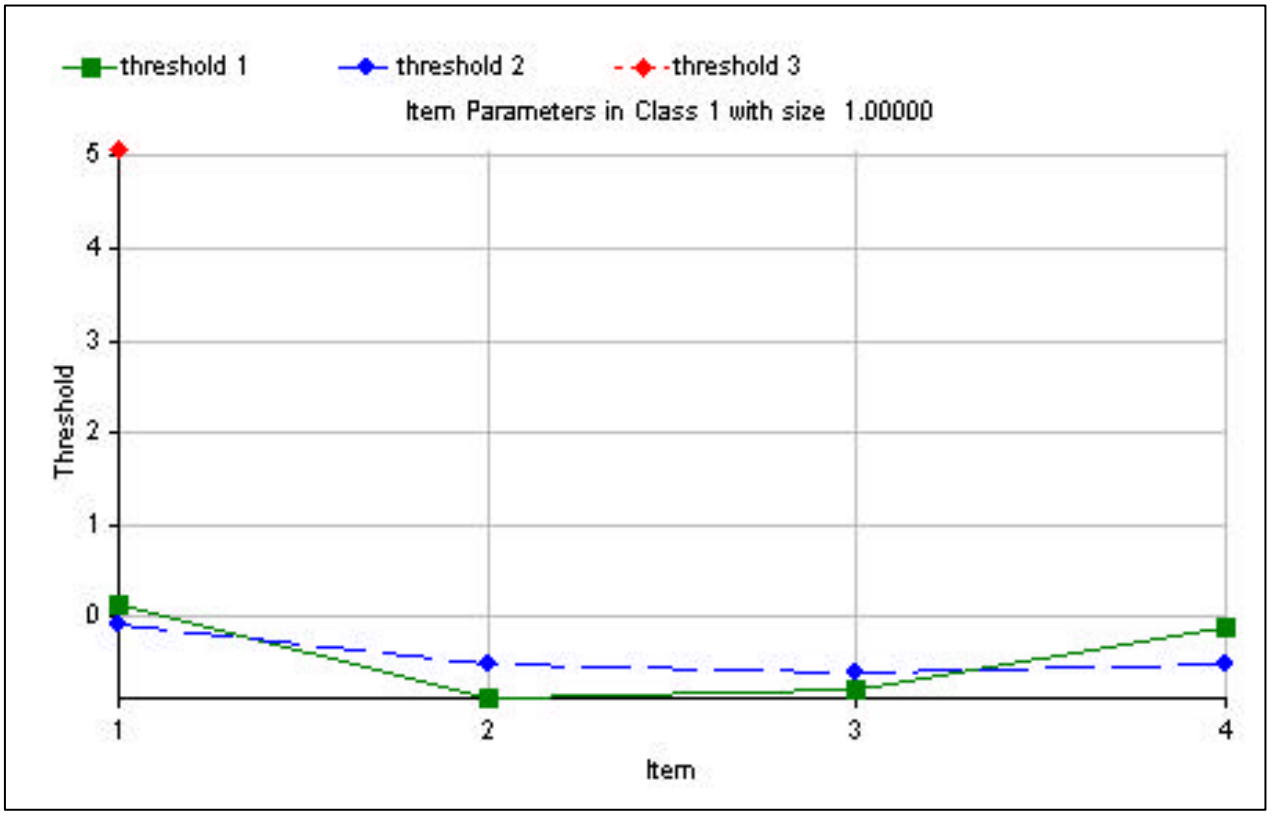


Abbildung 21: Aufgabenschwierigkeiten/ Schwellen von Aufgabe 1, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Die Analyse dieser Aufgaben erfolgte dann mit dem mehrkategoriellen Rasch-Modell, auch partial-credit-Modell genannt. Das Modell bestimmt aus den empirischen Daten für jede Teilaufgabe die Schwierigkeitsschwellen des Übergangs zwischen den einzelnen Kategorien. Mit Hilfe des Modells kann also den folgenden zwei Teilfragen nachgegangen werden: Wie groß ist die Schwierigkeit bei einer Teilaufgabe, die Kategorie 1 zu erhalten? Wie groß ist die Schwierigkeit bei dieser Teilaufgabe, die Kategorie 2 zu erreichen?

Für Aufgabe 1 lässt sich feststellen, dass die Teilaufgaben nur minimale Unterschiede in der Schwierigkeit der Lösung aufweisen. Die Aufgaben können somit nur zwischen denjenigen Studierenden diskriminieren, die diese Aufgaben lösen können, und denjenigen, die dies nicht können. Außerdem sind die Abstände zwischen den Schwierigkeitsschwellen so gering, dass die Diskriminierung der Studierenden anhand der unterschiedlichen Punktekategorien nicht möglich ist, denn aufgrund der Nähe der Schwierigkeitsschwellen können auch zufällige Effekte (Raten, falsch Raten) die richtige oder halbrichtige Lösung erzeugen. Voraussetzung ist nur, dass eine grundsätzliche Fähigkeit zur Lösung dieser Aufgabe vorhanden ist, um überhaupt in eine der Kategorien „Eins“ oder „Zwei“ zu gelangen.

Des Weiteren fällt auf, dass bei Teilaufgabe a) und d) das Erreichen der Kategorie „Zwei“ einfacher war, als die Schwierigkeit für Kategorie „Eins“. Dies kann mehrere Ursachen haben:

1. Durch die vorgenommene Umkodierung wurden entsprechend zu viele Studierende in die Kategorie 2 einsortiert.
2. Die Teilaufgaben sind widersprüchlich oder unverständlich formuliert und der Effekt beruht auf dem Erraten der Lösung.
3. Die Lehrkräfte haben die Teilaufgaben inkonsistent bewertet.

Obwohl die Analyse eine Bestätigung der Annahme liefert, dass Aufgabenteil a) am schwierigsten zu lösen war, bleibt auch hier festzustellen, dass diese Aufgabe und ihre Bewertung ungeeignet ist, die Leistungsfähigkeit der Studierenden zu bewerten.

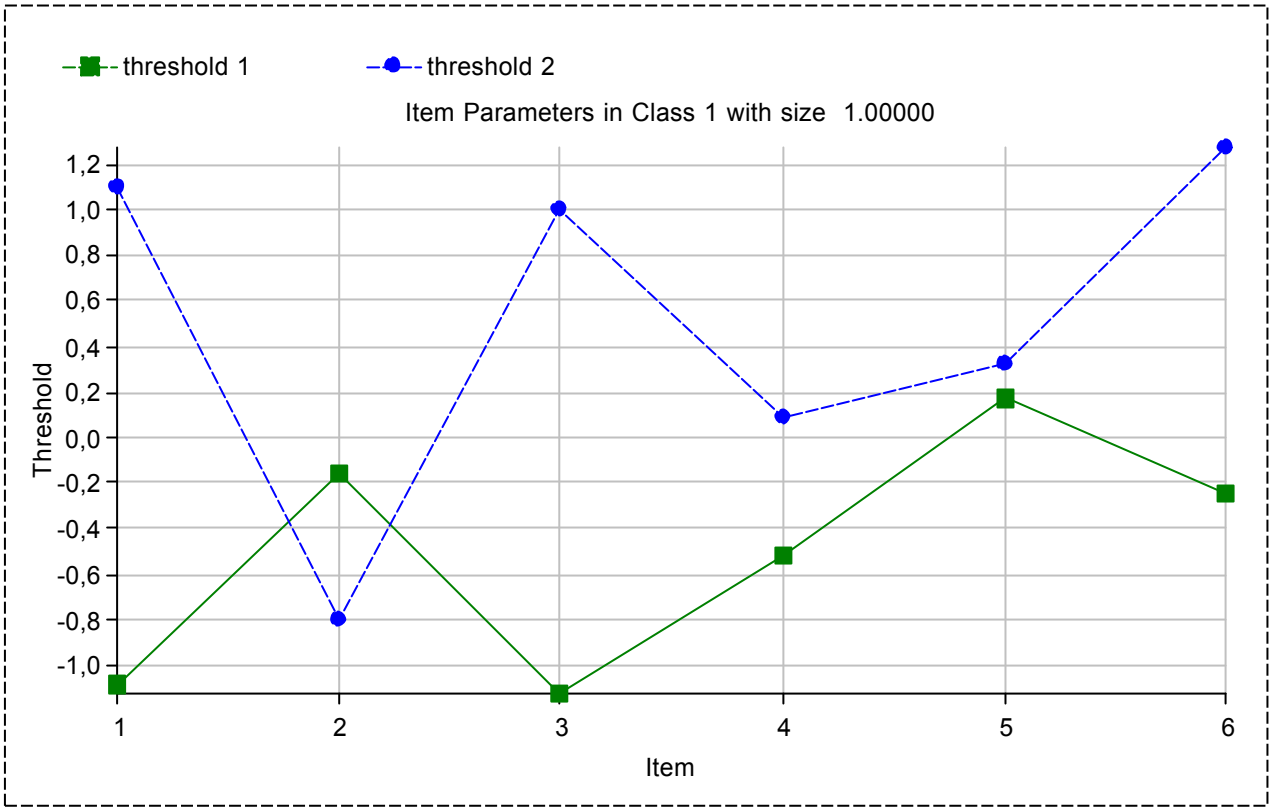


Abbildung 22: Aufgabenschwierigkeiten/ Schwellen von Aufgabe 2, Teil II - Grammatik (Vergleichsarbeit - Englisch)

Auch die Rasch-Analyse von Aufgabe 2 bestätigt die deskriptive Annahme, dass Aufgabenteil f) am schwierigsten zu lösen war. Des Weiteren wird erkennbar, dass bei dieser Aufgabe der Aufgabenteil b) am leichtesten zu lösen war. Die Aufgabenteile d) und e) befinden sich auf einem mittleren Schwierigkeitsniveau, Aufgabenteil a), c) und f) waren am schwierigsten zu lösen. Diese Aufgabe ist also in ihrer Gesamtheit in der Lage, drei unterschiedliche Leistungsniveaus bei den Studierenden zu identifizieren bzw. Studierende diesen Niveaus zuzuordnen.

Allerdings sind auch hier die Schwellenabstände zu beachten. Zum einen gibt es auch hier eine Teilaufgabe, Teilaufgabe b), bei der das Erreichen von Kategorie „Zwei“ einfacher als das Erreichen der Kategorie „Eins“ war. Außerdem sind die Schwellenabstände für die Teilaufgabe sehr unterschiedlich. Bei Teilaufgabe e) ist der Schwellenabstand sehr gering, so dass auch hier zufällige Effekte der richtigen oder halbrichtigen Lösung nicht ausgeschlossen werden können. Auch sind die Schwellenabstände von Teilaufgabe a) und c) sehr groß. Die Modellberechnung weist die halbrichtige Lösung dieser Teilaufgaben als die leichtesten Aufgaben aus, die richtige Lösung wird jedoch für beide Teilaufgaben als mit die Schwierigste ausgewiesen. Falls die Lehrkräfte oft einen halben Punkt vergeben haben, ließe sich dieser Effekt durch das Umkodieren erklären, da so die Anzahl der Studierenden in Kategorie „Eins“ unverhältnismäßig vergrößert wird. Es könnten aber auch grundsätzliche Probleme der Studierenden mit dieser Grammatikaufgabe sein, z. B. dass die Studierenden versucht haben, die Aufgabe irgendwie zu lösen und die Lehrkräfte diesen Lösungsversuchen eine entsprechende Punktwürdigkeit (ganzer oder halber Punkt) zusprechen. Im letztgenannten Fall würde sich genau der beschriebene Effekt durch die Umkodierung ergeben.

Für die Teilaufgaben d) und f) gilt: Die Schwierigkeit der Kategorie „Eins“ als auch der Kategorie „Zwei“ ist für Teilaufgabe f) größer als für Teilaufgabe d). Somit repräsentieren nur diese Teilaufgaben die Ansprüche, die an Teilaufgaben für eine Leistungsüberprüfung gestellt werden.

4.4 Erfahrungen bei der Dateneingabe

Die Korrekturanweisungen für die Bewertung der Vergleichsarbeiten waren zwar durch die Arbeitsgruppe einheitlich vorgegeben worden, doch ließen diese Anweisungen zu viel Spielraum, wie das Beispiel der Bewertung der Rechtschreibung bei Grammatikaufgaben zeigt. Hier sind eindeutige Regeln notwendig, damit unterschiedliche Bewertungen vermieden werden. Dies ist zweifellos der Kern für eine Vergleichbarkeit von Vergleichsarbeiten.

4.4.1 Bewertung freier Texte

Die Korrektur der freien Textteile konnte leider nicht nachvollzogen werden. Auch hier gab es von der Arbeitsgruppe ausgearbeitete Korrekturrichtlinien, doch war deren Anwendung insofern nicht nachvollziehbar, als dass die Benotung nicht auf einzelne Bewertungskriterien heruntergebrochen wurde, sondern nur global stattfand. Daher konnten die inhaltlichen Aufgaben nicht mittels einer Aufgabencharakteristik ausgewertet werden.

In Einzelfällen gab es auch eine inhaltlich nicht nachvollziehbare Benotung. Im Folgenden sind einige Einzelbeispiele dokumentiert, die veranschaulichen sollen, warum sich hier die Thematik einer „Zweikorrektur“ aufdrängt. Eine systematische Analyse dieser Problematik lag allerdings nicht im Aufgabenbereich des Projektes. Bei der Dateneingabe drängten sich aber einige solcher Beispiele auf, von denen Eines hier dokumentiert wird. Dabei geht es um die folgende Arbeitsaufgabe: “What do you think about the idea of teenagers living on their own at age eighteen? State your opinion and give reason. Write about 100 – 120 words.”

Beispiel 1:

I personally think that moving out at the age of eighteen is wrong because there is still so much to discover and experience, and so much that hasn´t been understood yet. My parents made me move out when I was seventeen, which was the worst thing to do. When I see young punks in Frankfurt my heart breaks, because I know how an early moving out from home can minimize the chance for a normal life. On the other hand, of course, there are a few teenagers who are willing and able to organize their lives by themselves early. But who knows which teenager will fail and which one won´t?

(Ausdruck: 32 von 32 Punkten)

Abbildung 23: Einzelbeispiel 1, Teil III - freier Text, Bewertung (Vergleichsarbeit - Englisch)

Beispiel 2:

In my opinion it is a good idea, because it is good for them to learn early the way of live. They should talk with their parents about the realising for the flat and a little help in the first time. The parents mostly know it if their children are independently enough to live in a own flat.

(Ausdruck: 32 von 32 Punkten)

Abbildung 24: Einzelbeispiel 2, Teil III - freier Text, Bewertung (Vergleichsarbeit - Englisch)

Beispiel 3:

I think it is not very good, when teenagers living on their parents, because they will make mistakes alone without parents. They children haven´t more experions as their parent. They need still love from mother and father.

(28 von 32 Punkten)

Abbildung 25: Einzelbeispiel 3, Teil III - freier Text, Bewertung (Vergleichsarbeit - Englisch)

Beispiel 4:

Well, I think it is good for young people when they move out from home in the age of 18.

They learn to manage their life and solve the problems on their own. Maybe, in the beginning it is not easy to live alone – you have to change many habbits. Life is very expensive and you must learn to save your money.

But all the problems you solve and pass are make you stronger and cleverer.

On that way you learn to take responsibility for you own and maybe for other people you live with in the future.

But I also think, that it is good, when do you have got parents who support you to manage your own live. It is good to have a place where you can go to, if you have big trouble in live. Sometimes, your parents' experience of life can help you to manage your own life better.

(Ausdruck: 32 von 32 Punkten)

Abbildung 26: Einzelbeispiel 4, Teil III - freier Text, Bewertung (Vergleichsarbeit - Englisch)

4.4.2 Rundungsfehler/ Rechenfehler bei der Punktevergabe

Bei der Berechnung der Gesamtnote wurden unterschiedliche Rundungsmethoden angewandt. Dies ist eine Frage konkreter Rundungsregeln im Rahmen der Korrekturanweisungen.

Weiterhin gab es schlichte Rechenfehler bei der Addition von Punkten bzw. der Errechnung der Gesamtnote. Solche Fehler könnten durch eine den Lehrkräften zur Verfügung gestellte Tabellenkalkulationsdatei, in die nur noch die Einzelergebnisse der jeweiligen Arbeitsaufgaben eingetragen werden müssen, vermieden werden.

5 Analyse der Vergleichsarbeit Mathematik (R4)

5.1 Notwendige Aufgabenprüfung und Bewertungstransformation

Um die Aufgaben der Vergleichsarbeit Mathematik (R4) einer Analyse nach dem Rasch-Modell zugänglich zu machen, muss zunächst eine Analyse der Aufgaben und eine Modifikation (Umkodierung) der Bewertung vorgenommen werden:

Die Vergleichsarbeit Mathematik R4 umfasste u. a. zwei inhaltsbezogene Teilbereiche. In einem ersten Teilbereich sollten die Leistungen der Studierenden bei Berechnungen an Körpern und Flächen überprüft werden. In einem zweiten Teil ging es um die Bearbeitung von quadratischen Gleichungen. Wie oben ausgeführt, erfordert die Anwendung des Rasch-Modells, dass sich Aufgaben auf ein und dieselbe Fähigkeit beziehen müssen, wenn Aufgaben bezüglich ihres Schwierigkeitsgrades analysiert werden sollen. Dies bedeutet für die hier zu untersuchende Vergleichsarbeit Mathematik (R4), dass die jeweiligen Aufgaben in den beiden o. g. Teilbereichen auch tatsächlich nur auf diese eine jeweils interessierende Fähigkeit rekurrieren. Daher erfolgt zunächst eine Überprüfung der Aufgaben dahingehend, ob sie diese Eigenschaft erfüllen.

