Grotjahn, R. (1989):

Der C-Test im Bundeswettbewerb Fremdsprachen

Thomas Finkenstaedt
Konrad Schröder
(Hrsg.)

ZWISCHEN EMPIRIE UND MACHBARKEIT

Erstes Symposium zum Bundeswettbewerb Fremdsprachen

Kloster Walberberg,
11. bis 13. April 1989

I&I

Augsburger I & I - Schriften
Herausgegeben von
Thomas Finkenstaedt und Konrad Schröder
Band 50

ISBN 3-923549-32-6

Im Auftrag der Universität herausgegeben
von Thomas Finkenstaedt und Konrad Schröder

Universität Augsburg
Universitätsstraße 10
8900 Augsburg

Herstellung arco-druck gmbh, Hallstadt

Inhalt

Für einen solchen Landeskunde-Test müßte allerdings noch eine geeignete Form gefunden werden. Sie zu finden, dürfte genauso schwer sein wie die Zusammenstellung geeigneter Themen. Ein Expertenteam sollte diese Frage erörtern.

1) Vgl. Klauer, Karl Josef: "Kontentvalidität". *Diagnostica* 30 (1984), Heft 1 : 1.

2) Vgl. *Einladung und erste Information zum Wettbewerb* 1986: 2.

3) Vgl. Bildung und Begabung e. V. (ed.): *Der Bundeswettbewerb Fremdsprachen.* Bonn - Bad Godesberg 1988: 10.

4) Vgl. Schröder, Konrad/Stütz, Wolfgang (eds.): *Der Bundeswettbewerb Fremdsprachen.* Berlin: Cornelsen 1988: IX.

5) Einladung und erste Information zum Wettbewerb 1986: 13.

6) Vgl. Klein-Braley, Christine/Raatz, Ulrich (eds.): *C-Tests in der Praxis.* Bochum: Ruhr-Universität Bochum 1985. (= AKS-Rundbrief 13/14)

7) Vgl. Doyé, Peter: *Typologie der Testaufgaben für den Englischunterricht.* München: Langenscheidt-Longman 1986.

Rüdiger Grotjahn

Der C-Test im Bundeswettbewerb Fremdsprachen: Eignung und Probleme (Französisch)

1. *Einleitung*

Die im Bundeswettbewerb Fremdsprachen bisher eingesetzten französischen C-Tests bestehen jeweils aus vier kurzen, in aufsteigender Schwierigkeit geordneten Texten. Ab dem zweiten Satz wird bei jedem zweiten Wort die zweite Hälfte getilgt. Die Wettbewerbsteilnehmer haben dann die getilgten Worthälften zu rekonstruieren. Die Zahl der Tilgungen pro Text beträgt 20. Die französischen C-Tests sind damit kürzer als z. B. die englischen C-Tests, die aus vier Texten mit jeweils 25 Lücken bestehen (zur Konstruktion und Evaluation von C-Tests vgl. z. B. Grotjahn 1987). Ein Beispiel für einen französischen C-Test-Text findet sich in Abschnitt 6.2 (siehe *Abbildung 1*).

Betrachtet man die statistischen Analysen der bisherigen Testdurchgänge im Bundeswettbewerb Fremdsprachen, erweist sich der französische C-Test als hochreliabel. Die Reliabilitäten (geschätzt mit Cronbachs α) lagen in der Regel um .9 (1986 sogar bei .95) und damit über den entsprechenden Werten für die übrigen Wettbewerbssprachen (vgl. auch den Beitrag von Raatz und Klein-Braley in diesem Bande). Wenn man berücksichtigt, daß es sich bei den im Bundeswettbewerb Fremdsprachen eingesetzten C-Tests um für den jeweiligen Wettbewerbsdurchgang neu entwickelte und nur ansatzweise vorerprobte Tests handelt, sind die erhaltenen Werte äußerst beeindruckend. Problematisch ist vor allem - und dies gilt im übrigen auch in bezug auf z. B. die englischen C-Tests - , daß bisher alle französischen C-Tests bezogen auf die Zielsetzung des Wettbewerbs zu leicht waren (die Schwierigkeiten lagen in der Regel zwischen 0.6 und 0.7) und meist auch nicht eindimensional gemessen haben. Auf Konsequenzen aus diesem Sachverhalt werde ich weiter unten eingehen.

Auch wenn insgesamt gesehen der (französische) C-Test sich im Rahmen des Bundeswettbewerbs bisher erstaunlich gut bewährt

hat, so gibt es neben den bereits erwähnten Problemen noch eine Reihe weiterer, grundsätzlicher Probleme. Als erstes soll nun - und zwar relativ kurz - auf folgende Fragen eingegangen werden:

(1) Nach welchen Kriterien sollten C-Test-Texte für den Bundeswettbewerb ausgewählt werden? Aus wieviel Texten sollte ein C-Test bestehen?
(2) Wie sollte der C-Test eingerichtet werden, d. h. welches Konstruktionsprinzip sollte angewendet werden?
(3) Welche Bedeutung hat die C-Test-Instruktion?
(4) Inwieweit ist es nötig, die für jeden Testjahrgang neu zusammengestellte C-Test-Version vorzutesten und statistisch zu analysieren?

