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	**Klaus Harney/ Christoph Fuhrmann** **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.