Um die Aufgaben mittels des Rasch-Modells auf ihre Schwierigkeit hin zu untersuchen, muss außerdem der Bewertung der Aufgaben die gleiche Bewertungsskala zugrunde liegen. Dies war nicht der Fall. Wir haben daher für unsere Auswertung eine Umkodierung der Bewertung der Aufgaben bzw. Teilaufgaben vornehmen müssen: Jede Teilaufgabe wurde mit „0“ codiert, sofern ihr bis zur Hälfte der erzielbaren Punkte zugesprochen worden war, und mit „1“ codiert, sofern ihr mindestens die Hälfte der möglichen Punktzahl zugesprochen war.

5.1.1 Teilbereich: Berechnung an Flächen und Körpern

5.1.1.1 Aufgabe 1)

Berechnungen an Flächen und Körpern		Punkte
1.	Das Becken eines Schwimmbades hat die Form eines Quaders. Es ist 50m lang, 20m breit und 3m hoch. a. Wie viel m^2 Fliesen werden für den Boden des Schwimmbades benötigt? b. Für die Materialberechnung werden zusätzlich 15% Verschnitt berechnet. Wie viel m^2 Fliesen müssen insgesamt eingekauft werden? c. Das Schwimmbecken wird bis zu einer Höhe von 2,8m mit Wasser gefüllt. In einer Minute fließen 5m^3 Wasser zu. Wie lange dauert der Füllvorgang? Geben Sie Ihr Ergebnis in Stunden und Minuten an.	8

Abbildung 27: Arbeitsaufgabe 1 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 1a):

Bei dieser Teilaufgabe soll die Fläche eines Rechtecks aus den gegebenen Seitenlängen berechnet werden. Somit überprüft die Aufgabe die Fähigkeit des Studierenden, Berechnungen an Flächen und Körpern durchzuführen.

Zu Aufgabe 1b):

Zu dieser Teilaufgabe sind zwei Bemerkungen zu machen. Erstens überprüft diese Aufgabe nicht die Fähigkeit, Berechnungen an Körpern und Flächen durchzuführen, denn bei dieser Aufgabe handelt es sich um eine Aufgabe zur Prozentrechnung. Zweitens ist die Lösbarkeit dieses Aufgabenteils abhängig von der Lösung von Teil a). Diese Abhängigkeit der Lösung widerspricht auch einer genannten Forderung des Rasch-Modells.

Zu Aufgabe 1c):

Diese Aufgabe besteht wiederum aus zwei Teilaufgaben, die jeweils unterschiedliche Fähigkeiten kontrollieren. Der erste Teil überprüft die Fähigkeit der Berechnungen an Flächen und Körpern, denn es soll ein Volumen aus den drei Grundseiten eines Körpers berechnet werden. Der zweite Teil ermittelt hingegen die Fähigkeit zur Lösung eines Dreisatzes.

Fazit für Aufgabe 1):

Aufgabenteil a) ist geeignet die geforderte Fähigkeit zu überprüfen. Ebenso der erste Teil von Aufgabe c), wenn bei einer Korrektur auf diese Zweiteilung eingegangen würde, d. h. wenn die Berechnung des Volumens und die Bewertung dieser Berechnung unabhängig von der Lösung des Dreisatzes erfolgen würde.

5.1.1.2 Aufgabe 2)

2. Berechnen Sie die gesuchten Stücke des dreiseitigen Prismas. Die Grundfläche wird durch ein gleichseitiges Dreieck begrenzt. Geben Sie Ihren Rechenweg an.
(Als Hilfsmittel sind zugelassen: Formelsammlung „Mathematik Realschule“, nicht programmierbarer und nicht grafikfähiger Taschenrechner. Schablonen dürfen nicht verwendet werden.)

Länge der Dreieckseite	Höhe der Grundfläche	Größe der Grundfläche	Höhe des Prismas	Volumen des Prismas
4,2 cm		7,64 cm^2		64,94 cm^3

6

Abbildung 28: Arbeitsaufgabe 2 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 2)

Wie bei der Korrekturanweisung durch die Arbeitsgruppe angegeben, lässt sich diese Aufgabe als zwei einzelne Aufgaben interpretieren, die jeweils die Fähigkeit der Berechnung an Flächen und Körpern überprüfen.

5.1.1.3 Aufgabe 3)

3.	Gegeben ist ein zusammengesetzter Körper mit quadratischer Grundfläche. (Maße sind in cm in einer entsprechenden Zeichnung angegeben.) a. Berechnen Sie das Volumen des zusammengesetzten Köpers. b. Ein cm^3 Stahl wiegt 7,85g. Welche Masse hat ein solcher Köper aus Stahl?	8

Abbildung 29: Arbeitsaufgabe 3 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 3a):

Da es sich bei dieser Aufgabe um die Berechnung eines Volumens handelt, wird durch diese Aufgabe die Fähigkeit des Studierenden, Berechnungen an Flächen und Körpern durchzuführen, überprüft.

Zu Aufgabe 3b):

Auch dieser Aufgabenteil ist in zweierlei Hinsicht ungeeignet, die geforderte Fähigkeit zu überprüfen. Zum einen handelt es sich bei der geforderten Leistung um eine reine Multiplikation (Volumenergebnis multipliziert mit dem spezifischen Gewicht) und zum anderen ist die Lösung dieser Aufgabe von der Richtigkeit des Ergebnisses von Teilaufgabe a) abhängig.

Fazit für Aufgabe 3):

Nur der Aufgabenteil a) leistet eine Überprüfung der geforderten Fähigkeit.

5.1.1.4 Aufgabe 4)

4.	Die Fläche eines runden Platzes ist 314m^2 groß. Berechnen Sie den Umfang des Platzes.	4

Abbildung 30: Arbeitsaufgabe 4 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 4):

Bei dieser Aufgabe handelt es sich um eine klassische Kreisberechnung. Somit wird durch diese Aufgabe die Fähigkeit des Studierenden, Berechnungen an Flächen und Körpern durchzuführen, überprüft.

5.1.1.5 Aufgabe 5)

Satz des Pythagoras		
5.	Die Cheopspyramide in Ägypten hat eine quadratische Grundfläche. Ihre Grundkante beträgt 230m, die Höhe h_s der Seitenfläche ist 186m lang. (Eine entsprechende Zeichnung war gegeben.) Berechnen Sie a. die Höhe der Pyramide und b. die Länge der Seitenkante s.	7

Abbildung 31: Arbeitsaufgabe 5 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 5):

In beiden Aufgabenteilen sollen Berechnungen an einer quadratischen Pyramide vorgenommen werden. Da beide Aufgaben unabhängig voneinander lösbar sind, eignen sich die Aufgabenteile gleichermaßen, um zu ermitteln, in welcher Weise die Studierenden über die Fähigkeit verfügen, Berechnungen an Flächen und Körpern durchzuführen.

5.1.2 Teilbereich: Quadratische Gleichungen

5.1.2.1 Aufgabe 7)

Quadratische Gleichungen		Punkte
7.	Lösen Sie die folgenden Gleichungen (Grundmenge ist R.) Eine unlösbare Gleichung ist möglich. a. $x^2 - 0{,}09 = 0$ b. $x^2 - 16x + 164 = 0$ c. $14x + 9{,}8 = -5x^2$ d. $(x+6)^2 = -2\,(x^2 + 15x + 1{,}5)$	14

Abbildung 32: Arbeitsaufgabe 7 (Vergleichsarbeit - Mathematik)

Zu Aufgabe 7):

Bei allen Teilaufgaben handelt es sich um Aufgaben, die mittels quadratischer Ergänzung oder „p-q-Formel“ gelöst werden können. Daher bilden die Aufgaben eindeutig die Fähigkeit des Studierenden ab, quadratische Gleichungen zu lösen.

5.2 Wahrscheinlichkeitstheoretische Analyse

5.2.1 Analyse der Aufgaben „Berechnung an Flächen und Körpern“

Aus der obigen Analyse ergibt sich, dass nur die (Teil-)Aufgaben 1a), 2 [nachträglich differenziert in 2a) und 2b)], 3a), 4), 5a) und 5b) für eine Betrachtung mit Hilfe des Rasch-Modells geeignet sind, wenn man sich auf die Fähigkeit beziehen will, Flächen und Körper zu berechnen.

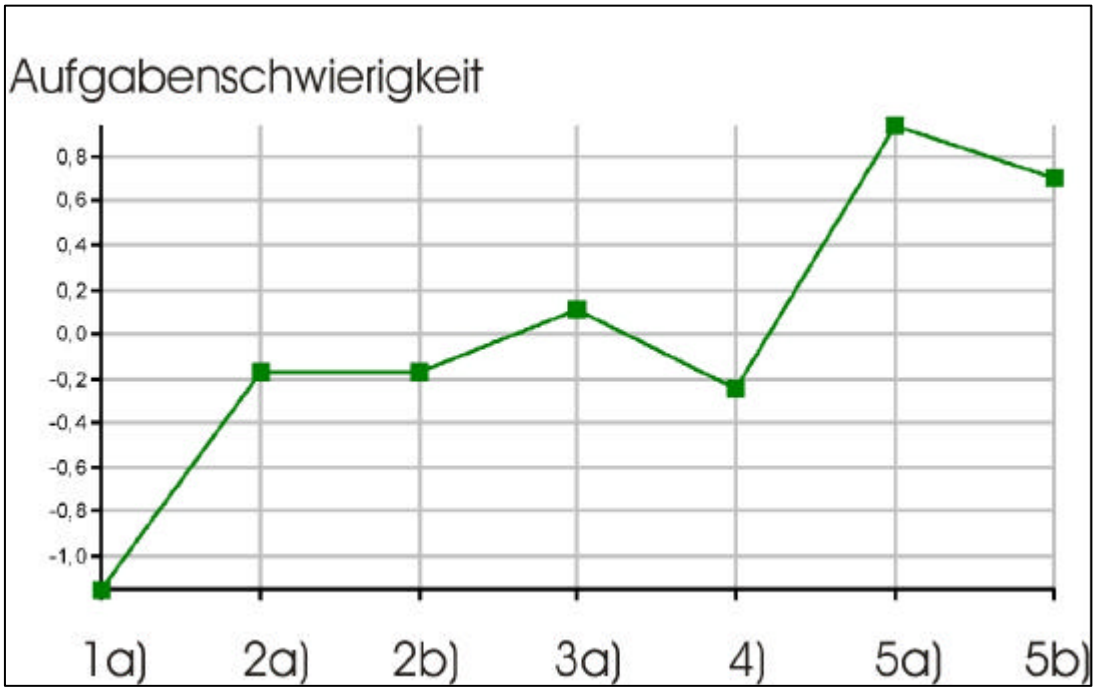


Abbildung 33: Aufgabenschwierigkeiten der ausgewählten Aufgaben zu Berechnungen an Flächen und Körpern (Vergleichsarbeit - Mathematik)

Die Analyse der Aufgabenschwierigkeit zeigt, dass die Aufgaben zur Berechnung der Fläche eines Rechtecks am leichtesten und die Aufgaben der Berechnungen von Höhen in einer Pyramide (Anwendung des Satzes von Pythagoras) für die Studierenden am schwersten waren. Des Weiteren zeigt sich, dass die Aufgaben nur drei unterschiedliche Schwierigkeitsniveaus aufweisen:

Aufgabe 1a) liegt auf einem leichten Schwierigkeitsniveau, die Aufgaben 2a), 2b), 3a) und 4) befinden sich auf einem mittleren Schwierigkeitsniveau und die Aufgaben 5a) und 5b) weisen ein hohes Schwierigkeitsniveau auf. Die Aufgaben repräsentieren somit nur drei Leistungsstufen, obwohl die Differenzierung der Leistung nach Schulnoten sechs Stufen aufweist.

Man kann also festhalten: Die hier genutzten Aufgaben können nicht zwischen Studierendenleistungen in dem für Schulnoten üblichen Bereich von sechs Leistungsstufen differenzieren.

Die folgende Grafik stellt dar, wie viele Studierende (Häufigkeit) jeweils eine bestimmte Anzahl von Aufgaben (0 bis 7) gelöst haben.

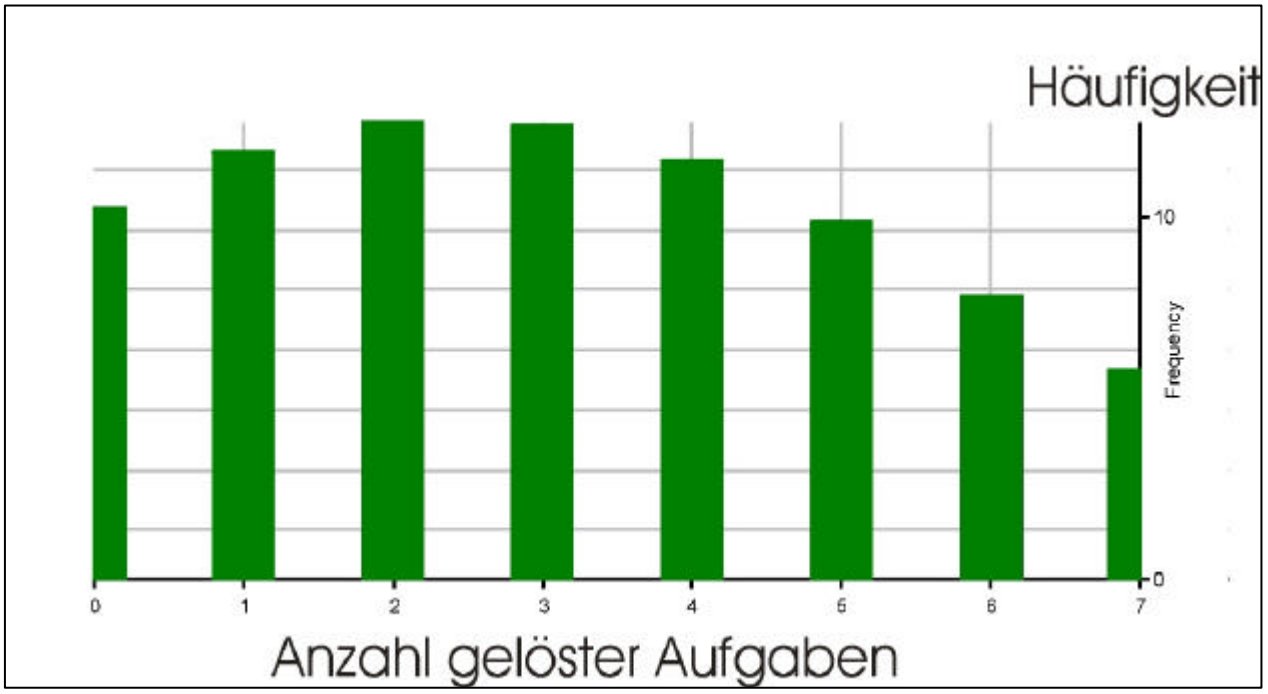


Abbildung 34: Lösungshäufigkeiten der Anzahl von gelösten Aufgaben der ausgewählten Aufgaben zu Berechnungen an Flächen und Körpern (Vergleichsarbeit - Mathematik)

Es ist zu erkennen, dass es einen nennenswerten Anteil an Studierenden gibt, die *keine einzige* Aufgabe lösen können (Anzahl gelöster Aufgaben = 0). Ebenso gibt es einen zwar kleineren, aber dennoch zu beachtenden Teil von Studierenden, die *alle* Aufgaben lösen (Anzahl gelöster Aufgaben = 7). Für die Studierenden, die keine Aufgabe gelöst haben, waren demnach alle gestellten Aufgaben zu schwer. Der Erkenntnisgewinn über die Leistungsfähigkeit dieser Gruppe von Studierenden ist nur derjenige, dass sie leistungsschwächer als die minimalste Leistungsanforderung ist, die durch die Aufgaben überprüft wird. Dies gilt in umgekehrter Weise auch für diejenigen Studierenden, die *alle* Aufgaben gelöst haben. Diese Studierenden erbringen zwar die maximale Leistung, die durch die Aufgaben bezogen auf das Thema überprüft wird, doch diese Studierenden könnten auch über eine noch höhere Leistungsfähigkeit verfügen.

Man kann also festhalten: Die Aufgaben erfassen nicht das gesamte Spektrum der Leistungsfähigkeit der Studierenden! Die Leistungen der Studierenden in den Randbereichen sind durch die vorliegenden Aufgaben mit maximalem bzw. minimalem Schwierigkeitsniveau nicht eindeutig diagnostizierbar. Um zu einer Lernunterstützung und Lerndiagnostik für die Studierenden der Randbereiche zu gelangen, müssten die Aufgaben sowohl die untersten als auch die höchsten Schwierigkeitsniveaus abdecken, welche die Studierenden zu bearbeiten vermögen. Eine diagnostisch fundierte Förderung der Studierenden in beiden Randbereichen der Leistung ist erst möglich, wenn Aufgaben diese Leistungsniveaus adäquat erfassen.