Im Anschluß an die Diskussion der vier angeführten Problembereiche möchte ich etwas ausführlicher auf das Problem der Korrektur von C-Tests eingehen und auch einige empirische Befunde zu dieser, insbesondere unter Praktikabilitätsgesichtspunkten wichtigen Frage präsentieren. Die folgenden Ausführungen beziehen sich zwar schwerpunktmäßig auf den Bereich 'Französisch', vieles gilt jedoch auch für die anderen Wettbewerbssprachen.

2. *Textauswahl*

Das Procedere bei der Textauswahl im Bereich Französisch war meist folgendermaßen: Ich habe ca. 10 C-Test-Texte in zwei oder drei 10. Klassen Bochumer Gymnasien vorgetestet. Aufgrund der Ergebnisse habe ich dem zuständigen Ausschuß sechs bis sieben Texte vorgelegt. Der Ausschuß hat dann aus diesen Texten vier Texte für die endgültige Testversion ausgewählt. Bei der Erstellung und statistischen Analyse der C-Tests hat eine Absprache zwischen den Erstellern der verschiedenen C-Tests für die einzelnen Wettbewerbssprachen in bezug auf Eignung und Probleme der jeweiligen Testversion nur ansatzweise und lediglich aufgrund persönlicher Initiativen stattgefunden. Deshalb sind auch die Erfahrungen, die ein bestimmter C-Test-Ersteller mit seinen Versionen gemacht hat, bisher nur sehr bedingt in die Arbeit der übrigen C-Test-Ersteller eingeflossen.

Bei der Textauswahl war u. a. auf folgendes zu achten:

(1) *Sind die Texte sprachlich dem Lernstand der Testpopulation angemessen?*
Dies bedeutet, daß die Texte hinsichtlich Grammatik und Lexik möglichst weitgehend den bis zum Testzeitpunkt im Unterricht vermittelten Inhalten entsprechen sollten. So sollten die Texte z. B. nur dann Konjunktive enthalten, wenn dieses Phänomen im Unterricht auch (mehrheitlich) behandelt worden ist. Auf die Schwierigkeiten, dieser Forderung in einem bundesweiten Test zu entsprechen, sei lediglich hingewiesen.

(2) *Sind die Texte altersadäquat?*
Dies bedeutet u. a., daß hochabstrakte Texte nicht für die Testpopulation des Bundeswettbewerbs in Frage kommen.

(3) *Ist die aus der Textauswahl resultierende Testversion im Schwierigkeitsgrad optimal in bezug auf die Hauptzielsetzung des Bundeswettbewerbs, nämlich die möglichst verläßliche Auswahl der besten Wettbewerbsteilnehmer?*
Punkt 3 impliziert, daß die Texte insgesamt nicht zu leicht sein dürfen, da sie dann nicht mehr genügend im oberen Leistungsbereich differenzieren.

Betrachtet man die bisherige Praxis, kann festgehalten werden, daß ich dem Ausschuß stets C-Test-Versionen vorgelegt habe, die einen m. E. vertretbaren Kompromiß hinsichtlich der Kriterien 'Lernstand', 'Altersadäquatheit' und 'Differenzierungsfähigkeit im oberen Leistungsbereich' darstellten. Der Ausschuß hat jedoch offensichtlich vor allem die Kriterien 'Lernstand' und 'Altersadäquatheit' sowie allgemeine motivationale Kriterien - schwierige Texte könnten demotivieren - bei der endgültigen Auswahl zugrunde gelegt. Als Folge sind alle bisherigen französischen C-Test-Versionen für eine optimale Differenzierung im oberen Leistungsbereich zu leicht. Dies gilt übrigens noch deutlicher für das Englische (vgl. den Beitrag von Raatz und Klein-Braley). Angesichts dieser Tatsache ist m. E. ein dringendes Desiderat zu klären, welche Ziele genau mit dem Einsatz der C-Tests verfolgt werden sollen und wie die C-Tests diesen Zielen optimal angepaßt werden können.

Auf jeden Fall möchte ich dafür plädieren, zukünftige C-Test-Versionen zumindest um einen fünften, sehr schwierigen Text zu erweitern. Auf diese Weise ließe sich die Diskriminationsfähigkeit im oberen Leistungsspektrum verbessern. Aus Ökonomiegründen könnte man vermutlich dann darauf verzichten, den ersten, sehr leichten Text überhaupt auszuwerten. Dieser Text hätte dann lediglich eine Aufwärmfunktion. Wie die bisherigen Testdurchläufe zeigen, trägt Text 1 in der Regel auch nur sehr wenig oder überhaupt nicht zur Gesamtreliabilität bei. So hat z. B. beim Wettbewerbsdurchgang 1986 die Nichtberücksichtigung von Text 1 keine Auswirkung auf die Cronbachs α, während bei Nichtberücksichtigung eines der übrigen Texte α von .95 auf .92 fällt.