5.2.2 Analyse der Aufgaben „Berechnung von quadratischen Gleichungen“

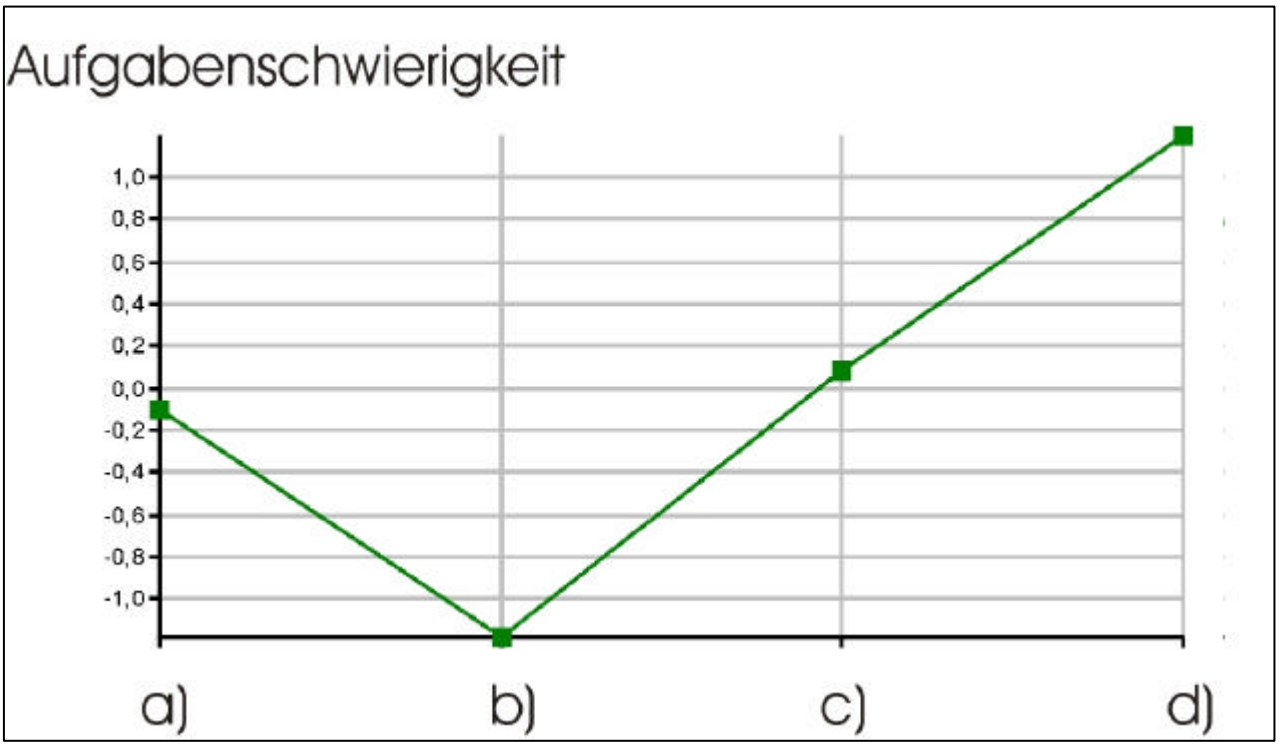


Abbildung 35: Aufgabenschwierigkeiten von Aufgabe 7 zur Lösung von quadratischen Gleichungen (Vergleichsarbeit - Mathematik)

Die Sichtung der Aufgaben ergab, dass sich alle vier Aufgabenteile (7a-d) für die Überprüfung der Fähigkeit zur Lösung von quadratischen Gleichungen eignen, daher konnte eine Analyse durchgeführt werden.

Die Auswertung nach dem Rasch-Modell identifiziert bei Teilaufgabe b) das geringste Anforderungsniveau. Bei dieser Teilaufgabe handelt es sich um die Standardform einer quadratischen Gleichung. Die Lösung dieser Aufgabe ergibt sich daher durch einfaches Einsetzen der Werte in die p-q-Formel. Die Studierenden müssen also nur ein schematisches Verfahren anwenden.

Die Modellauswertung identifiziert Aufgabenteil d) als die schwerste Teilaufgabe. Zum einen ist eine binomische Formel auszurechnen und zum anderen steht auf der rechten Seite des Gleichheitszeichens vor der Klammer ein Minuszeichen. Das Auflösen dieser Klammer scheint der Grund zu sein, dass dieser Aufgabenteil für viele Studierende eine besondere Herausforderung darstellt.

Die Aufgabenteile a) und c) entsprechen einander in ihrem mittleren Schwierigkeitsgrad. In Aufgabenteil a) müssen die Studierenden für die Anwendung der p-q-Formel erkennen, dass der Faktor des gemischten Terms gleich Null ist bzw. sie müssen eine Äquivalenzumformung der Formel vornehmen, die auf einer Addition beruht, um dann den Wert in den Taschenrechner eingeben zu können. Aufgabenteil c) ergibt durch die Äquivalenzumformung einer Addition die Standardform zur Anwendung der „a-b-c-Formel“, d. h. nach einer leichten Umformung kann wieder ein schematisches Verfahren zur Lösung der Gleichung angewendet werden. Die erforderlichen mathematischen Operationen zur Lösung von Aufgabenteil a) und c) können daher als vergleichbar angesehen werden, wodurch sich der ähnliche Schwierigkeitsindex beider Aufgaben erklären ließe.

Man kann also festhalten: Der Schwierigkeitsgrad dieser Aufgabenteile würde sich gemäß der zuvor angestellten Überlegungen an der vorzunehmenden Äquivalenzumformung festmachen lassen. Eine leichte Aufgabe lässt sich durch simples Einsetzen in ein auswendig gelerntes Lösungsschema lösen, schwierigere Aufgaben lassen erst nach Umformung die Anwendung des Lösungsschemas zu. Dabei nimmt die Aufgabenschwierigkeit mit der Komplexität der anzuwendenden Umformung zu.

Identifiziert werden auch hier nur drei Schwierigkeitsniveaus: ein leichtes, ein mittleres und ein schweres. Somit ergibt sich auch in diesem Zusammenhang das bereits oben festgestellte Problem der Zuweisung von drei Schwierigkeitsstufen zu einer Skala von sechs Schulnoten.

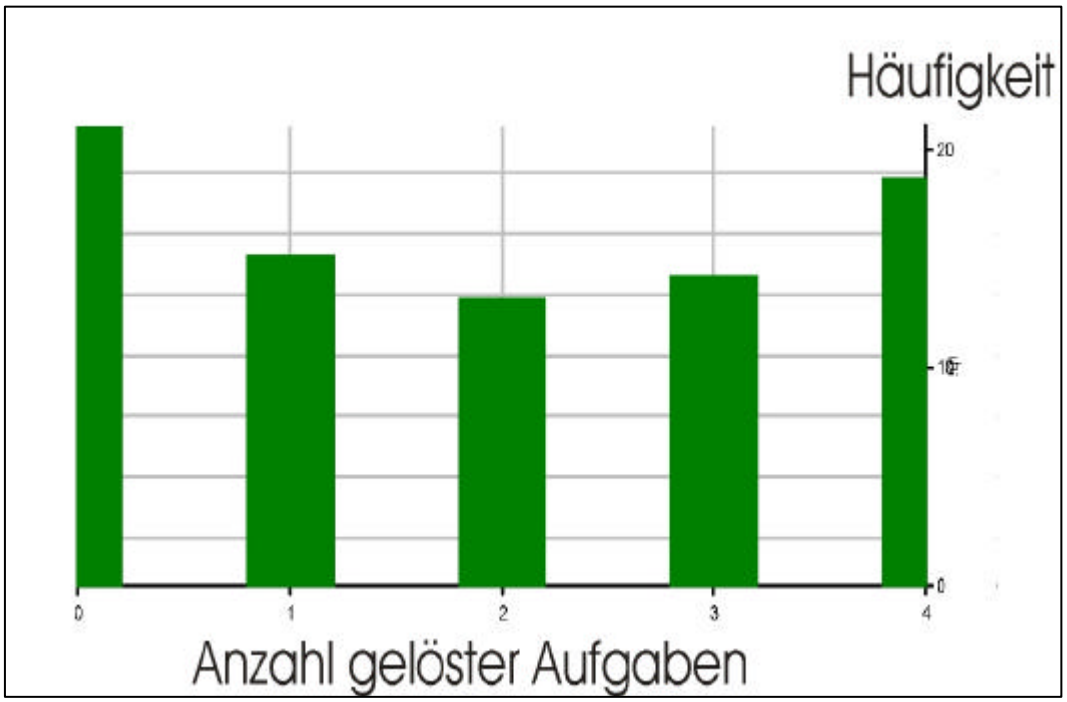


Abbildung 36: Lösungshäufigkeiten der Anzahl von gelösten Aufgaben von Aufgabe 7 zur Lösung quadratischer Gleichungen (Vergleichsarbeit - Mathematik)

Die Häufigkeitsverteilung der Anzahl gelöster Aufgaben pro Studierendem weist - wie bei den Aufgaben zu Berechnungen an Flächen und Körpern - hohe Anteile von Studierenden aus, die keine Aufgabe bzw. alle Aufgaben gelöst haben. Somit gelten die gleichen Überlegungen bezüglich der diagnostischen Möglichkeiten an den Rändern des Leistungsniveaus.

6 Analyse der Vergleichsarbeit Deutsch (E2)

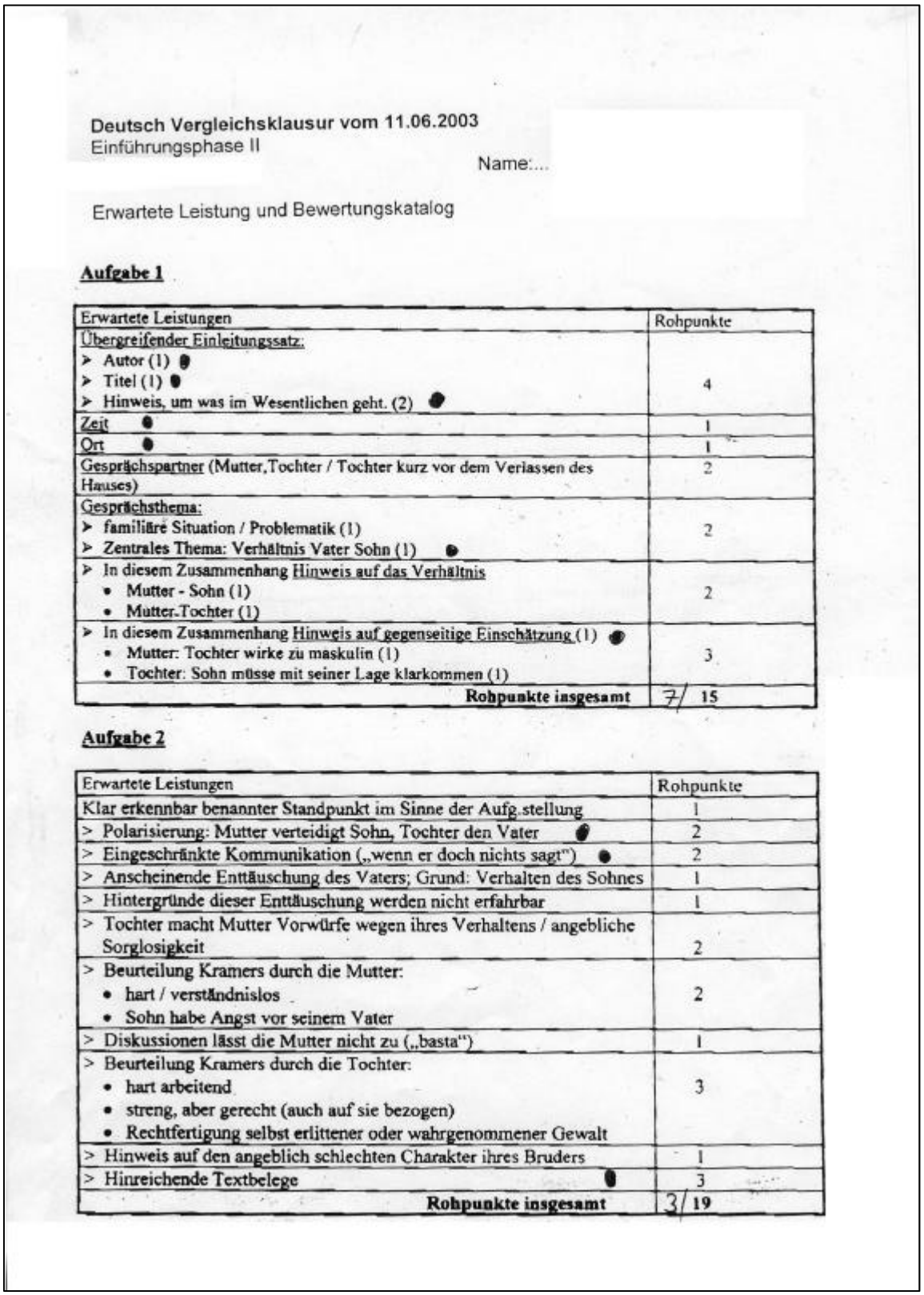
6.1 Das Problem der (fehlenden) Explikation/ Nachvollziehbarkeit von Bewertungskriterien

Die folgenden Anmerkungen und Bewertungsbeispiele zur Analyse der Vergleichsarbeit Deutsch (E2) stehen beispielhaft für die Schwierigkeiten der Auswertung der Vergleichsarbeiten Deutsch in den anderen Schulformen der SfE bzw. für die Schwierigkeiten der Auswertung freier Textteile unter Berücksichtigung der vorliegenden Bewertungen durch die Lehrkräfte.

Für die Analyse der Aufgaben oder aufgabenübergreifenden Fähigkeiten ist es erforderlich, dass die Bewertung der Aufgaben oder Fähigkeiten einheitlich erfolgt. Von den Arbeitsgruppen wurden dazu weitgehende Bewertungsvorschläge erstellt. Die nachstehenden Bewertungsbeispiele stellen einen Querschnitt der angetroffenen Bewertungsstile dar und sollen zugleich dokumentieren, warum eine nach Aufgaben oder Fähigkeiten differenzierende Auswertung gar nicht erst möglich war. Die Anordnung der Beispiele richtet sich nach dem Grad der Differenziertheit bzw. Nachvollziehbarkeit der Bewertungen.

Es ist zu erkennen, dass die Bewertungsstile von einer sehr gut dokumentierten, die einzelnen Anforderungen überprüfenden und Transparenz erzeugenden Bewertung bis hin zu einer Bewertung reichen, die nur aus der Note bzw. KMK-Punktzahl besteht. Eine Bewertung im Stile der beiden ersten Beispiele ist die notwendige Voraussetzung, um eine statistische Auswertung mittels des Rasch-Modells prinzipiell zu ermöglichen. Da solche Bewertungsformen jedoch nur in Ausnahmefällen praktiziert wurden, konnte keine Aufgaben- oder Anforderungsdiagnostik vorgenommen werden.

6.2 Beispiele für Bewertungsstile



Deutsch Vergleichsklausur vom 11.06.2003
Einführungsphase II

Name:...

Erwartete Leistung und Bewertungskatalog

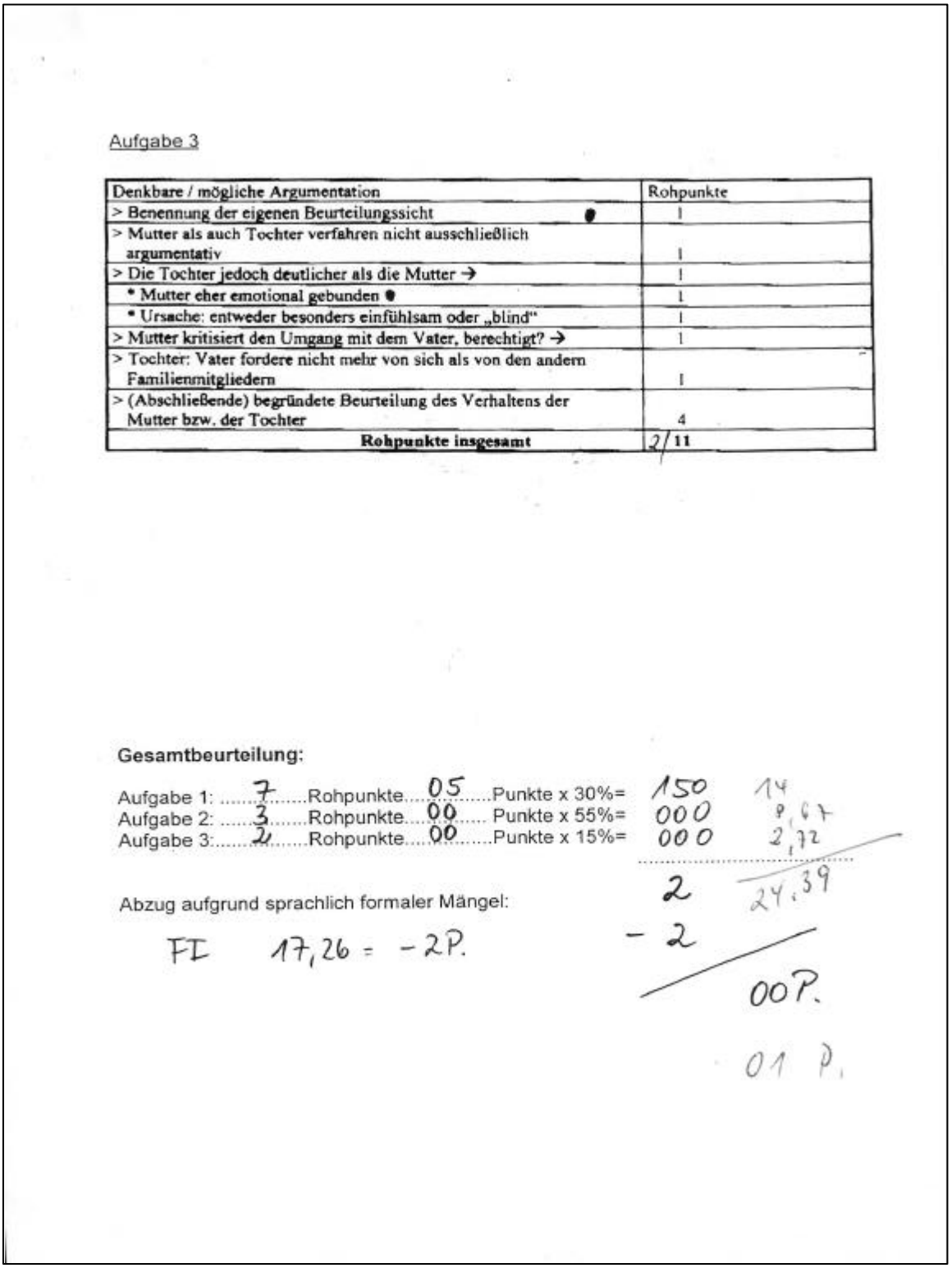
Aufgabe 1

Erwartete Leistungen	Rohpunkte
Übergreifender Einleitungssatz: ➢ Autor (1) ➢ Titel (1) ➢ Hinweis, um was im Wesentlichen geht. (2)	4
Zeit	1
Ort	1
Gesprächspartner (Mutter,Tochter / Tochter kurz vor dem Verlassen des Hauses)	2
Gesprächsthema: ➢ familiäre Situation / Problematik (1) ➢ Zentrales Thema: Verhältnis Vater Sohn (1)	2
➢ In diesem Zusammenhang Hinweis auf das Verhältnis • Mutter - Sohn (1) • Mutter-Tochter (1)	2
➢ In diesem Zusammenhang Hinweis auf gegenseitige Einschätzung (1) • Mutter: Tochter wirke zu maskulin (1) • Tochter: Sohn müsse mit seiner Lage klarkommen (1)	3
Rohpunkte insgesamt	7/ 15

Aufgabe 2

Erwartete Leistungen	Rohpunkte
Klar erkennbar benannter Standpunkt im Sinne der Aufg.stellung	1
> Polarisierung: Mutter verteidigt Sohn, Tochter den Vater	2
> Eingeschränkte Kommunikation („wenn er doch nichts sagt")	2
> Anscheinende Enttäuschung des Vaters; Grund: Verhalten des Sohnes	1
> Hintergründe dieser Enttäuschung werden nicht erfahrbar	1
> Tochter macht Mutter Vorwürfe wegen ihres Verhaltens / angebliche Sorglosigkeit	2
> Beurteilung Kramers durch die Mutter: • hart / verständnislos • Sohn habe Angst vor seinem Vater	2
> Diskussionen lässt die Mutter nicht zu („basta")	1
> Beurteilung Kramers durch die Tochter: • hart arbeitend • streng, aber gerecht (auch auf sie bezogen) • Rechtfertigung selbst erlittener oder wahrgenommener Gewalt	3
> Hinweis auf den angeblich schlechten Charakter ihres Bruders	1
> Hinreichende Textbelege	3
Rohpunkte insgesamt	3/ 19

Abbildung 37: Detaillierte tabellarische Umsetzung einer Korrektur mit starker inhaltsbezogener Ausrichtung der Bewertung (Vergleichsarbeit - Deutsch)



Aufgabe 3

Denkbare / mögliche Argumentation	Rohpunkte
> Benennung der eigenen Beurteilungssicht	1
> Mutter als auch Tochter verfahren nicht ausschließlich argumentativ	1
> Die Tochter jedoch deutlicher als die Mutter →	1
* Mutter eher emotional gebunden	1
* Ursache: entweder besonders einfühlsam oder „blind"	1
> Mutter kritisiert den Umgang mit dem Vater, berechtigt? →	1
> Tochter: Vater fordere nicht mehr von sich als von den andern Familienmitgliedern	1
> (Abschließende) begründete Beurteilung des Verhaltens der Mutter bzw. der Tochter	4
Rohpunkte insgesamt	2/11

Gesamtbeurteilung:

Aufgabe 1: ...7.....Rohpunkte....05.....Punkte x 30%= 150 14
Aufgabe 2: ...3.....Rohpunkte....00..... Punkte x 55%= 000 8,67
Aufgabe 3:...2.....Rohpunkte....00.....Punkte x 15%= 000 2,72

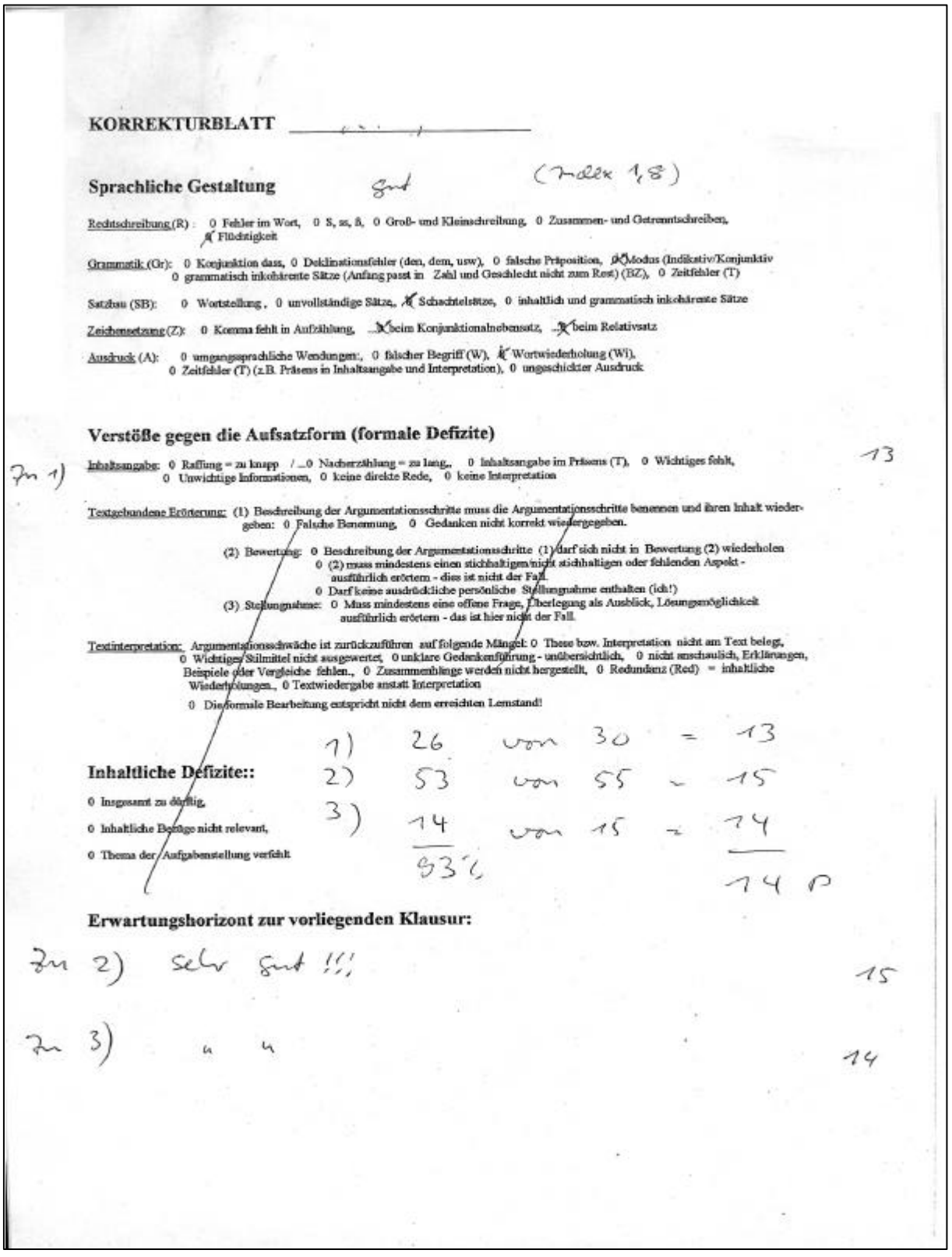
2 24,39
- 2
00P.

Abzug aufgrund sprachlich formaler Mängel:

FI 17,26 = -2P.

01 P.

Abbildung 38: Detaillierte tabellarische Umsetzung einer Korrektur mit starker inhaltsbezogener Ausrichtung der Bewertung (Vergleichsarbeit - Deutsch)



KORREKTURBLATT _____

Sprachliche Gestaltung gut (Index 1,8)

Rechtschreibung (R): 0 Fehler im Wort, 0 S, ss, ß, 0 Groß- und Kleinschreibung, 0 Zusammen- und Getrenntschreiben, 0 Flüchtigkeit

Grammatik (Gr): 0 Konjunktion dass, 0 Deklinationsfehler (den, dem, usw), 0 falsche Präposition, 0 Modus (Indikativ/Konjunktiv 0 grammatisch inkohärente Sätze (Anfang passt in Zahl und Geschlecht nicht zum Rest) (BZ), 0 Zeitfehler (T)

Satzbau (SB): 0 Wortstellung, 0 unvollständige Sätze, 0 Schachtelsätze, 0 inhaltlich und grammatisch inkohärenste Sätze

Zeichensetzung (Z): 0 Komma fehlt in Aufzählung, 0 beim Konjunktionalnebensatz, 0 beim Relativsatz

Ausdruck (A): 0 umgangssprachliche Wendungen:, 0 falscher Begriff (W), 0 Wortwiederholung (Wi), 0 Zeitfehler (T) (z.B. Präsens in Inhaltsangabe und Interpretation), 0 ungeschickter Ausdruck

Verstöße gegen die Aufsatzform (formale Defizite)

zu 1) Inhaltsangabe: 0 Raffung = zu knapp / 0 Nacherzählung = zu lang,, 0 Inhaltsangabe im Präsens (T), 0 Wichtiges fehlt, 0 Unwichtige Informationen, 0 keine direkte Rede, 0 keine Interpretation — 13

Textgebundene Erörterung: (1) Beschreibung der Argumentationsschritte muss die Argumentationsschritte benennen und ihren Inhalt wiedergeben: 0 Falsche Benennung, 0 Gedanken nicht korrekt wiedergegeben.

(2) Bewertung: 0 Beschreibung der Argumentationsschritte (1) darf sich nicht in Bewertung (2) wiederholen 0 (2) muss mindestens einen stichhaltigen/nicht stichhaltigen oder fehlenden Aspekt - ausführlich erörtern - dies ist nicht der Fall. 0 Darf keine ausdrückliche persönliche Stellungnahme enthalten (ich!)

(3) Stellungnahme: 0 Muss mindestens eine offene Frage, Überlegung als Ausblick, Lösungsmöglichkeit ausführlich erörtern - das ist hier nicht der Fall.

Textinterpretation: Argumentationsschwäche ist zurückzuführen auf folgende Mängel: 0 These bzw. Interpretation nicht am Text belegt, 0 Wichtiges Stilmittel nicht ausgewertet, 0 unklare Gedankenführung - unübersichtlich, 0 nicht anschaulich, Erklärungen, Beispiele oder Vergleiche fehlen., 0 Zusammenhänge werden nicht hergestellt, 0 Redundanz (Red) = inhaltliche Wiederholungen., 0 Textwiedergabe anstatt Interpretation

0 Die formale Bearbeitung entspricht nicht dem erreichten Lernstand!

Inhaltliche Defizite::

0 Insgesamt zu dürftig,

0 Inhaltliche Belege nicht relevant,

0 Thema der Aufgabenstellung verfehlt

1) 26 von 30 = 13
2) 53 von 55 = 15
3) 14 von 15 = 14
93 %
14 P

Erwartungshorizont zur vorliegenden Klausur:

zu 2) sehr gut !!! — 15

zu 3) " " — 14

Abbildung 39: Tabellarische Umsetzung einer Korrektur unter stärkerem Einbezug formaler Aspekte („Sprachliche Gestaltung“) (Vergleichsarbeit - Deutsch)

Zu Aufg. 1): Sie haben die formalen Vorgaben der Inhaltsangabe im Großen und Ganzen erfüllt. (Die Angabe der Rahmendaten ist unvollständig)

Der Textinhalt ist zutreffend erfasst, die Wiedergabe aber nicht deutlich genug – zum Teil ist sie zu stark verkürzt, zum Teil ist die sprachliche Darstellung zu ungenau.

Zu Aufg. 2): Sie verdeutlichen an mindestens drei Beispielen, wie der Verfasser der Textvorlage den Leser zu überzeugen versucht; Ihre Analyse ist inhaltlich weitgehend stimmig, die sprachliche Darstellung weist jedoch Schwächen auf, teilweise ist die Analyse dadurch auch undeutlich.

Zu Aufg. 3): Sie nehmen Bezug auf die Textvorlage und formulieren einen eigenen Standpunkt; Sie begründen Ihren Standpunkt aber nur in sehr groben Ansätzen, Ihre Argumentation ist dadurch nicht überzeugend..

Sprachlich-formale Fehler (Fehlerindex 6,5→ Abzug von 01 Punkt)

(ca. 450 W.)

Abbildung 40: Kommentierte Begründung der Bewertung (maschinengeschrieben) (Vergleichsarbeit - Deutsch)

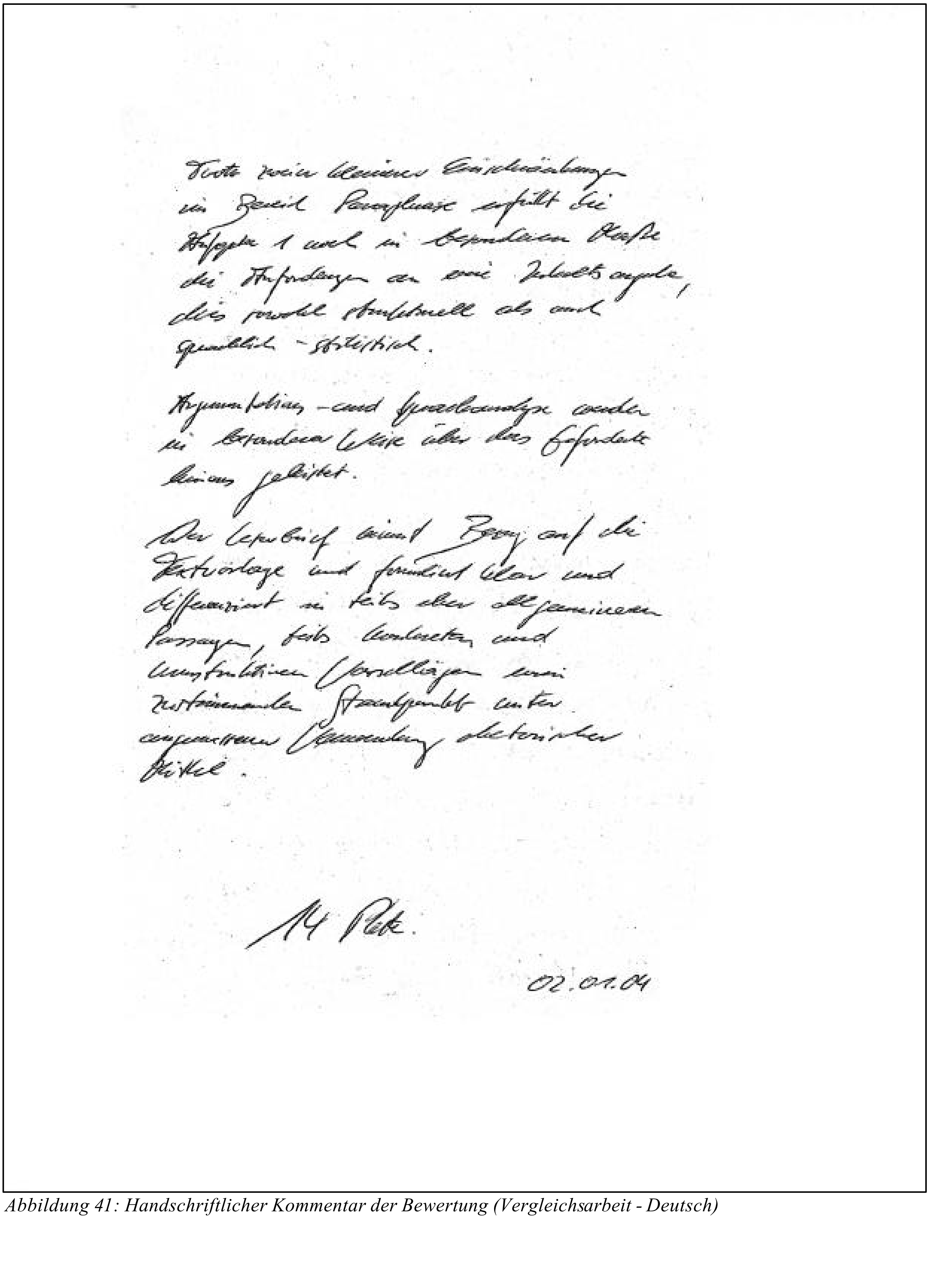
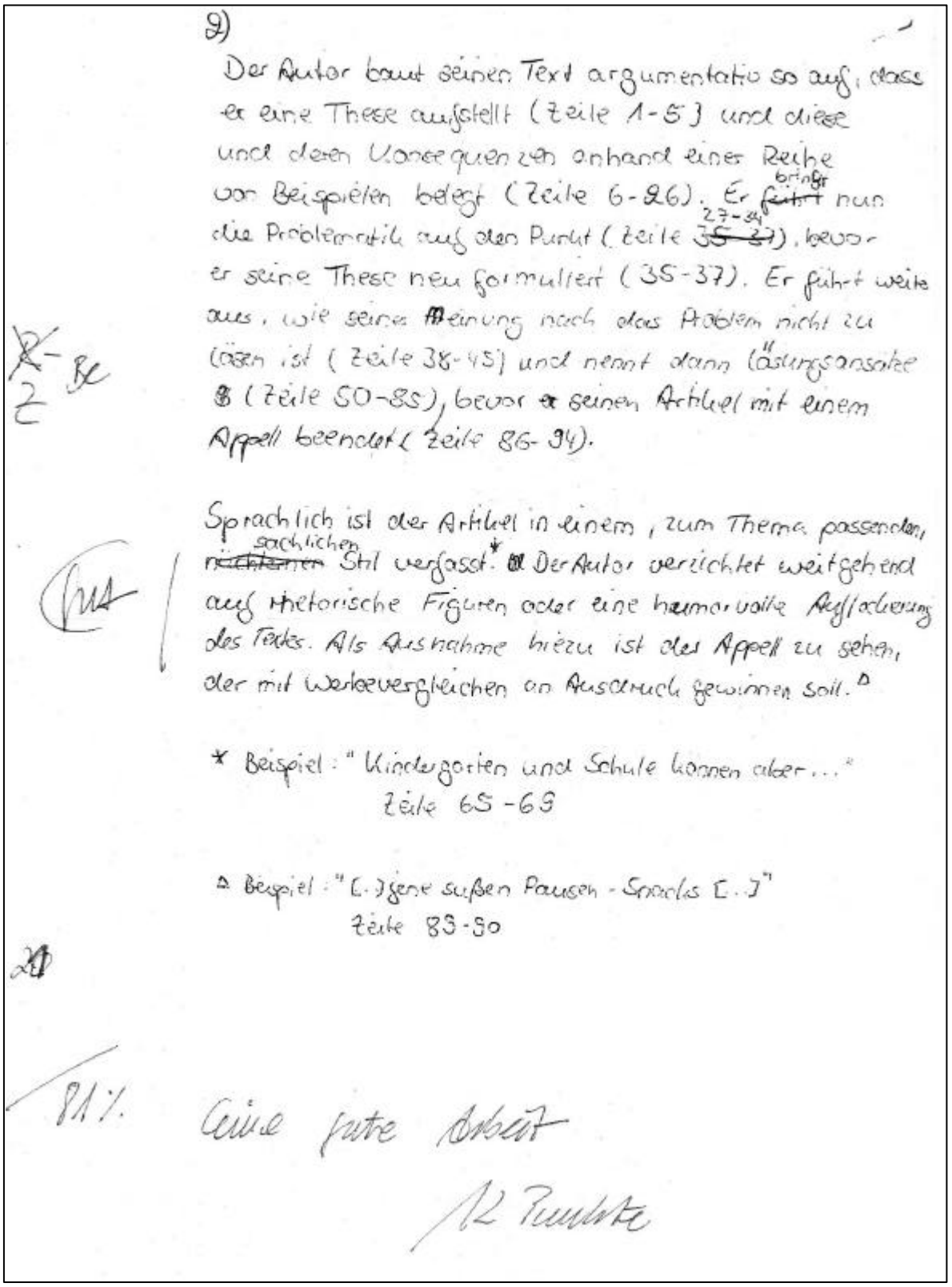


Abbildung 41: Handschriftlicher Kommentar der Bewertung (Vergleichsarbeit - Deutsch)


9)

Der Autor baut seinen Text argumentativ so auf, dass er eine These aufstellt (Zeile 1-5) und diese und deren Konsequenzen anhand einer Reihe von Beispielen belegt (Zeile 6-26). Er führt bringt nun die Problematik auf den Punkt (Zeile 27-34), bevor er seine These neu formuliert (35-37). Er führt weiter aus, wie seiner Meinung nach das Problem nicht zu lösen ist (Zeile 38-49) und nennt dann Lösungsansätze (Zeile 50-85), bevor er seinen Artikel mit einem Appell beendet (Zeile 86-94).