3. *Konstruktionsprinzip*

Betrachtet man die bisher erstellen C-Tests in den verschiedenen Wettbewerbssprachen, fällt auf, daß es nicht unerhebliche Unterschiede im zugrundegelegten Konstruktionsprinzip gibt. Explizite und zumindest ansatzweise wissenschaftlich abgesicherte Begründungen für die Unterschiede im Vorgehen liegen meines Wissens nicht vor. So gibt es z. B. Diskrepanzen in der Behandlung von einsilbigen Wörtern und von Eigennamen bei der Tilgung. Außerdem gibt es C-Test-Versionen, in denen nicht nur Teile von Wörtern, sondern zusätzlich auch ganze Wörter getilgt worden sind. Weiterhin wird in einer Reihe von C-Tests jeder getilgte Buchstabe durch einen einzelnen Punkt symbolisiert. Das zuletzt geschilderte Vorgehen ist m. E. nicht unproblematisch, da es hierdurch - zumindest in einem gewissen Umfang - zu einer im Rahmen einer fremdsprachlichen Aufgabenstellung unerwünschten Buchstabenzählerei auf Seiten der Schüler kommt.

4. *Die Instruktion*

In einer Vielzahl kognitionspsychologischer Arbeiten ist nachgewiesen worden, daß das Verhalten der Probanden bei Konstanz der eigentlichen Problemlösungsaufgabe in entscheidender Weise von der jeweiligen Instruktion abhängen kann. In Anbetracht dieser Tatsache bedürfen m. E. die vorhandenen Unterschiede in den C-Test-Instruktionen in den verschiedenen Wettbewerbssprachen einer expliziten Begründung.

So ist z. B. der Hinweis, daß die *Hälfte* eines jeden zweiten Wortes fehlt, nicht unproblematisch, da dies zu der bereits erwähnten Buchstabenzählerei führen kann. Belege, daß dies tatsächlich der Fall sein kann, finden sich in einigen an der Ruhr-Universität Bochum aufgenommen Protokollen des "lauten Denkens" beim C-Test-Lösen (vgl. hierzu Feldmann, Grotjahn und Stemmer 1986).

Noch problematischer ist es, wenn die schriftliche Instruktion ganz weggelassen wird, wie dies 1988 bei den französischen, englischen, russischen und italienischen C-Tests der Fall war. Dann ist weder sicher gestellt, daß alle Schüler die Texte unter den gleichen Voraussetzungen bearbeiten, noch daß die Aufgabenstellung überhaupt richtig verstanden wird.

5. *Vortest und statistische Analysen*

Wie meine Erfahrungen mit der Erstellung von französischen C-Tests für unterschiedlichste Probandengruppen zeigen, schätzen sogar Lehrende, die die jeweilige Testgruppe sehr gut kennen, den Schwierigkeitsgrad von C-Test-Texten in der Regel zu hoch ein. So hört man immer wieder die Aussage "Das ist doch viel zu schwer". Analysiert man dann empirisch die Testschwierigkeit, stellt man meist fest, daß der Test sogar noch zu leicht ist. Eine Abschätzung der Schwierigkeit von C-Test-Texten ist deshalb *ohne* Vortest auch bei sehr viel Erfahrung auf Seiten des Testkonstrukteurs nur sehr bedingt möglich. Ähnliches gilt in bezug auf Trennschärfen und Reliabilitäten. Obwohl Vortestergebnisse z. B. aus zehnten Klassen von Gymnasien nur sehr bedingt aussagekräftig hinsichtlich der Zielpopulation des Bundeswettbewerbs sind, erhält man dennoch zumindest einige Hinweise hinsichtlich der Eignung einzelner Texte.

Weiterhin erwiesen sich die französischen (und englischen) C-Tests bei den statistischen Nachanalysen der jeweiligen Jahrgangsresultate meist auch nicht als eindimensional. Als Folge ist die Interpretation der Testresultate nicht unproblematisch. Die Gründe für die fehlende Eindimensionalität sind bisher alles andere als klar. Liegt die fehlende Eindimensionalität z. B. an einer nicht ausreichenden Erprobung und Dimensionalitätsprüfung

der jeweiligen Vortestversion? Oder sind z. B. die Teilnehmer am offiziellen Testdurchgang, die in ihrem Antwortverhalten stark von der Gruppennorm abweichen, für Abweichungen von der Eindimensionalität verantwortlich? Die Beantwortung dieser Fragen ist m. E. eines der Desiderate weiterer Forschung zum Bundeswettbewerb.

Insgesamt gesehen erscheint es aufgrund der bisherigen Erfahrungen dringend geboten, die in jedem Jahr neu entwickelten C-Test-Versionen gründlich vorzutesten und sowohl die Vortestergebnisse als auch die eigentlichen Testergebnisse einer ausführlichen statistischen Analyse zu unterziehen.

6. *Korrekturproblematik*

6.1 *Inkonsistenzen bei der C-Test-Korrektur*

Ich komme nun zu meinem Hauptanliegen, nämlich der Korrekturproblematik. Betrachtet man die von den Auswertern des Bundeswettbewerbs korrigierten C-Tests, fällt auf, daß es sowohl innerhalb einzelner Sprachen als auch zwischen den Sprachen nicht unerhebliche Inkonsistenzen bei der Bewertung der Lösungen gibt. So werden beispielsweise Rechtschreibefehler von manchen Auswertern als ganzer Fehler, von anderen Auswertern lediglich als halber Fehler gewertet. Ob diese unterschiedliche Gewichtung z. B. negative Auswirkungen auf die Güte der jeweiligen C-Test-Version hat, ist bisher nicht untersucht worden. Weiterhin scheint vielfach auch keine Einigkeit darüber zu bestehen, was überhaupt als Rechtschreibefehler zu werten ist. Ähnliches gilt für die Wertung von Varianten als korrekte Lösung.