Sprachlich ist der Artikel in einem, zum Thema passenden, ~~nüchternen~~ sachlichen Stil verfasst.* Der Autor verzichtet weitgehend auf rhetorische Figuren oder eine humorvolle Auflockerung des Textes. Als Ausnahme hierzu ist der Appell zu sehen, der mit Vergleichen an Ausdruck gewinnen soll.△

* Beispiel: "Kindergarten und Schule können aber..."
Zeile 65-69

△ Beispiel: "(...) jene süßen Pausen-Snacks (...)"
Zeile 83-90

81%. Eine gute Arbeit

R Pundsack

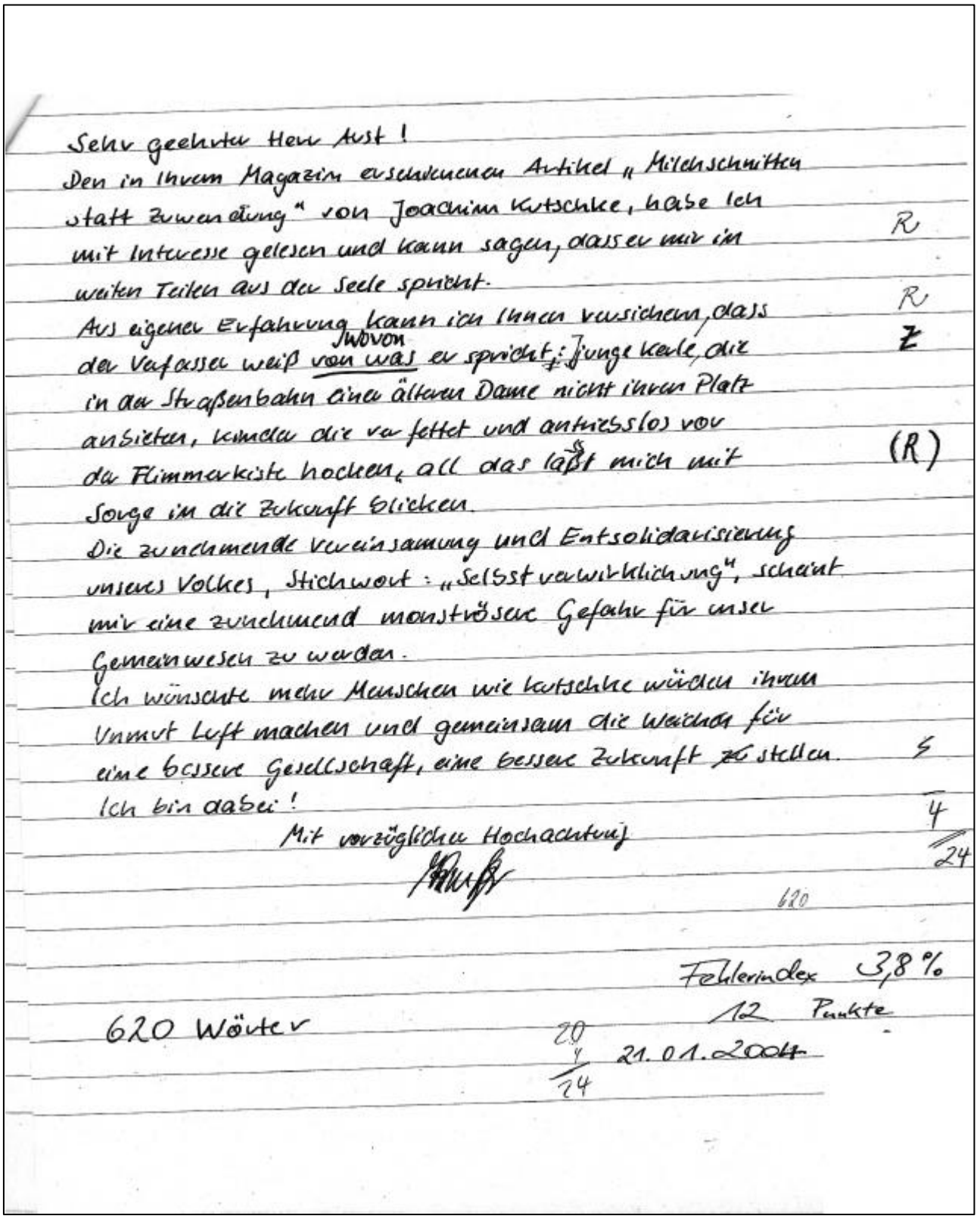

Sehr geehrter Herr Aust!
Den in Ihrem Magazin erschienenen Artikel „Milchschnitten statt Zuwendung" von Joachim Kutschke, habe ich mit Interesse gelesen und kann sagen, dass er mir in weiten Teilen aus der Seele spricht. R
Aus eigener Erfahrung kann ich Ihnen versichern, dass der Verfasser weiß wovon er spricht: Junge Kerle, die in der Straßenbahn einer älteren Dame nicht ihren Platz anbieten, komplett übergewichtig und antriebslos vor der Flimmerkiste hocken, all das läßt mich mit Sorge in die Zukunft blicken. R Z
Die zunehmende Vereinsamung und Entsolidarisierung unseres Volkes, Stichwort: „Selbstverwirklichung", scheint mir eine zunehmend monströse Gefahr für unser Gemeinwesen zu werden. (R)
Ich wünschte mehr Menschen wie Kutschke würden ihrem Unmut Luft machen und gemeinsam die Weichen für eine bessere Gesellschaft, eine bessere Zukunft stellen. S
Ich bin dabei!
Mit vorzüglicher Hochachtung
4/24
620
Fehlerindex 3,8 %
12 Punkte
620 Wörter
20/24
21.01.2004

Abbildung 43: Unkommentierte Bewertung (Vergleichsarbeit - Deutsch)

7 Teststatistik und Professionspraxis – Erfahrungen im Feld

Nachdem im Herbst 2004 die Vergleichsarbeiten der Fächer Englisch und Deutsch einer Auswertung unterzogen worden waren bzw. im Fall des Faches Deutsch eine Nichtauswertbarkeit konstatiert werden musste, wurde eine Ergänzung der Projektfragestellung notwendig:

Es stellte sich die Frage, inwiefern die Formate der Vergleichsarbeiten sowie die Klausurentwürfe und deren Musterlösungen *prinzipiell* den Anforderungen einer teststatistischen Auswertbarkeit entsprechen. Dieser Frage wurde anhand von zwei aktuellen Klausurentwürfen der Fächer Englisch und Deutsch für die Abendhauptschulen (H2) durch die Projektgruppe nachgegangen. Es zeigte sich, dass die Formate, d. h. die Rahmenvoraussetzungen der Klausurentwürfe keinerlei Hindernis sind. Daraufhin wurden die Klausurentwürfe und ihre Musterlösungen dahingehend analysiert, inwiefern sie eine Aufgabendiagnostik mit Hilfe der probabilistischen Testtheorie zulassen. Bei beiden Entwürfen traf dies nur zum Teil zu, wobei die Projektgruppe anhand einiger Veränderungsvorschläge, die nur beispielhaften Charakter hatten, zeigen konnte, dass dies prinzipiell möglich ist (siehe Anhang).

Die Prüfung der Formate und Klausurentwürfe war Teil des Versuchs, die Prinzipien einer statistisch auswertbaren Konstruktion von Klausuraufgaben bzw. Bewertungsprinzipien in den Erstellungsprozess der Vergleichsarbeiten zu implementieren, um so in Zukunft Analysen zu ermöglichen, die dann eine Bearbeitung der Projektfragestellung überhaupt erst erlauben. Dazu wurde im Dezember 2004 von der Projektgruppe ein fünfseitiges Papier erstellt („Die statistische Auswertung von Vergleichsarbeiten – Anforderungen an die Erstellung von Klausuraufgaben“; siehe Anhang I) und den Arbeitsgruppen, welche die Vergleichsarbeiten erstellen, zugänglich gemacht. Von Seiten des Projektes bestand die Hoffnung, so die wenigen grundlegenden Prinzipien der Aufgaben- und Bewertungskonstruktion in die zukünftigen Erstellungsprozesse von Vergleichsarbeiten einfließen lassen zu können.

Die Reaktion auf diesen Versuch, - anhand einiger operativer Vorschläge eine statistische Analyse überhaupt erst zu ermöglichen - war unerwartet heftig und größtenteils ablehnend. Dabei reichten die Begründungen für die ablehnende Haltung von einfachen organisatorischen Einwänden über die Ablehnung spezifischer Auswertungsmethoden bis hin zur grundsätzlichen Infragestellung der Auswertbarkeit an sich. Auch schlichte inhaltliche Missverständnisse waren zu verzeichnen. Eine schriftliche Stellungnahme der Arbeitsgruppen, welche die Vergleichsarbeiten erstellen, sowie eine darauf bezogene Stellungnahme des Projektteams verdeutlichen recht gut die kontroverse Diskussion und werden daher in diesem Bericht als Beispiel für die Diskussion dokumentiert (siehe Anhang II).

Vor dem Hintergrund dieser Ereignisse lässt sich für die Fragestellung der Implementation festhalten, dass der Erstellungsprozess von teststatistisch auswertbaren Vergleichsarbeiten als einer notwendigen Voraussetzung für die Möglichkeit, die Güte von Vergleichsarbeiten im Sinne einer präzisen Leistungsdiagnostik und einer Vergleichbarkeit von Fachleistungen zu analysieren, nicht von Lehrkräften allein geleistet werden kann. Hier ist eine wissenschaftliche Unterstützung in teststatistischer und gegebenenfalls fachdidaktischer Hinsicht bereits im Erstellungsprozess erforderlich.

8 Abbildungsverzeichnis

9 Literatur

ARNOLD, KARL-HEINZ (2001): „Qualitätskriterien für die Messung von Schulleistungen“. In: Weinert, Franz E. (Hrsg.): *Leistungsmessungen in Schulen*. Weinheim, Basel. S.117-130.

BACKHAUS, KLAUS (2000): *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung*. 9. vollständig überarbeitete Auflage. Berlin.

BIRKEL, CLAUDIA; BIRKEL, PETER (2002): „Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss“. In: *Psychologie in Erziehung und Unterricht*. Jahrgang 49, Heft 3. S.219-224.

BORTZ, JÜRGEN (1999): *Statistik für Sozialwissenschaftler*. 5. vollständig überarbeitete Auflage. Berlin.

DITTON, HARTMUT (2001): „Der Leistungsbegriff in empirischen Schulqualitätsstudien“. In: Freitag, Christine; Solzbacher, Claudia (Hrsg.): *Anpassen, verändern, abschaffen? Schulische Leistungsbewertung in der Diskussion.* Bad Heilbrunn. S.39-57.

FISCHER, GERHARD H. (1974): *Einführung in die Theorie psychologischer Tests – Grundlagen und Anwendungen*. Bern.

FREITAG, CHRISTINE (2001): „Die Schulreform in England und ihre Auswirkungen auf die Leistungsbewertung“. In: Freitag, Christine; Solzbacher, Claudia (Hrsg.): *Anpassen, verändern, abschaffen? Schulische Leistungsbewertung in der Diskussion.* Bad Heilbrunn. S.59-75.

FREITAG, CHRISTINE; SOLZBACHER, CLAUDIA (2001): „Anpassen, verändern, abschaffen?“. In: Freitag, Christine; Solzbacher, Claudia (Hrsg.): *Anpassen, verändern, abschaffen? Schulische Leistungsbewertung in der Diskussion.* Bad Heilbrunn. S.7-15

GROFFMANN, KARL-JOSEF; MICHEL, LOTHAR (1982): „Theoretische Grundlagen psychometrischer Tests“. In: Groffmann, Karl-Josef; Michel, Lothar (Hrsg.): *Grundlagen psychologischer Diagnostik*. Göttingen.

HELLER, KURT A., HANY, ERNST A. (2001): „Standardisierte Schulleistungsmessungen“. In: Weinert, Franz E.: *Leistungsmessungen in Schulen*. Weinheim, Basel. S.87-101.

HELMKE, ANDREAS (2004): *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze.

INGENKAMP, KARLHEINZ (1995): *Die Fragwürdigkeit der Zensurengebung*. 9. Auflage. Weinheim.

INGENKAMP, KARLHEINZ (1997): *Lehrbuch der pädagogischen Diagnostik: Texte und Untersuchungsberichte*. 4. Auflage. Weinheim.

JÄGER, REINHOLD S. (1996): *Beurteilung und Beurteilungsprobleme in der Schule*. Landau.

JÜRGENS, EIKO (1998): *Leistung und Beurteilung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht*. 4. erweiterte Auflage. Sankt Augustin.

JÜRGENS, EIKO; SACHER, WERNER (2000): *Leistungserziehung und Leistungsbeurteilung. Schulpädagogische Grundlegung und Anregungen für die Praxis*. Neuwied.

KLAUER, KARL-JOSEF (1978): „Perspektiven pädagogischer Diagnostik“. In: Klauer, Karl-Josef (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Düsseldorf. S.3-14.

KLIEME, ECKHARD (2004): „Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und Befunde“. In: *Zeitschrift für Pädagogik*. Jahrgang 50, Heft 5. S.625-634.

KRÜSSEL, HERRMANN (2001): „Lernen und Leistungsbewertung als Elemente einer konstruktivistischen Lernkultur“. In: Freitag, Christine; Solzbacher, Claudia (Hrsg.): *Anpassen, verändern, abschaffen? Schulische Leistungsbewertung in der Diskussion.* Bad Heilbrunn. S.123-145.

PEEK, RAINER (2001): „Die Bedeutung vergleichender Schulleistungsmessungen für die Qualitätskontrolle und Qualitätsentwicklung von Schulen und Schulsystemen“. In: Weinert, Franz E. (Hrsg.): *Leistungsmessungen in Schulen*. Weinheim, Basel. S.323-335.

REISS, KRISTINA (2004): „Bildungsstandards und die Rolle der Fachdidaktik am Beispiel der Mathematik“. In: *Zeitschrift für Pädagogik.* Jahrgang 50, Heft 5. S.635-649.

ROST, JÜRGEN (2004): „Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen“. In: *Zeitschrift für Pädagogik*. Jahrgang 50, Heft 5. S.662-678.

RHEINBERG, FALCO (2001): „Bezugsnormen und schulische Leistungsbeurteilung“. In: Weinert, Franz E. (Hrsg.): *Leistungsmessungen in Schulen*. Weinheim, Basel. S.59-71.

ROHWER, GÖTZ; PÖTTER, ULRICH (2002): *Methoden sozialwissenschaftlicher Datenkonstruktion*. Weinheim.

ROST, JÜRGEN (2004): *Lehrbuch Testtheorie – Testkonstruktion*. 2. vollständig überarbeitete und erweiterte Auflage. Bern.

ROST, JÜRGEN; SPADA, HANS (1978): „Probabilistische Testtheorie“. In: Klauer, Karl-Josef (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Düsseldorf. S.59-97.

SCHRADER, FRIEDRICH-WILHELM; HELMKE, ANDREAS (2001): „Alltägliche Leistungsbeurteilung durch Lehrer“. In: Weinert, Franz E. (Hrsg.): *Leistungsmessungen in Schulen*. Weinheim, Basel. S.45-58.

STEYER, ROLF; EID, MICHAEL (2001): *Messen und Testen: mit Übungen und Lösungen.* 2. korrigierte Auflage. Berlin.

SÜLLWOLD, FRITZ (1983): „Pädagogische Diagnostik“. In: Groffmann, Karl-Josef; Michel, Lothar (Hrsg.): *Intelligenz- und Leistungsdiagnostik*. Göttingen. S.307-386.

TENORTH, HEINZ-ELMAR (2004): „Bildungsstandards und Kerncurriculum – Systematischer Kontext, bildungstheoretische Probleme“. *Zeitschrift für Pädagogik*. Jahrgang 50, Heft 5, S. 650-661.

WEINERT, FRANZ E. (2001): „Perspektiven der Schulleistungsmessung – mehrperspektivisch betrachtet“. In: Weinert, Franz E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim, Basel. S.270-284.

WEINERT, FRANZ E.(2001): „Vergleichende Leistungsmessung in Schulen – Eine umstrittene Selbstverständlichkeit“. In: Weinert, Franz E. (Hrsg.): *Leistungsmessungen in Schulen*. Weinheim, Basel. S.17-31.

WESTMEYER, HANS (1978): „Grundbegriffe: Diagnose, Prognose, Entscheidung“. In: Klauer, Karl-Josef (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Düsseldorf. S.15-26.

WIECZERKOWSKI, WILHELM; SCHÜMANN, MICHAEL (1978): „Klassische Testtheorie“. In: Klauer, Karl-Josef (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Düsseldorf. S.41-58.

Anhang I:

„Die statistische Auswertung von Vergleichsarbeiten – Anforderungen an die Erstellung von Klausuraufgaben“

(inkl. einer Anwendung auf zwei Beispiele: Klausurentwürfe der Vergleichsarbeiten in den Fächern Deutsch und Englisch der Abendhauptschule im WS 2004/05)

Verbundprojekt
„Steuerung von Schulen des Zweiten Bildungswegs (Schulen für Erwachsene) in Hessen“

Die statistische Auswertung von Vergleichsarbeiten – Anforderungen an die Erstellung von Klausuraufgaben

An den hessischen Schulen für Erwachsene sind im Dezember 2003 Vergleichsarbeiten in den Fächern Mathematik, Deutsch und Englisch geschrieben worden, die z.Z. im Rahmen des Projektes „Steuerung von Schulen des Zweiten Bildungswegs (Schulen für Erwachsene) in Hessen“ (kurz: Projekt Bildungssteuerung) ausgewertet werden. Ziel dieser Auswertung ist es u.a., ein Verfahren zu erarbeiten, mit dem die SfE selbständig wissenschaftlich fundierte Aufgaben bzw. Aufgabenformate für Vergleichsarbeiten (weiter)entwickeln können. Das Ziel eines solchen Verfahrens ist die vergleichsorientierte Rückmeldung von Lernständen und Lernbedürfnissen, die sich an definierten Kompetenzen bzw. Kompetenzstufen orientiert. Ein solches Verfahren setzt voraus, dass spezifische statistische Anforderungen bei der Erstellung der Aufgaben und Bewertungsvorlagen beachtet werden.

1) Hinweise zur Logik des statistischen Auswertungsverfahrens

Das Ziel von Vergleichsarbeiten besteht darin, die Fachleistungen von Studierenden unabhängig von der einzelnen Schule, der jeweiligen Klasse und auch unabhängig vom Unterrichtenden selbst vergleichbar zu machen. In statistischer Hinsicht sollen die Studierenden durch Vergleichsarbeiten entsprechend ihrer individuellen Fähigkeiten in einem Unterrichtsfach unterschieden werden können.

Das Grundprinzip des hier vorgestellten Verfahrens besteht darin, Aufgaben nach ihrem Schwierigkeitsgrad zu unterscheiden und dann in einer Vergleichsarbeit Aufgaben verschiedener Schwierigkeitsgrade zum Einsatz zu bringen. Die Einschätzung des Schwierigkeitsgrades richtet sich danach, wie groß der Anteil der Studierenden ist, der eine jeweilige Aufgabe lösen kann. Dies unterstellt, dass einfache Aufgaben von vielen Studierenden, mittelschwere Aufgaben von einem Teil der Studierenden und schwere Aufgaben nur von wenigen Studierenden gelöst werden.

Auf diese Weise werden z.B. auch

- missverständliche,
- zu leichte oder
- zu schwere

Aufgaben identifiziert. Missverständliche/ unverständliche Aufgaben werden von allen Studierenden gleich gut oder schlecht gelöst, unabhängig von ihrer Kompetenz, da die Lösung der Aufgabe erraten wird. Die Häufigkeit der korrekten Lösung schwankt somit in statistischer Hinsicht zufällig innerhalb der untersuchten Gruppe. Wird eine Aufgabe von allen Studierenden gelöst, so lässt diese Aufgabe keine Aussage über Leistungsunterschiede zu. Sie macht daher keinen Sinn im Rahmen einer vergleichenden Perspektive. Dasselbe gilt umgekehrt für eine zu schwere Aufgabe, wenn sie von niemandem gelöst werden kann. Eine zu schwere Aufgabe kann nur helfen, die Leistung der Studierenden nach oben hin abzuschätzen. In allen drei hier genannten Fällen informie-

ren solche Aufgaben nicht über das zu untersuchende Merkmal Fachkompetenz, d.h. sie lassen Unterschiede in der Fachkompetenz nicht deutlich werden.

Die Anzahl der gelösten Aufgaben drückt im Rahmen der Logik der PISA-Untersuchung die Kompetenz der Studierenden aus (genauer gesagt: sie ist ein Indikator dafür – mehr nicht). Umgekehrt wird mit der Anzahl der Personen, die eine Aufgabe lösen können, der Schwierigkeitsgrad der jeweiligen Aufgabe zum Ausdruck gebracht. Die Verschränkung dieser beiden Gesichtspunkte ist für die Leistungsmessung entscheidend. Aus ihr erwächst die Einschätzung der Verwendbarkeit von Aufgabenstellungen: Aufgaben, die von einigen der fähigen Studierenden nicht gelöst werden, von anderen fähigen Studierenden dagegen gelöst werden, darf es demnach nicht geben. Solche Aufgaben werden ausgeschieden, da sie die fragliche Kompetenz nicht messen, sondern eine andere Fähigkeit. Wendet man diese formale Struktur der PISA-Aufgabenbewertung an, dann können Aufgaben der Vergleichsarbeiten unter dem Aspekt ihrer Lösungshäufigkeit erprobt, bewertet und archiviert werden.

Die Kompetenz in einem Unterrichtsfach (z.B. „Englischkompetenz“) setzt sich in der Regel aus verschiedenen Teilkompetenzen zusammen, die im Rahmen von Vergleichsarbeiten abgefragt werden (z.B. Grammatik, sprachlicher Ausdruck etc.). Vergleichsarbeiten haben das Ziel, die „Gesamtkompetenz“ zu erfassen, womit die Unterstellung einher geht, dass die Teilkompetenzen insgesamt ähnlich stark ausgeprägt sind. Wenn eine Vergleichsarbeit diese Teilkompetenzen diagnostisch sinnvoll abfragt, dann sollte es dementsprechend enge, statistisch erfassbare, Zusammenhänge geben. [1]

2) Konkrete Anforderungen an die Konstruktion von Aufgaben und Bewertungskriterien bei Vergleichsarbeiten unter Berücksichtigung der PISA-Logik:

Punktzahlen für Teilaufgaben und die Konstruktion von Bewertungskriterien

Statistisch-formale Voraussetzung zur Identifikation des Schwierigkeitsgrads von Teilaufgaben ist es, dass innerhalb jeder Aufgabe die Teilaufgaben mit jeweils der gleichen Punktzahl bewertet werden.

Diese Anforderung lässt sich z.B. realisieren, indem die Bewertung der Teilaufgaben in genügend kleine Korrekturschritte unterteilt wird. Im kleinschrittigsten Fall wird bei der Korrektur nur noch überprüft, ob eine geforderte Anforderung vorhanden/erfüllt ist oder nicht. So wäre z.B. bei einer Lückentextaufgabe zur Überprüfung einer Grammatikregel im Englischen die erste Anforderung, dass die Studierenden die Grammatikregel richtig anwenden und die zweite Anforderung, dass das Wort selbst richtig geschrieben wird. Würde für die Erfüllung der Anforderung Grammatikregel und Rechtschreibung jeweils ein Punkt vergeben, so ergäbe sich als Bewertung für das korrekte ausfüllen einer Lücke eine Gesamtpunktzahl von 2, als Addition der Bewertung der beiden Anforderungen.[2]

---

[1] Der Zusammenhang von Gesamtkompetenz und Note kann durch Verfahren wie z.B. der Regressionsrechnung oder der Faktorenanalyse überprüft werden. Dabei wird im Fall der Faktorenanalyse überprüft, ob es einen gemeinsamen Faktor gibt, der den verschiedenen Aufgaben/Teilen einer Vergleichsarbeit zugrunde liegt.

[2] Soll die Richtigkeit der Rechtschreibung keinen Einfluss auf die Bewertung der Grammatikfähigkeiten haben, so wäre es an dieser Stelle möglich nur die Grammatikpunkte zu betrachten und die Punktzahlen für die Rechtschreibung außen vor zu lassen. Die Rechtschreibung könnte z.B. als gesonderte Kompetenz zwischen den Aufgaben verglichen werden.

Es sind aber auch Bewertungskategorien möglich, die einen größeren Bewertungsspielraum zulassen, z.B. „voll erfüllt“, „erfüllt“, „nicht erfüllt“ bzw. analoge Bewertungen in Punkten. Dabei ist jedoch zu beachten, dass die Bewertungskategorien von Aufgaben mit erweitertem Bewertungsspielraum so definiert werden, dass die Korrektur möglichst eindeutig wird, d.h., dass es aufgrund der exakten Definition von Anforderungen für eine Bewertungskategorie keine Missverständlichkeiten gibt. Für eine statistische Auswertung ist es auch hier notwendig, dass die Anzahl von Bewertungskategorien für Teilaufgaben einer Aufgabe bei allen Teilaufgaben identisch sein muss. Außerdem muss definiert werden, ob eine Bewertung mit Teilpunkten (halbe Punkte oder ggfls. sogar Viertelpunkte) möglich ist.

An dieser Stelle sei noch einmal darauf hingewiesen, dass die o.g. Anforderungen nicht bedeuten, dass möglichst kleinschrittige Aufgaben entwickelt werden müssen, sondern dass die beschriebene Kleinschrittgkeit sich auf die Bewertungskriterien und die korrespondierende Punktvergabe, d.h. die Korrektur der Vergleichsarbeiten bezieht.

Skalierung von Gesamtpunktzahlen für Aufgaben

Wenn die Aufgaben insgesamt eine einheitliche Gesamtkompetenz überprüfen sollen, dann gilt die obige Forderung analog auch für die Aufgaben selbst. Die maximal erzielbaren Punkte für die Aufgaben der Vergleichsarbeiten müssen für alle Aufgaben gleich sein. Eine unterschiedliche Skalierung der Punktzahlen pro Aufgabe ist in Hinsicht auf eine statistische Auswertung problematisch:

Tabelle 1: Probleme bei unterschiedlichen Skalen bzw. unterschiedlichen Gesamtpunktzahlen

Arbeitsaufgabe 1 (es sind 5 Punkte erreichbar): Punktzahlen, die von Studierenden erreicht werde können	Durch Dehnung der Skala (hier z.B. Multiplikation mit Faktor 2) werden die Punkte einer Arbeitsaufgabe derjenigen Arbeitsaufgabe mit doppelter Punktzahl angeglichen	Arbeitsaufgabe 2 (es sind 10 Punkte erreichbar): Punktzahlen, die von Studierenden erreicht werden können	Punktzahlen, die durch die unterschiedliche Skalierung von Studierenden bei der Arbeitsaufgabe 1 nicht erreicht werden können:
1		1	1
2		2	
3		3	3
4		4	
5		5	5
		6	
		7	7
		8	
		9	9
		10	

Bei einer Stauchung der Gesamtpunktzahl auf die kleinste erreichbare Punktzahl gehen Informationen verloren, denn gleiche Punktzahlen werden zu einer Punktzahl zusammengefasst. Eine Streckung der Gesamtpunktzahlen auf die Punktzahl einer Aufgabe, die die höchste Gesamtpunktzahl hat, erzeugt unbesetzte Kategorien, die eine korrekte Auswertung behindern: Wenn z.B. die mögliche Punktbewertung einer Aufgabe von ein bis fünf Punkten (nur ganze Punkte) auf die Punktewertung von eins bis zehn gestreckt wird, so werden die Kategorien „eins“, „drei“, „fünf“, „sieben“

und „neun“ nicht besetzt sein. Es gibt dann also keine Studierenden, die diese Kategorien in dieser Aufgabe belegen. Bei anderen Aufgaben hingegen gibt es Studierende in diesen Kategorien (vgl. Tabelle 1).

Somit scheidet sowohl eine Streckung als auch eine Stauchung unterschiedlicher Punkteskalen als Anpassungsverfahren zwischen den Aufgaben aus.

Gewichtung

Eine Gewichtung von Aufgaben braucht nicht zu erfolgen, da aufgrund des unterschiedlichen Schwierigkeitsgrads der Aufgaben diese „gewichtet“ sind. Über die Anzahl der gelösten Aufgaben ergibt sich die Differenzierung zwischen den Aufgaben. Das selbe gilt auch für die Teilaufgaben. Auch die Teilaufgaben sollen eine Schwierigkeitsstrukturierung leisten, nur eben innerhalb der Aufgabe.

Wenn allerdings eine Gewichtung von Aufgaben zur Notengebung benötigt wird, so kann dennoch eine Gewichtung vorgenommen werden, sofern diese erst nach Abschluss aller Korrekturen bzw. der Punktvergabe erfolgt (s.u. das Beispiel Abschlussprüfung Deutsch AHS Winter 2004/05).

Freie Textteile

Besondere Anforderungen ergeben sich bei der Bewertung freier Textteile. Gerade hier ist es notwendig, möglichst kleinschrittige Bewertungskriterien festzulegen. Sinnvoll ist ein erarbeiteter Bewertungskatalog, der soweit wie möglich alle Anforderungen und (Fehler-)Eventualitäten berücksichtigt. Dazu gehört z.B. die Bewertung inhaltlicher Aspekte, sprachlicher Fähigkeiten, Rechtschreibung etc. Die Erstellung dieser Bewertungskataloge kann immer nur annähernd sein, da unmöglich alle möglichen Bewertungsprobleme vorausgesehen werden können (und die Kreativität der Studierenden bei der Erstellung von Fehlern letztlich dem Bewertungskatalog immer einen Schritt voraus sein dürfte).

Folgeaufgaben

Die Lösung verschiedener Aufgaben sollte möglichst nicht voneinander abhängen. Durch die Abhängigkeit der Aufgaben voneinander wird ein Schwierigkeitsgrad von Folgeaufgaben erzwungen, der nicht zutreffend sein muss.

Tabelle 2: Beispiel für **eine mögliche** Anordnung von Bewertungskriterien und Punktzahlen

	Alle Arbeitsaufgaben insgesamt gleiche Punktzahl	Alle Teilaufgaben insgesamt gleiche Punktzahl	Skalierung z.B.: „Ja/Nein“ oder „voll erfüllt“ bis „nicht erfüllt“
Teilkompetenz 1 erzielbare Pkt: 16	Arbeitsaufgabe 1 erzielbare Pkt: 8	Teilaufgabe 1 erzielbare Pkt: 2	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig
		Teilaufgabe 2 erzielbare Pkt: 2	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig
		Teilaufgabe 3 erzielbare Pkt: 2	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig
		Teilaufgabe 4 erzielbare Pkt: 2	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig
	Arbeitsaufgabe 2 erzielbare Pkt: 8	Teilaufgabe 1 erzielbare Pkt: 4	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig Korrekturschritt 3, 1 Pkt, wenn richtig Korrekturschritt 4, 1 Pkt, wenn richtig
		Teilaufgabe 2 erzielbare Pkt: 4	Korrekturschritt 1, 1 Pkt, wenn richtig Korrekturschritt 2, 1 Pkt, wenn richtig Korrekturschritt 3, 1 Pkt, wenn richtig Korrekturschritt 4, 1 Pkt, wenn richtig
Teilkompetenz 2 erzielbare Pkt: 16	Arbeitsaufgabe 3 erzielbare Pkt: 16	Teilaufgabe 1 erzielbare Pkt: 16	Korrekturschritt 1, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 2, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 3, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 4, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 5, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 6, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 7, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig Korrekturschritt 8, 2 Pkt, wenn richtig, 1 Pkt, wenn teilweise richtig

3) Anwendung der statistischen Anforderungen auf zwei konkrete Beispiele

Die Umsetzung der o.g. statistischen Anforderungen auf die Korrektur von Vergleichsarbeiten soll im folgenden anhand der Abschlussprüfungen für die AHS in den Fächern Englisch und Deutsch im WS 2004/05 verdeutlicht werden. In beiden Fällen zeigt sich, dass die entwickelten Korrekturvorgaben der Abschlussprüfungen schon weitgehend den Anforderungen entsprechen und nur geringe Modifikationen notwendig sind, um die Anforderungen in Gänze zu erfüllen. Im folgenden werden mögliche Varianten einer Modifikation beispielhaft vorgeschlagen, wobei entsprechend mehr Varianten denkbar sind, als hier angeführt werden.

Die Musterklausuren befinden sich im Anhang.

3a) Mögliche Modifikationen der Bewertungskriterien bzw. der Punktvergabe bei der Abschlussarbeit Englisch im Bildungsgang Abendhauptschule (WS 2004/05)

Betrachtung von Teil II:

Teil II dient der Überprüfung der Grammatikfähigkeiten und des Sprachschatzes der Studierenden. Dabei behandeln die Aufgaben II a) bis II c) grammatikalische Probleme, Aufgabe II d) überprüft den Wortschatz der Studierenden. Es stellt sich daher die Frage, ob die Aufgabe II d) überhaupt mit den Aufgaben II a) bis II c) verglichen werden soll. Dies ist jedoch eine fachdidaktische Frage.