Wie Bauer in seinem Beitrag zu diesem Bande anhand von Beispielen nachweist, findet sich auch bei der Auswertung der englischen C-Tests ein erhebliches Maß sowohl an interpersonaler als auch an intrapersonaler Varianz. Das Problem der nicht ausreichenden Auswertungsreliabilität stellt sich übrigens - z. T. in weit stärkerem Maße - auch in anderen Testteilen des Wettbewerbs und ist, wie die Arbeit von Legenhausen (1988) zeigt, ein Erbe aus dem Schülerwettbewerb Fremdsprachen. Geht man davon aus, daß die Schülerzahlen und damit auch die Anzahl der zu

korrigierenden C-Tests weiter steigt, dann sollte das Problem der optimalen Auswertungsmethode sowie das hiermit zusammenhängende Problem der nicht ausreichenden inter- und intrapersonalen Auswertungskonstanz möglichst bald gelöst werden.

6.2 *Fehlerkategorien und deren Unschärfe*

Um erste empirische Hinweise zur Lösung der Problematik zu erhalten, habe ich insgesamt 80 C-Tests aus Hessen aus den Jahren 1987 und 1988 - in Zusammenarbeit mit zwei Französinnen - erneut ausgewertet. Dabei wurden folgende sechs Lösungskategorien unterschieden:

(1) unausgefüllt
(2) orthographisch richtiges Original
(3) grammatisch und/oder inhaltlich nicht akzeptabel
(4) orthographisch richtige Variante
(5) orthographisch falsches Original
(6) orthographisch falsche Variante

Es ist darauf hinzuweisen, daß die sechs Kategorien nicht nur für mich, sondern auch für die beiden Französinnen mit einer gewissen Unschärfe behaftet sind. So waren wir uns z. B. keineswegs immer einig, ob eine Lösung z. B. als Orthographiefehler oder als Morphologiefehler zu betrachten ist. Auch bei der Bewertung von Varianten gab es Probleme. So gab es z. B. grammatisch korrekte Lösungen, die im Mikrokontext als korrekte Variante anzusehen waren, nicht jedoch im Makrokontext. Da jedoch alle 80 C-Tests von ein und demselben Auswerter bewertet worden sind, gibt es im Gegensatz zu den vom Bundeswettbewerb durchgeführten Korrekturen zumindest keine interpersonellen Inkonsistenzen. Außerdem wurde die intrapersonale Inkonsistenz dadurch verringert, daß die inhaltliche Bedeutung der Kategorien sowohl vor Auswertungsbeginn als auch im Verlauf der Auswertung ausführlich diskutiert worden ist.

Anhand von Text 1 der C-Test-Version des Testjahrgangs 1987 sollen nun die unterschiedenen Kategorien verdeutlicht und einige Korrekturprobleme veranschaulicht werden. Der Text ist in Abbildung 1 wiedergegeben. Die Zeilen sind numeriert und die orthographisch richtigen Originallösungen (Lösungskategorie 2) jeweils unterstrichen und kursiv gesetzt.

1 Le docteur Philippe Chevreul travaille à Bayeux, une petite
2 ville de Normandie, depuis 20 ans. Maintenant, il a une bonne
3 clientèle, qu'il connaît bien. Il habite une grande maison
4 dans le centre de la ville. Il a son cabinet de consultation au
5 premier étage. Tous les matins de huit heures à midi, il va
6 travailler à l'hôpital de la ville. Pendant ce temps, sa femme
7 répond au téléphone et note les visites à domicile de l'après-
8 midi.

Abb. 1: ***Text 1 aus dem Testjahrgang 1987***

Hier einige bei den Wettbewerbsteilnehmern gefundene Lösungen und deren Bewertung:
Als orthographisch falsches Original (Kategorie 5) wurde die Lösung "pendent " für "pendant " in Zeile 6 gewertet. Das Wort "pendent" existiert zwar im Französischen als 3. Person Plural des Präsenz von "pendre", ist jedoch kontextuell absolut unangemessen und zudem den Wettbewerbsteilnehmern vermutlich nicht bekannt.

Ähnliches gilt für "habit " anstelle von "habite " in Zeile 3. Hier sind wir davon ausgegangen, daß der Kontext eindeutig ein Verb verlangt (und nicht das vielen Wettbewerbsteilnehmern möglicherweise unbekannte Substantiv "habit") und daß auch schwache Wettbewerbsteilnehmer das Verb *habiter* und die Endungen der Verben auf *-er* kennen. Aufgrund dieser Überlegung haben wir "habit " als orthographischen "Flüchtigkeitsfehler" (Kategorie 5) gewertet.

Wie die beiden Beispiele zeigen, bedarf es im Einzelfall weitreichender und häufig problematischer Interpretationen sowie vielfach auch einer detaillierten Kenntnis des Leistungsstands der Testteilnehmer, um zu einer validen Trennung zwischen Orthographiefehlern und grammatisch und/oder inhaltlich nicht akzeptablen Lösungen zu kommen. Die Problematik entsprechender Lösungskategorien insbesondere im Fall wechselnder Korrektoren und bei Korrekturen einer großen Zahl von C-Tests von ein und demselben Korrektor unter Zeitdruck dürfte offensichtlich sein.

Ein interessanter Fall ist die mehrfach aufgetretene Lösung "habite dans ". Die Lösung ist sowohl grammatisch als auch inhaltlich akzeptabel. Sie stellt jedoch einen Verstoß gegen das C-Test-

Prinzip dar, da nicht nur ein Teil eines Wortes, sondern zusätzlich noch ein ganzes Wort rekonstruiert worden ist. Das Auftreten von Lösungen dieser Art läßt sich (weitgehend) vermeiden, indem man die Testteilnehmer entsprechend instruiert. Leider hat den Teilnehmern im vorliegenden Fall keine standardisierte, schriftliche Instruktion vorgelegen. Wir haben deshalb die Lösung der Kategorie 4 (orthographisch richtige Variante) zugeordnet.

Um einen eindeutigen Grammatikfehler handelt es sich bei dem Plural "vont " für "va " in Zeile 5: Der Schüler hat nicht berücksichtigt, daß das vorangehende Personalpronomen *il* im Singular steht. Problematischer ist jedoch die Lösung "veut " anstelle von "va ". Im Mikrokontext ist "veut " akzeptabel, nicht jedoch im Makrokontext (zumindest nach einstimmiger Auffassung der drei Auswerter). Der Fall wurde deshalb der Kategorie 3 zugeordnet. Auch nicht ganz unproblematisch ist die Lösung "vient " für "va ". Geht man jedoch z. B. davon aus, daß der Text von einem Kollegen des Dr. Chevreul am Krankenhaus von Bayeux erzählt wird, dann ist die Lösung als orthographisch richtige Variante zu bewerten (Kategorie 4).

Ein sehr problematischer Fall ist die Lösung "Toutes les matières " für "Tous les matins " in Zeile 5. Das Wort "matières " ist inhaltlich eindeutig inakzeptabel (Kategorie 3). Im Zusammenhang mit "matières " ist jedoch "Toutes ", und nicht die Originallösung "Tous " grammatisch korrekt. Wertet man trotzdem "Toutes " als inakzeptabel, da es weder grammatisch noch inhaltlich in den Kontext der Originallösung "matins " paßt, wird der Schüler doppelt "bestraft". Entsprechende Fälle von Abhängigkeiten zwischen C-Test-Lösungen sind übrigens in den romanischen Sprachen nicht gerade selten (vgl. auch den Beitrag von Joppich in diesem Bande). Im vorliegenden Fall haben wir die Lösung "Toutes " im Zusammenhang mit "matières " als korrekt gewertet.

Neben den diskutierten Beispielen finden sich noch weitere Beispiele für Auswertungsprobleme in Text 1. Auch die anderen bisher im Bundeswettbewerb Fremdsprachen eingesetzten französischen Texte enthalten eine Reihe von z. T. sehr problematischen Fällen. Insgesamt gesehen dürfte deutlich geworden sein, daß die

Berücksichtigung von Varianten und Orthographiefehlern - zumindest im Französischen - notwendigerweise zu Reliabilitätsproblemen bei der Auswertung führt.

6.3 *Einige statistische Ergebnisse*

Tabelle 1 zeigt die prozentualen Häufigkeiten der sechs Auswertungskategorien für die Jahrgänge 1987 und 1988 sowie für die zusammengefaßten Jahrgänge, aufgeteilt am Leistungsmedian. Bei der Berechnung des Medians wurden lediglich orthographisch richtige Originale als korrekt gewertet. Alle Berechnungen in der vorliegenden Arbeit wurden mit Hilfe von SPSS^x (Programmversion 2.2) durchgeführt.

Tabelle 1: Häufigkeiten (%) von sechs Auswertungskategorien

Kategorie	1987 N=41	1988 N=39	87/88 N=80	>Md 87/88 N=39	<Md 87/88 N=41
unausgefüllt	12.9	16.0	14.4	4.0	24.3
orthographisch richtiges Original	60.2	60.4	60.3	72.9	48.3
grammatisch/inhaltlich nicht akzeptabel	18.9	16.0	17.5	16.0	19.0
orthographisch richtige Variante	1.3	1.7	1.5	1.5	1.5
orthographisch falsches Original	6.5	5.6	6.1	5.4	6.7
orthographisch falsche Variante	0.1	0.3	0.2	0.2	0.2

Md: Median

Wie Tabelle 1 zeigt, spielen Varianten - zumindest bei den beiden untersuchten Testversionen - glücklicherweise praktisch keine Rolle. Auch von Schülern der höheren Leistungsgruppe werden kaum Varianten gefunden. Auffallend ist, daß mehr als 6 % grammatisch korrekter Lösungen Orthographiefehler enthalten, wobei erwartungsgemäß in der schwächeren Leistungsgruppe etwas mehr Orthographiefehler zu finden sind.