Auf der kleinschrittigsten Korrekturebene weisen die Aufgaben II a) bis II d) jeweils für sich betrachtet einheitliche Bewertungen der Unteraufgaben aus, z.B. II c.1) bis II c.6) je ein Punkt bei richtiger Beantwortung der Unteraufgabe. Dadurch erfüllen diese Unteraufgaben per se die Bedingungen der statistischen Auswertbarkeit im Sinne der PISA-Logik.

Die Aufgaben innerhalb von Teil II weisen allerdings unterschiedliche maximale Punktzahlen auf:

II a) 10 Punkte, II b) 5 Punkte, II c) 6 Punkte, II d) 5 Punkte.

Für eine Anpassung dieser Punktzahlen bietet sich zunächst an, die Aufgabe II c) zu kürzen, indem eine Unteraufgabe gestrichen wird. Da für jede Unteraufgabe maximal ein Punkt vergeben wird, würde sich dann auch für II c) eine maximale Punktzahl von 5 Punkten ergeben.

Aufgabe II a) könnte in zwei gleichberechtigte Teile von je 5 Punkten umgewidmet werden: 5 Punkte bei richtiger Beantwortung aller Unteraufgaben bzgl. Zeit und Grammatik und 5 Punkte bei richtiger Beantwortung aller Unteraufgaben bzgl. Fragewort und Wortstellung.

Zu lösen ist nun noch das Problem, dass bei der Bewertung von II a) „Fragewort und word order“ halbe Punkte als Bewertung zugelassen sind, also auch der Wert von z.B. 4,5 bei dieser Aufgabe erreicht werden kann, ebenso wie bei II d), wohingegen bei den anderen Aufgaben nur ganze Werte entstehen. Eine mögliche Lösung wäre es, ebenfalls auch halbe Punkte bei der Korrektur der anderen Aufgaben zuzulassen oder bei Aufgabe II a) Fragewort und word order als zusammengehörig zu interpretieren und auch dort nur ganze Punkte zuzulassen. Dies ist jedoch wieder eine fachdidaktische Entscheidung. Überträgt man diese Regel allerdings auf II d) so ergibt sich dort eine zu hohe maximal erreichbare Punktzahl, d.h. II d) müsste entsprechend gekürzt werden.

Betrachtung von Teil I und III:

In Teil I und III wird die Bewertung von Inhaltsaspekten und sprachlicher Kompetenz unterschieden. In Teil I b) sollen die Studierenden zu sechs inhaltlichen Aspekten, in I c) zu sieben und in III) zu zehn inhaltlichen Aspekten Stellung nehmen. Bewertet werden die Aspekte jeweils mit einem Punkt. Es ergeben sich somit unterschiedliche maximal zu erreichende Punktzahlen. Eine Egalisierung der maximal erreichbaren Punktzahlen kann z.B. durch eine Verlängerung der Aufgaben I b) und I c) auf zehn zu untersuchende Aspekte erfolgen oder z.B. durch Streichung von Inhaltsaspekten auf eine für alle Aufgaben gemeinsame Größe.

Bei der Überprüfung der sprachlichen Kompetenz stellt sich das gleiche Problem. In Teil I b) sind nur 6 Punkte für die Sprachkompetenz zu erzielen in Teil III) zehn. Würde Aufgabe I b) auf zehn inhaltliche Aspekte erweitert so würden sich dort auch maximal zehn erreichbare Punkte für die sprachliche Kompetenz ergeben. Wird die Verkürzung der Aufgaben als adäquatere Lösung zur Anpassung der maximal erzielbaren Punktzahl angesehen, so müsste die Bewertungsskala der sprachlichen Bewertung in Teil III) entsprechend angepasst werden.

Da die Unteraufgaben bei allen Aufgaben jeweils mit einem Punkt bewertet werden, sind die Unteraufgaben der jeweiligen Aufgaben innerhalb der Aufgabe auf ihre Eignung hin statistisch überprüfbar.

Will man die Teile I bis III der selbst miteinander vergleichen, dann müssen die maximal erzielbaren Punkte für alle drei Teile gleich sein. Ungeachtet fachdidaktischer Überlegungen wäre eine mögliche Lösung die folgende: Ausgehend von den maximal zu erzielenden 20 Punkten in Teil III könnte I b) auf zehn inhaltliche Aspekte erweitert werden und dafür Aufgabe I c) gestrichen werden. Somit ergäben sich für Teil I auch maximal 20 Punkte. In Teil II müsste II c) um eine Unteraufgabe gekürzt werden und eine der Aufgaben II b) bis d) wegfallen.

Das Problem der Bewertung von Rechtschreibfehlern wird umgangen indem leichte Fehler nicht bewertet werden. Es wäre jedoch ggfls. zu präzisieren, wann ein Wort als erkennbar gilt.

3b) Mögliche Modifikationen der Bewertungskriterien bzw. der Punktvergabe bei der Abschlussarbeit Deutsch im Bildungsgang Abendhauptschule (WS 2004/05)

Fehlende Kriterien/ Skala

Wie aus der Musterarbeit ersichtlich ist, wurden für die Aufgaben 1) und 2) die Kriterien zur Bewertung der inhaltlichen Leistungen der Studierenden definiert. Es fehlt allerdings eine Skala, anhand derer bewertet werden kann, in welchem Umfang die inhaltlichen Leistungen erbracht wurden, z.B. „voll erfüllt", „teilweise erfüllt", „gar nicht erfüllt" o.ä.

Problematisch sind Doppelanforderungen wie „finden eine zum Hauptteil passende Einleitung und einen Schluss" anstatt hier nach der passenden Einleitung und dem passenden Schluss zu differenzieren.

Bei Aufgabe 3) fehlt eine Explikation der Bewertungskriterien und einer Skala.

Richtung der Skalierung/ Fehlerindex

Die Verstöße gegen die sprachliche Richtigkeit werden kleinschrittig definiert. Der Wert der sprachlichen Richtigkeit wird über einen Fehlerindex errechnet, d.h. je höher der Wert desto schlechter ist das Ergebnis. Bei der Bewertung der inhaltlichen Aspekte ist dies i.d.R. genau umgekehrt, wenn z.B. die numerische Kodierung der Skala „nicht erfüllt" bis „voll erfüllt" mit 0, 1, 2 erfolgen würde, d.h. je höher der Wert, desto besser das Ergebnis. Diese unterschiedliche Orientierung der Bewertungen kann berücksichtigt werden, indem ein maximaler Fehlerindex definiert ist, ab dessen Höhe es fachdidaktisch keinen Sinn mehr macht, schlechte Leistungen zu differenzieren. Die Invertierung dieser Bewertungsskala würde dann zum gewünschten Ergebnis der Gleichorientierung der Skalen führen.

Mögliche Gewichtung der Anforderungen (Anforderungsebene)

Soll nun die sprachliche Richtigkeit mit der inhaltlichen Leistung und der äußeren Gestaltung verglichen werden, so müsste in diesen Anforderungen jeweils die gleiche maximale Punktzahl erreichbar sein. Eine Gewichtung dieser Anforderungen, wie in der Musterarbeit definiert, bliebe hiervon unberücksichtigt. Wenn z.B. für die sprachliche Richtigkeit, die inhaltliche Anforderung und die formale Gestaltung jeweils 30 Punkte maximal erzielbar sind und die Gewichtung der Anforderungen für die Zusammensetzung der Note für die sprachliche Richtigkeit 30% betragen würde, für die inhaltliche Anforderung 60% und für die formale Gestaltung 10%, so ergäbe sich für die Notenvergabe folgendes Bild: die für die Notenvergabe relevante maximale Punktezahl wäre dann 30 Punkte * 0,3 + 30 Punkte *0,6 + 30 Punkte * 0,1 = 28 Punkte.

Vergleichbarkeit von Aufgaben (Aufgabenebene)

Eine Vergleichbarkeit der Aufgaben selbst ist zu erzielen, wenn in diesen die jeweiligen maximal erzielbaren Punkte für den Inhalt, die Sprachrichtigkeit und die formale Gestaltung jeweils identisch sind. Für die Sprachrichtigkeit ergeben sich keine Probleme, da ein Fehlerindex definiert ist. Dieser müsste nur für jede Aufgabe separat berechnet werden. Eine inhaltliche „Gleichbewertung" der Aufgaben würde erreicht, wenn die Anzahl der geforderten Kriterien gleich wäre und eine Be-

wertung der Erfüllung oder Nichterfüllung dieses Aspekts wie oben beschrieben skaliert würde. Die Bewertung für die formale Gestaltung könnte gesetzt werden. Eine für die Notengebung vorzunehmende Gewichtung wie oben beschrieben ist auch in diesem Fall unkritisch für die Auswertung.

Abschlussarbeit Englisch

WS 2004/2005

Bildungsgang Abendhauptschule

Themenbereich: family and friends

(Prototyp)

I) Reading comprehension

a) Read the text - Lesen Sie den Text

This is our normal day
I'm Jacky Brown. I'm a music teacher. My husband, Tony, is an actor (Schauspieler). This is our normal day. I get up at seven o'clock and make tea. My husband and I have tea in bed. Then I have a shower and have my breakfast. Tony and our daughter, Sonia, have their breakfast later. I leave home at eight; I start work at half past. Sonia's school starts at nine o'clock and Tony takes her to school. He goes to the theatre in the afternoon.
I usually meet Sonia at four o'clock and get home at half past four. After dinner we often watch TV or play games together. I prepare a snack for Tony because he isn't back from work before midnight.We have a glass of red wine together and talk about the day, then we go to bed.

(145 words)

b) Answer the questions in full sentences - Beantworten Sie die Fragen in vollständigen Sätzen.

1. What is Tony's job?

_____

2. When does Sonia's school start?

_____

3. How does Sonia get to school?

_____

4. Why does Jacky prepare a snack for Tony?

_____

5. Where does Tony work?

_____

6. Who is at home at eight thirty a.m.?

_____

c) Tick '*true', 'false'* or *'not in the text'* - Kreuzen Sie '*richtig', 'falsch'* oder *nicht im Text* an.

Statement	true	false	not in the text
Tony has to work every morning.			
He normally comes home after twelve p.m.			
Sonia is Tony's sister.			
Tony works at the local opera.			
Jacky works at a high school.			
Jacky and Tony are happily married.			
Tony has got a driving licence.			

II) Grammar and language

a) Ask for the underlined part of the sentences - Fragen Sie nach den unterstrichenen Satzteilen

1._____?

Jacky's mother lives <u>in Brighton</u>

2._____?

She is <u>60 years</u> old.

3. _____?

She works in a hospital <u>because she is a doctor.</u>

4. _____?

They like <u>ice-cream.</u>

5. _____?

I get up <u>at seven in the morning.</u>

b) Use the personal pronoun - Verwenden Sie die Personalpronomen

1. Tony works at a theatre -_____ is an actor.
2. Tony and Jacky are married -_____ have got one child.
3. Sonia needs a new book. - _____ costs 10 Euro.
4. Jacky says to Tony: “Hurry up, _____ are late!”
5. Sonia goes to school. -_____ is a good student.

c) Complete these sentences. Use the same verb in the second part. - Vervollständigen Sie die Sätze. Verwenden Sie das gleiche Verb im zweiten Teil des Satzes.

1. She was at school yesterday.	They _____ at home.
2. They are 40.	She _____ 25.
3. I will go to America next year.	She _____ to Spain next month.
4. We are writing a test at the moment.	The teacher _____ a letter.
5. I have got a bike.	He _____ a car.
6. They live in Brighton.	She _____ in London.

d) Translate the words - Übersetzen Sie diese Wörter

1. Vater =_____
2. Mutter =_____
3. Tochter =_____
4. Sohn =_____
5. Tante =_____
6. Ehemann =_____
7. Onkel =_____
8. Bruder =_____
9. Ehefrau =_____
10. Schwester =_____

III. Text production.

In the text you read about Jacky Brown's normal day. But yesterday everything was different. Jacky visited her mother in Brighton. Before she left she told Tony what he and Sonia should do. This is the plan Jacky gave Tony:
(In dem Text haben Sie über Jacky Browns normalen Tag gelesen. Aber gestern war alles anders. Jacky besuchte ihre Mutter in Brighton. Bevor sie abfuhr, sagte sie Tony, was er und Sonia tun sollten. Dies ist der Plan, den Jacky ihnen gab:)

- *8.00 Uhr morgens aufstehen*
- *ein Bad nehmen*
- *Sonia nicht zur Schule fahren.*
- *Sonia nimmt den Bus*
- *Abendessen kochen*
- *Sonia macht ihre Hausaufgaben*
- *sie hat keine Zeit zum Spielen*
- *Fernsehen schauen bis neun Uhr*
- *Jacky um neun Uhr anrufen*
- *um 10 Uhr ins Bett gehen*

**Now write down what Tony and Sonia did yesterday. Write about 70 words-
Schreiben Sie, was Tony und Sonia gestern taten. Schreiben Sie ca. 70 Wörter.**

Start like this:
Yesterday morning Tony....

Hinweise zur Bewertung

Kein Punktabzug bei leichten Rechtschreibfehlern. (Wort muss erkennbar sein.)

zu I)

b) pro Antwort 2 Punkte
Inhalt -1 P.
Sprache - 1P.
=12 P.

c) pro richtiges Kreuz 1 Punkt
= 7 P.

zu II)

a) pro Frage 2 Punkte
Fragewort und word order - je ½ P.
Zeit und Grammatik - 1 P.
= 10 P.

b) pro Personalpronomen 1 Punkt
= 5 P.

c) pro Verbform 1 Punkt
= 6 P.

d) 1/2 Punkt pro Vokabel
= 5 P.

zu III)

Inhalt: 1 Punkt pro vorgegebenem Aspekt
= 10 P

Sprache:

9/10 P. - abwechslungsreicher Satzbau, idiomatische Ausdrucksweise, grammatisch korrekt
7/8 P. - abwechslungsreich im Ausdruck, Grammatisch weitgehend korrekt
5/6 P. - schlicht, aber angemessen im Ausdruck, keine schwerwiegenden grammatischen Fehler
3/4 P. - eingeschränkter Ausdruck, etliche grammatische Fehler, die die Verständlichkeit insgesamt kaum beeinträchtigen
1/2 P. - sehr eingeschränkter Ausdruck, Fehler beeinträchtigen die Verständlichkeit deutlich
0 P. - Ausdruck, Satzbau und Formenbildung so schwach, dass keine Verständlichkeit gegeben ist

= 20 P.

Gesamtpunktzahl: 65 Punkte

% / Punkte	65 - 57	56 - 48	47 - 39	38 - 30	29 - 13	12 - 0
Note	1	2	3	4	5	6

Musterarbeit für die schriftliche Abschlussarbeit Deutsch an Abendhauptschulen WS 2004/5

Bildergeschichte

Aufgabenvorschlag:

Bildvorlage: E.O. Plauen

Aufgabenstellungen:

1. Gestalten Sie zu den vorgegebenen Bildern einen Text! Verfassen Sie eine Einleitung, einen Hauptteil und einen Schluss und finden Sie eine treffende Überschrift! (60%)
2. Beschreiben Sie die Beziehung des Sohnes zu seinem Vater! (15%)
3. Sicherlich sind auch andere Verhaltensweisen des Sohnes in dieser Situation denkbar. Beschreiben Sie diese! (15%)

Gestalten Sie Ihre Arbeit formal korrekt und optisch ansprechend (10%)

Erwartete Leistungen

I. Inhaltlich

Zu 1:

Die Studierenden
- erkennen die Kernaussage der Bildergeschichte
- geben die bildliche Darstellung sachlich korrekt wieder
- gestalten die Geschichte anschaulich und kreativ aus
- stellen die Geschichte zusammenhängend dar
- finden eine treffende Überschrift
- finden eine zum Hauptteil passende Einleitung und einen Schluss

Musterlösung

Der rettende Einfall

Mike hatte für die Mathearbeit nichts gelernt, weil so schönes Wetter war und er lieber mit seinen Freunden ins Schwimmbad gegangen war. Er hatte, schon bevor der Lehrer die Arbeit zurückgab, ein flaues Gefühl im Magen. Jetzt stand der Lehrer mit strafendem Blick vor ihm, er hatte wirklich eine Sechs geschrieben. Mike schlich nach Hause, weil er Angst hatte seinem Vater die Arbeit zu zeigen, aber das musste er, denn eine Sechs musste immer unterschrieben werden. Er überlegte und überlegte. Da hatte er einen rettenden Einfall. Zu Hause angekommen, setzte er sich im Wohnzimmer auf den Boden, verband sich die Augen und tat so, als ob er Unterschriften-Schreiben mit verbundenen Augen üben wollte. Sein Vater schaute ihm eine Weile zu. „Kannst du das auch?“, fragte Mike seinen Vater. Der verband sich die Augen, legte sich auf den Boden und schrieb mehrfach seinen Namen. Ohne dass der Vater es bemerkte, schob Mike, der inzwischen das Tuch abgenommen hatte, sein Matheheft dazwischen und der Vater schrieb seine Unterschrift ins Heft. Jetzt nahm der Vater das Tuch ab und überprüfte die Unterschriften. Mike packte schnell und unbemerkt das Heft in die Mappe.
Am nächsten Tag zeigte er mit klopfendem Herzen dem Lehrer die Unterschrift, auch der merkte nichts. Das war ja noch einmal gut gegangen. Aber ein schlechtes Gewissen hatte Mike schon, dass er seinen Vater so ausgetrickst hatte. Für die nächste Arbeit, nahm er sich vor, wollte er ganz bestimmt lernen.

Zu 2.:

Die Studierenden
- beschreiben die Beziehung Vater-Sohn zutreffend (z.B. Angst vor Strafe/ keine Offenheit/ kein Vertrauen)
- erläutern bzw. begründen ihre Beschreibung

Musterlösung

Die Beziehung des Sohnes zu seinem Vater, die man aus dieser Bildergeschichte interpretieren kann, ist von Angst des Sohnes gegenüber dem Vater geprägt. Der Sohn traut sich nicht aus Angst vor einer Strafe seinem Vater eine schlechte Note zu präsentieren. Wahrscheinlich hat er schon schlechte Erfahrungen gemacht und scheint nicht genug Vertrauen zu ihm zu haben. Deshalb entscheidet er sich den Vater lieber mit einer List zu „betrügen" als die Wahrheit zu sagen.