In Tabelle 2 wurden fünf Auswertungsmethoden, d. h. Methoden zur Berechnung des Gesamtpunktwertes unterschieden und Mittelwerte (M), Standardabweichungen (SD), Schwierigkeiten (p) und Reliabilitäten (α) berechnet. Als 'korrekt' gewertet wurden:

Methode A: Originale ohne Orthographiefehler
Methode B: Originale ohne Orthographiefehler; Varianten ohne Orthographiefehler
Methode C: Originale mit oder ohne Orthographiefehler
Methode D: Originale mit oder ohne Orthographiefehler; Varianten mit oder ohne Orthographiefehler
Methode E: von den Auswertern des Bundeswettbewerbs Fremdsprachen vergebene Punktzahl

Tabelle 2: Mittelwerte (M), Standardabweichungen (SD), Schwierigkeiten (p) und Reliabilitäten (α) für fünf Auswertungsmethoden

	1987 N=41				1988 N=39			
	M	SD	p	α	M	SD	p	α
Methode A	48.2	11.6	.60	.89	48.3	12.7	.60	.92
Methode B	49.2	11.5	.62	.89	49.7	12.8	.62	.92
Methode C	53.4	10.7	.67	.86	52.8	12.1	.66	.91
Methode D	54.5	10.6	.68	.85	54.4	12.3	.68	.91
Methode E	49.8	11.5	.62	.89	49.7	12.9	.62	.92

als richtig gewertet:
Methode A: Originale ohne Orthographiefehler
Methode B: Originale ohne Orthographiefehler; Varianten ohne Orthographiefehler
Methode C: Originale mit oder ohne Orthographiefehler
Methode D: Originale mit oder ohne Orthographiefehler; Varianten mit oder ohne Orthographiefehler
Methode E: von den Auswertern des Bundeswettbewerbs vergebene Punktzahl

Wie aus Tabelle 2 ersichtlich ist, sind beide Testversionen hinreichend reliabel, wobei die Werte für 1988 - möglicherweise bedingt durch die etwas größeren Streuungen - geringfügig höher sind. Wertet man Orthographiefehler als 'korrekt', wird der Test erwartungsgemäß leichter, und es kommt außerdem zu einem geringfügigen Absinken der Streuungen und - möglicherweise hierdurch bedingt - der Reliabilitäten. Die Wertung von Orthographiefehlern als 'korrekt' hat somit eine potentiell negative Auswirkung auf die Meßgüte des Tests insbesondere im oberen Leistungsbereich.

Das Absinken der Streuungen bei der Wertung von Orthographiefehlern als 'korrekt' läßt sich folgendermaßen erklären: Berechnet man den Produkt-Moment-Korrelationskoeffizienten zwischen der Auswertungsmethode A (orthographisch richtiges Original) und der Auswertungskategorie 5 (orthographisch falsches Original), ergibt sich für 1987 $r_{a5} = -.50$ und für 1988 $r_{a5} = -.35$, wobei die einseitigen Irrtumswahrscheinlichkeiten .0005 bzw. 0.015 betragen. Dies bedeutet, daß die Zahl der orthographisch falschen Originallösungen mit steigender Testleistung (gemessen anhand der Zahl der orthographisch korrekten Originallösungen) abnimmt. Für die Auswertungsmethode C z. B. gilt nun, daß sich der Punktwert jedes Wettbewerbsteilnehmers (X_c) additiv zusammensetzt aus der Zahl der richtigen Lösungen nach Methode A (X_a) sowie der Zahl der orthographisch falschen Originale entsprechend Auswertungskategorie 5 (X_5), d. h. es gilt $X_c = X_a + X_5$. Hieraus folgt für die Streuung der Auswertungsmethode C: $s_c^2 = s_a^2 + s_5^2 + 2s_a s_5 r_{a5}$, d. h. die Streuung setzt sich additiv zusammen aus den Streuungen der Variablen 'Zahl der orthographisch richtigen Originale' und 'Zahl der orthographisch falschen Originale' sowie aus der Kovarianz zwischen den beiden genannten Variablen. Wird r_{a5} negativ, führt dies ab einer bestimmten Größenordnung notwendigerweise zu einem Absinken der Streuung der Auswertungsmethode C. Eine analoge Argumentation gilt für die Auswertungsmethode D. Da der varianzreduzierende Effekt der Wertung von Orthographiefehlern als 'korrekt' somit kein Zufallseffekt ist, sollte er bei der Entscheidung über die künftige Korrekturpraxis mit berücksichtigt werden.

In *Tabelle 3* sind zur Charakterisierung des Ausmaßes der Übereinstimmung zwischen den fünf Auswertungsmethoden in der oberen Dreiecksmatrix die Spearman Rangkorrelationskoeffizienten r_s und in der unteren Dreiecksmatrix die entsprechenden Werte für den Kendall Rangkorrelationskoeffizienten τ_b aufgeführt. Die beiden Koeffizienten unterscheiden sich insbesondere dadurch, daß bei der Berechnung von τ_b alle Inversionen der Rangordnung gleich gewichtet werden, während bei der Berechnung von r_s einzelne große Rangplatzdifferenzen ein besonderes Gewicht erhalten (vgl. Hays 1973, S. 792 f.). Zusätzlich wurden für Tabelle 3 und alle weiteren Tabellen auch noch Pearson

Produkt-Moment-Korrelationskoeffizienten berechnet. Da die erhaltenen Werte weitgehend mit den Werten für r_s übereinstimmten, wurde auf eine Wiedergabe verzichtet.

Tabelle 3: Spearman Rangkorrelationen (obere Dreiecksmatrix) und Kendall Rangkorrelationen (untere Dreiecksmatrix) zwischen fünf Auswertungsmethoden (1987; N=41)

	Methode A	Methode B	Methode C	Methode D	Methode E
Methode A	----	.998	.981	.980	.986
Methode B	.983	----	.979	.980	.985
Methode C	.918	.909	----	.996	.976
Methode D	.919	.915	.974	----	.974
Methode E	.931	.921	.893	.892	----

Alle Spearman Korrelationen in Tabelle 3 sind extrem hoch und unterscheiden sich numerisch nur minimal. Die Kendall Korrelationen zwischen Methode A und B sowie zwischen Methode C und D stimmen weitgehend mit den entsprechenden Spearman Korrelationen überein. Hierin spiegelt sich die Tatsache wider, daß die Zahl der von den Wettbewerbsteilnehmern gefundenen akzeptablen Varianten sehr gering ist und daß als Folge Unterschiede in der Bewertung von akzeptablen Varianten keinen Einfluß auf die Rangfolge haben. Die übrigen Kendall Korrelationen sind jedoch geringfügig niedriger als die jeweiligen Spearman Korrelationen. Insgesamt weisen die Korrelationen darauf hin, daß die Bewertung von Orthographiefehlern als 'korrekt' zu - allerdings geringen - Änderungen in der Rangfolge der Schüler zu führen scheint und daß damit zumindest im Einzelfall die Auswertungsmethode über die Preisvergabe mit entscheiden kann.

Auffallend sind auch die vergleichsweise etwas geringeren Korrelationen der Methoden A bis D mit Methode E, d. h. mit den Punktzahlen, die von den Korrektoren des Bundeswettbewerbs vergeben worden sind. In diesem Sachverhalt dürfte sich die bereits diskutierte inter- und intrapersonale Varianz der Bewertungsmaßstäbe bei der Korrektur durch den Bundeswettbewerb Fremdsprachen widerspiegeln.

Tabelle 4 zeigt analog zu Tabelle 3 die Korrelationen zwischen den fünf Auswertungsmethoden für das Jahr 1988. Ein Vergleich

der beiden Tabellen ergibt keine auffallenden Unterschiede. Dies kann als ein Hinweis auf die Stichprobenunabhängigkeit der erhaltenen Koeffizienten gewertet werden.

Tabelle 4: Spearman Rangkorrelationen (obere Dreiecksmatrix) und Kendall Rangkorrelationen (untere Dreiecksmatrix) zwischen fünf Auswertungsmethoden (1988; N=39)

	Methode A	Methode B	Methode C	Methode D	Methode E
Methode A	----	.996	.988	.986	.990
Methode B	.971	----	.981	.984	.991
Methode C	.929	.913	----	.997	.980
Methode D	.924	.925	.975	----	.979
Methode E	.945	.944	.894	.902	----

In Tabelle 5 sind wiederum die Schüler aus den Jahren 1987 und 1988 zusammengefaßt und anhand des Medians der Methode A in zwei Leistungsgruppen aufgeteilt.

Tabelle 5: Spearman Rangkorrelationen (obere Dreiecksmatrix) und Kendall Rangkorrelationen (untere Dreiecksmatrix) zwischen fünf Auswertungsmethoden, aufgeteilt am Median der Auswertungsmethode A (1987/88; N=80)

(a) obere Leistungsgruppe (N=39)

	Methode A	Methode B	Methode C	Methode D	Methode E
Methode A	----	.982	.923	.919	.974
Methode B	.936	----	.906	.920	.964
Methode C	.797	.776	----	.989	.904
Methode D	.802	.800	.947	----	.904
Methode E	.896	.878	.766	.766	----

(b) untere Leistungsgruppe (N=41)

	Methode A	Methode B	Methode C	Methode D	Methode E
Methode A	----	.991	.959	.959	.972
Methode B	.955	----	.945	.959	.969
Methode C	.866	.831	----	.989	.948
Methode D	.863	.859	.948	----	.951
Methode E	.895	.879	.829	.843	----

Es fällt auf, daß Unterschiede in der Wertung von Orthographiefehlern anscheinend eher in der oberen Leistungsgruppe und damit in dem für den Bundeswettbewerb entscheidenden Bereich zu Abweichungen in der Rangfolge der Schüler führen. Dieser Befund stimmt im übrigen in der Tendenz mit den Ergebnissen einer Untersuchung bei Studienanfängern des Französischen (Lehramt und Magister) an der Ruhr-Universität Bochum überein (vgl. Grotjahn 1987, S. 241). Da die erhaltenen Unterschiede zwischen den Leistungsgruppen jedoch sehr gering sind, möchte ich, solange das Ergebnis nicht durch weitere Studien im Rahmen des Bundeswettbewerb Fremdsprachen bestätigt worden ist, auf eine weitergehende Interpretation verzichten.

7. *Schlußbemerkung*

Dieser Beitrag hat gezeigt, daß es in bezug auf die Problembereiche Textauswahl, Konstruktionsprinzip, Instruktion, Vortest, Korrektur und statistische Analysen eine Reihe von zu klärenden Fragen gibt. Dies gilt insbesondere für die Korrektur. Hier ist zu fragen, welche Konsequenzen für die zukünftige Korrekturpraxis zu ziehen sind, falls sich die erhaltenen Resultate anhand größerer Stichproben replizieren lassen sollten.