Zu 3.:

Hier gibt es verschiedene Möglichkeiten, die von den Studierenden genannt werden können. Naheliegend ist der Vorschlag, dass der Sohn seinem Vater die Wahrheit sagt und sich dem Unmut des Vaters aussetzt. Er kennt die Reaktion seines Vaters, weiß aber aus Erfahrung, dass der Vater sich wieder beruhigen wird. Der Sohn muss also aufgrund seiner Ehrlichkeit kein schlechtes Gewissen haben.

Eine weitere Möglichkeit wäre, dass der Junge zu seinem Lehrer geht, ihn bittet auf die Unterschrift des Vaters unter die schlechte Arbeit zu verzichten und ihm verspricht, das nächste Mal für die Klausur zu lernen.

II. Sprachlich

Die Studierenden
- beachten die Regeln der Grammatik (z.B. Satzbau, Zeitform, direkte und indirekte Rede)
- verwenden treffende Adjektive und Verben, abwechslungsreiche Satzanfänge
- vermeiden Wortwiederholungen
- beachten die Regeln der Rechtschreibung und Zeichensetzung

III. Äußere Gestaltung

Die Studierenden gestalten ihre Arbeit formal korrekt und optisch ansprechend.

Anforderungen für die Note ausreichend.

Insgesamt müssen von 100% mindestens 46% erreicht sein. Dies ist der Fall, wenn Folgendes geleistet wird.

Bei 1.: Die Überschrift, die Einleitung, der Hauptteil und der Schluss passen ansatzweise zusammen und ergeben einen „roten Faden".
Der Hauptteil lässt erkennen, dass die Bildfolge im Wesentlichen verstanden und nachvollziehbar erzählt wird.
Eine Zeitebene wird weitgehend eingehalten.
Bei 2: Das Vater-Sohn Verhältnis wurde ansatzweise plausibel beschrieben.
Bei 3: Eine Alternative wurde ansatzweise formuliert.

Allgemein:

Fehlerindex

Abzug bei gehäuften Verstößen gegen die Sprachrichtigkeit:
Die Leistung im Bereich der sprachlichen Richtigkeit (Rechtschreibung, Zeichensetzung und Grammatik) sind integraler Bestandteil der Arbeiten im Deutschunterricht. In den Vergleichsarbeiten erfolgt daher

Der Fehlerindex errechnet sich nach der Formel $\frac{\text{Fehler x 100}}{\text{Wortzahl}}$

Folgende Tabelle gibt Ihnen Aufschluss über den Fehlerindex:

0 – 1,5	1%
1,6 – 2,9	2%
3,0 – 4,5	3%
4,6 – 6,0	4%
6,1 – 7,9	5%
8,0 – 9,5	6%
9,5 – 10,9	7%
11,0 – 12,5	8%
13,6 – 14,9	9%
15,0 – 16,0	10%

Anhang II:

Brief der Arbeitsgruppen „Vergleichsarbeiten“ sowie die darauf bezogene Stellungnahme des Projektteams (Synopse vom 31.1.05)

Synopse Stellungnahmen Vergleichsarbeiten (Stand 31.1.05)

Die Arbeitsgruppen für die Abschlussprüfung Abendhauptschule Englisch, Deutsch, Mathematik die Abschlussprüfung Abendrealschule Englisch, Deutsch, Mathematik die zentralen Vergleichsarbeiten in Q2 in Englisch, Deutsch, Mathematik	**Christoph Fuhrmann/ Klaus Harney/ Sascha Koch (31.1.05):** **Stellungnahme zum Schreiben der Arbeitsgruppen für die Abschlussprüfungen und zentralen Vergleichsarbeiten vom 14.01.2005** *Unsere Hinweise beziehen sich auf den Abschnitt 2 der Arbeitsgruppen-Kritik, da der erste Abschnitt außerhalb unseres Auftrags- und Einflussbereichs liegt. Ausgangspunkt unserer Überlegungen sind die Antwortformate der uns vorgelegten Vergleichsarbeiten. Deren Antwortformate sind unserer Auffassung nach im Hinblick auf die leistungsbezogene Vergleichsabsicht verbesserungsfähig. Die Leistung der Arbeitsgruppen bei der Aufgabenfindung stellen wir nicht in Abrede – ganz im Gegenteil. Unsere Hinweise beziehen und bezogen sich auf die Weiterentwicklung der Korrekturpraxis. Sie entsprechen der Praxis der Korrektur der Vergleichsarbeiten in NRW .*
[…] **2. Anforderungen an die Erstellung von Klausuraufgaben nach Prof. Harney** Unzweifelhaft ist nach unserer Auffassung: Ein Test **ist** nur dann valide, wenn er die Unterrichtsrealität abbildet, er **wird** valide (gemacht), wenn die Testanforderungen den Unterricht entsprechend normieren. Letzteres wird durch Herrn Min'R Hochstätter mithilfe des Verbundprojekts „Steuerung von Schulen des Zweiten Bildungswegs (Schulen für Erwachsene) in Hessen" umzusetzen versucht. Die folgende Kritik an dem Papier *„Die statistische Auswertung von Vergleichsarbeiten – Anforderungen an die Erstellung von Klausuraufgaben*" (Anlage) betrifft die Arbeitsgruppen zur Erstellung von Abschlussarbeiten in AHS und ARS und von zentralen Vergleichsarbeiten in Q 2 zwar graduell unterschiedlich, sie gilt jedoch für alle.	Hier wird u.E. ein eher unspezifischer Begriff von Validität benutzt. Validität kann nur auf ein Merkmal bezogen werden, welches getestet wird. Im Fall von zentralen Vergleichsarbeiten werden Studierendenkompetenzen getestet. Aussagen über die Unterrichtsrealität oder ähnliche Faktoren sind ohne die Erhebung weiterer empirischer Gegebenheiten höchstens indirekt möglich.

- Es wird behauptet, der Schwierigkeitsgrad einer jeweiligen Aufgabe ergebe sich aus der Anzahl der Personen, die sie lösen können. Daraus wird gefolgert, dass es Aufgaben, die „von einigen der fähigen Studierenden nicht gelöst werden, von anderen fähigen Studierenden dagegen gelöst werden, nicht geben darf." Es sollen damit Unterschiede in der Fachkompetenz deutlich werden. Diese Anforderung unterstellt zum einen, dass Studierende in Teilkompetenzen gleich stark bzw. schwach seien. Das ist aber keineswegs der Fall. So können sogenannte Nullanfänger in Englisch (z. B. Russlanddeutsche) sehr stark im Erkennen isolierter Grammatikphänomene, aber sehr schwach im Produzieren von Texten sein. Andere Studierende wiederum können gut Texte produzieren, sind in den Grammatikteilen jedoch schwach, wiederum andere sind sowohl in Textproduktion wie auch in Grammatik stark (schwach). Auch in Q 2 ist die Bewältigung der Aufgabentypen Zusammenfassung, Analyse und eigenständige, aber abgeleitete Stellungnahme keineswegs auf die unterstellte Gleichheit „fähiger Studierender" zu reduzieren. Diese Anforderung unterstellt zum anderen, dass Tests unter normorientierten Gesichtspunkten erstellt werden. **Prüfungsaufgaben** werden jedoch nach den vorgegebenen Lernzielen in den **Lehrplänen** und **FAPAS** erstellt, also kriteriumsorientiert.	Zum ersten Absatz: Die vorgeschlagene Logik des Rasch-Modells geht davon aus, dass schwierige Aufgaben nur von fähigen Studierenden gelöst werden. Faktisch wäre es auch möglich, dass z.B. durch Raten, Mogeln etc. im konkreten Einzelfall diese Annahme nicht zutrifft. Dies ist jedoch in der Logik des Rasch-Verfahrens vorgesehen. Solche Einzelfälle sind in statistischer Hinsicht zufällig und werden daher über die wahrscheinlichkeitstheoretische Bearbeitung aufgefangen. Zum zweiten Absatz: Hier handelt es sich um ein Missverständnis. Unser Verfahrensvorschlag bezieht sich darauf, dass Studierende nicht in <u>einer</u> Teilkompetenz zugleich stark bzw. schwach sein können. In mehreren Teilkompetenzen können sie das selbstverständlich. Das angeführte Beispiel (Nullanfänger) verweist auf solche unterschiedlich ausgeprägten Teilkompetenzen, um deren angemessene Erfassung es ja gerade geht. Zum dritten Absatz: Auch wir gehen davon aus, dass die Inhalte der Vergleichsarbeiten über die Lehrpläne bestimmt werden. Unsere Verfahrensvorschläge beziehen sich allein auf die Konstruktion von Aufgaben und Bewertungskriterien, nicht auf Inhalte (s. Beispiele in unserem Arbeitspapier).
- Mit diesem Schwierigkeitsgrad korrespondiert die Forderung nach gleicher Punktzahl innerhalb der Teilaufgaben, aber auch die nach einer Gleichgewichtung der Aufgaben selbst. Ob eine gleiche Punktzahl innerhalb der Teilaufgaben angemessen ist, darf nicht eine Entscheidung der Testanforderung sein, sondern ist vielmehr eine bewusste pädagogische Entscheidung des Lehrers. Manche Teilaufgaben sind weniger wichtig. In der traditionellen Aufgabenstellung der Q- Phase ist die Forderung nach gleicher Gewichtung der Aufgaben besonders fragwürdig und rechtswidrig. Die **FAPAS** fordern, bezogen auf die drei **Anforderungsbereiche**, einen klaren Schwerpunkt der zu erbringenden Prüfungsleistungen im **Anforderungsbereich II**. Eine gleiche Punktzahl für die jeweiligen Teilaufgaben ist deshalb entweder unzulässig oder aber die Aufgaben werden in ein Schema gepresst, das ihnen von der Anforderungsstruktur her nicht gemäß ist.	Für das Verfahren ist es notwendig, dass die pro (Teil-)Aufgaben zu erzielenden Punkte jeweils gleich sind. (Dies ist nur bei solchen Aufgaben notwendig, die miteinander verglichen werden sollen.) Davon unberührt ist eine mögliche zusätzliche Gewichtung von Aufgaben zwecks Notenermittlung: Sind (Teil-)Aufgaben in fachdidaktischer Hinsicht unterschiedlich relevant, ist ihre entsprechend unterschiedliche Gewichtung problemlos möglich. Der Vorschlag lautet, zunächst jeweils gleiche Punktzahlen für die Erfüllung von jeweiligen (Teil-)Aufgaben zu vergeben. Dies ist die Voraussetzung für eine mögliche Gewichtung *im Rahmen des Korrigierens bzw. der Notenerstellung*.

- Aus der Anforderung nach gleicher Punktzahl wird die nach genügend kleinen Korrekturschritten abgeleitet. Zwar wird diese eingeforderte Eindeutigkeit der Korrektur im Schlussteil (freie Textteile) wieder etwas relativiert, es bleibt jedoch die Forderung bestehen, „möglichst kleinschrittige Bewertungskriterien festzulegen“ und „einen Bewertungskatalog zu erarbeiten“, der „möglichst alle Anforderungen und (Fehler)Eventualitäten berücksichtigt.“ Da die Autoren selbst zugeben, dass „die Erstellung dieser Bewertungskataloge immer nur annähernd sein (kann)“, ziehen wir im Sinne der Eindeutigkeit der Formulierung und der Aussage die Schlussfolgerung, dass man von dieser Anforderung tunlichst Abstand nehmen sollte. Im Fach Deutsch z.B. wäre ein solcher Bewertungskatalog zudem geradezu absurd.	Der Kern einer Vergleichbarkeit von Studierendenleistung ist abgesehen von der Nutzung identischer Aufgaben die Bewertung dieser Aufgaben nach einheitlichen Kriterien. Wenn verschiedene Lehrkräfte Aufgaben bewerten, dann stellt die Explikation sowie die Kleinschrittigkeit, d.h. Differenziertheit der Bewertung die Grundlage einer Vergleichbarkeit her. Der Arbeitspraxis von Lehrkräften unterliegt – nicht nur bei Klausuren – grundsätzlich eine mehr oder weniger explizierte und differenzierte Bewertung von Studierendenleistungen. Im Falle von zentralen Vergleichsarbeiten muss eine Explikation und Kleinschrittigkeit so weit wie möglich vorhanden sein, um eine Bewertung personenindifferent zu machen, d.h. unabhängig von Lehrkraft *und* Studierendem. Wir stimmen zu, dass die Erstellung von Bewertungskriterien je nach Aufgabentyp unterschiedlich schwierig ist. Bewertungskriterien für freie Textteile sind zweifelsfrei anspruchsvoller in ihrer Erstellung als solche anderer Aufgabentypen. Wie das o.g. Beispiel der Klausur für das Fach Deutsch zeigt, ist dies jedoch zweifelsohne möglich. Die kleinschrittige und explizite Nennung von Bewertungskriterien ist bereits teilweise Praxis in den zentralen Vergleichsarbeiten der SfE (siehe das Beispiel der zentralen Abschlussklausur WS 04/05 im Fach Deutsch der Abendhauptschule, das in unserem Arbeitspapier kommentiert wird). Wir weisen nur darauf hin, dass diese Kleinschrittigkeit und Explikation der Bewertungskriterien systematisch vorgenommen werden muss.

- Dass die „Anforderungen“ sich über **FAPAS** und **Lehrpläne** hinwegsetzen, ja sie schlicht ignorieren, wird besonders in der abschließenden Forderung deutlich, nach der die Lösungen verschiedener Aufgaben nicht voneinander abhängen sollten. Die Zusammenfassung der zentralen Aussagen eines Textes, ihre Analyse und Beurteilung **müssen** aufeinander bezogen sein. Das gilt besonders für die Q-Phase und das Abitur, aber auch für die Klausuren in der Abendhaupt- und Abendrealschule, in denen die Aufgaben als textbasierte sich möglichst auf den oder die Ausgangstexte beziehen sollen. Dieser Bezug macht geradezu die Qualität der Aufgabenstellung aus!	Hier besteht u.E. ein Missverständnis darüber, was die genannte Forderung meint, dass die Lösungen verschiedener Aufgaben nicht voneinander abhängig sein dürfen. Dies lässt sich an dem Beispiel der Arbeitsgruppe erläutern, bei dem die Sachlogik (Anforderung des internen Bezugs der Aufgaben) mit der Bewertungslogik (jede Aufgabe wird für sich allein gewertet) verwechselt wird: Angenommen die drei Aufgaben „1) Textzusammenfassung“, „2) Analyse“ und „3) Beurteilung“ beziehen sich auf einen (einzigen) Text. Sofern hier ein interner Bezug hergestellt werden **muss** (z.B. laut Lehrplan), handelt es sich um gleich mehrere, konkurrierende implizite Kompetenzmodelle, beispielsweise: a) Textzusammenfassung ist sehr schwer, Analyse mittelmäßig schwer, eine Beurteilung einfach b) Textzusammenfassung ist einfach, Analyse schwerer, Beurteilung am schwersten, c) der Schwierigkeitsgrad der Aufgaben steht in keinem Zusammenhang, d) alle drei Aufgaben haben denselben Schwierigkeitsgrad e) … Wenn nun jede Aufgabe für sich bewertet würde, dann könnte die vorgeschlagene teststatistische Auswertung empirisch überprüfen, ob bzw. welches der denkbaren impliziten Kompetenzmodelle sich in der Leistung der Studierenden widerspiegelt. Weiterhin: Sofern hier ein interner Bezug hergestellt werden **muss** (z.B. laut Lehrplan), ist dies ein *Bewertungskriterium*, das abgeprüft werden muss, d.h. jeder Teil müsste daraufhin bewertet werden, ob Bezüge zu den beiden anderen Teil hergestellt wurden (s.o.). In diesem Fall könnte eine teststatistische Auswertung z.B. herausarbeiten, bei welchen Aufgaben dies schwerer fällt und bei welchen leichter. Das von uns angesprochene Problem, dass Lösungen nicht aufeinander aufbauen dürfen, ist in dem o.g. Beispiel nicht virulent.

- In den „Anforderungen“ heißt es, dass nach dem Vorbild der PISA-Aufgaben Kompetenzen und Kompetenzstufen getestet werden sollen. Um dieses Ziel erfüllen zu können, bedarf es aber eines **theoretisch fundierten Kompetenzmodells**. Dieses ist für die PISA-Aufgaben in jahrelanger Arbeit von Testspezialisten und Fachdidaktikern erarbeitet worden. **Unsere Lehrpläne enthalten jedoch keine Kompetenzmodelle**, und wir verfügen weder über das Know-how noch die Zeit, derartige wissenschaftliche Modelle entwickeln zu können.	Kompetenzmodelle sind u.E. in jeder unterrichtlichen Aufgabenstellung existent - ob explizit oder implizit. Gleiches gilt für Lehrpläne. Der Auffassung der Arbeitsgruppe, Lehrpläne beruhten nicht auf Kompetenzmodellen, können wir uns nicht anschließen: Lehrpläne enthalten implizite Kompetenzmodelle, die in ihre Konstruktion eingehen. Unsere Vorschläge stellen darauf ab, die der Lehr- und Klausurpraxis eigene Verschränkung von Kompetenzzurechnung, Aufgabenstellung und Leistungsbeurteilung transparent und analysierbar zu machen. Für die Erstellung von Kompetenzmodellen als einem pragmatischen und ständig empirisch zu testenden Bezugspunkt ist eine Verschränkung von Testspezialisten und Fachdidaktikern zweifellos sinnvoll, wobei hier eine professionsbezogene ebenso wie eine universitätsdisziplinäre Fachdidaktik angesprochen ist.