Sollte nachgewiesen werden können, daß die aus den unterschiedlichen Korrekturmethoden resultierenden Abweichungen in der Rangfolge der Wettbewerbsteilnehmer nicht vor allem auf Reliabilitätsprobleme bei der Korrektur zurückzuführen sind, ist zu klären, ob und in welchem Maße der Bundeswettbewerb Entscheidungen über Preisträger mit abhängig machen will von der Beherrschung der Orthographie der jeweiligen Wettbewerbssprache.

Sollten Reliabilitätsprobleme jedoch die Hauptursache sein, dann böte es sich an, Orthographiefehler in Zukunft grundsätzlich als 'inkorrekt' zu werten. Eine Ausnahme könnte man eventuell im Fall von eindeutigen Akzentfehlern machen. Zusätzlich könnte man auch noch die ebenfalls mit Auswertungsproblemen behafteten akzeptablen Varianten als 'inkorrekt' werten oder zumindest nur solche Varianten als korrekt werten, die nach Überprüfung durch Muttersprachler oder vergleichsweise kompetente Beurteiler als akzeptable Varianten auf den den Korrektoren auszuhän-

digenden Lösungsbögen aufgelistet sind. Die Vorteile dieses Verfahrens, das die Auswertungsökonomie deutlich erhöhen würde, dürfte angesichts steigender Teilnehmerzahlen unmittelbar einsichtig sein. Zudem hätte dieser Auswertungsmodus auch noch den wünschenswerten Effekt, daß die Tests schwerer würden und dadurch besser im oberen Leistungsbereich differenzieren würden.

Literaturhinweise

Feldmann, U./Grotjahn, R./Stemmer, B.: "Was messen Sprachtests eigentlich? Überlegungen zur introspektiven Validierung von C-Tests." In: Seminar für Sprachlehrforschung der Ruhr-Universität Bochum (ed.): *Probleme und Perspektiven der Sprachlehrforschung*. Bochumer Beiträge zum Fremdsprachenunterricht in Forschung und Lehre. Frankfurt/M.: Scriptor 1988: 325 - 338.

Grotjahn, R.: "How to Construct and Evaluate a C-Test: A Discussion of Some Problems and Some Statistical Analyses." In: Grotjahn, R./Klein-Braley, C./ Stevenson, D. K. (eds.): *Taking Their Measure: The Validity and Validation of language Tests*. Bochum: Brockmeyer 1987: 219 - 253.

Hays, W. L.: *Statistics for the Social Sciences*. London & New York: Holt, Rinehart and Winston 1973.

Legenhausen, L.: "Fehler-Fuzziness und Bewertungsvarianz." In: Finkenstaedt, Th./ Weller, F.-R. (eds.): *Schrittweise zur Validität*. Der Schülerwettbewerb im Stifterverband für die Deutsche Wissenschaft 1979 - 1984. Augsburg: Universität 1988: 235 - 251. (= Augsburger I & I - Schriften 41)

Bernd Kielhöfer

Über die Schwierigkeiten, eine Geschichte zu schreiben. Die semi-kreative Aufgabenstellung Französisch.

O. *Vorüberlegungen zum Profil der Aufgabe*

Bevor ich die Aufgabenstellungen konkret analysiere, möchte ich einige Vorüberlegungen zu den Anforderungen an die semi-kreative Aufgabe im Rahmen des Wettbewerbs anstellen, gewissermaßen ihr Profil ermitteln:

1) Die semi-kreative Aufgabe soll komplementär zu den anderen Aufgaben des Wettbewerbs sein. Das Ausmaß der Komplementarität läßt sich durch Interkorrelationen feststellen, die Korrelationsquoeffizienten sollten m. E. zwischen r = 0,4 - 0,6 liegen.

2) Die Besonderheit der semi-kreativen Aufgabe besteht in der minimalen Lenkung oder, positiv gesagt, in der Gestaltungsfreiheit. Die Schüler sollen Gelegenheit erhalten, ihr eigenes sprachliches Ausdrucksvermögen voll zu entfalten. Dieses Merkmal ist wohl mit dem Attribut "semi-kreativ" gemeint.

3) Die semi-kreative Aufgabe soll stimulierend und motivierend sein. Diese Anforderung gehört ja generell zur Philosophie des Wettbewerbs, sie gilt aber speziell für diesen Aufgabenteil, da sie in anderen Bereichen (in den Tests z. B.) nur schwer einzulösen ist.

   Originalität der Aufgabenstellung, aber auch die altersgemäße Wahl der Inhalte und Formen sind wichtige Faktoren, die zur Stimulation beitragen.

4) Die semi-kreative Aufgabe soll selektiv sein. Der Schwierigkeitsgrad sollte an der oberen Grenze liegen, da es zum Wesen des Wettbewerbs gehört, mit einer solchen Hürde die Besten herauszufinden.

   Zusammenfassend könnte man sagen, daß die Aufgabe eine stimulierende Herausforderung sein soll, zu zeigen, was man kann.