CHUN NR. 24/2009, S. 21-45

Was ist Chinesischkompetenz?

Standardisierte Chinesischprüfungen und Untersuchungen zur Sprachstandsmessung

Florian Meyer

1. Einleitung: Zur zunehmenden Relevanz von Sprachtests

Sprachtests[1] sind in vielen Ländern (vor allem den USA) zu einem integralen Teil des Bildungssystems geworden. Die Resultate solcher Tests entscheiden immer häufiger darüber, Studenten zu einem Studium an einer Universität zuzulassen,[2] ihnen ein Stipendium zu geben, sie in das passende Sprachprogramm einzugliedern oder auch, einen Bewerber für einen Arbeitsplatz letztlich einzustellen oder nicht. In Deutschland sind seit Beginn des neuen Jahrtausends auch Sprachstandstests vor der Einschulung für Kinder mit Migrationshintergrund und Sprachtests für Zuwanderer hinzugetreten.[3] Allerdings werden hierzulande sowohl im schulischen als auch universitären Fremdsprachenunterricht bisher nur äußerst selten formelle Tests (vgl. Abschnitt 3.5.) eingesetzt, die einen über den jeweiligen Lernkontext hinausgehenden Vergleich der fremdsprachlichen Kompetenz erlauben. Im Gegenteil: Häufig werden – auf der Basis wenig transparenter Kriterien – Noten vergeben, die je nach Lehrer, Lerngruppe, Schule oder Bundesland für völlig unterschiedliche fremdsprachliche Leistungen stehen können. Daraus resultiert u. a., dass ein potenzieller Arbeitgeber anhand vorgelegter Zeugnisse kaum dazu in der Lage ist, zu halbwegs verlässlichen Einschätzungen der fremdsprachlichen Fähigkeiten und Fertigkeiten[4] eines Bewerbers zu gelangen (Grotjahn 2003:4).

---

[1] Bei Sprachtests kann zwischen *achievement tests* und *proficiency tests* unterschieden werden. *Achievement tests* stellen definierte Leistungen fest, oft bezogen auf einen Sprachkurs oder Sprachunterricht. *Proficiency tests* erfassen die Fähigkeiten zum allgemeinen Gebrauch einer Sprache, unabhängig davon, wie diese erworben wurden. (Vollmer 2007:365)

[2] Beispiele sind etwa der *Test of English as a Foreign Language* (TOEFL), der häufig für ein Studium an nordamerikanischen Universitäten verlangt wird, oder der *Test Deutsch als Fremdsprache* (TestDaF), mit dem Studierende, die ein Studium an einer deutschen Universität aufnehmen wollen, die erforderlichen deutschen Sprachkenntnisse nachweisen können.

[3] Sprachtests für Zuwanderer stehen stark in der Kritik, da sie oft als politisches Instrument und so genannte "gate keeper" dienen. (Kühn/Lehker/Timmermann 2005:7) Seit 2007 müssen in Nordrhein-Westfalen alle Vierjährigen einen Sprachtest, Delfin 4, ablegen.

[4] In der Sprachstandsmessung wird zwischen den Begriffen Fähigkeit (engl. *ability*) und Fertigkeit (engl. *skill*) unterschieden. Die Fähigkeit bezeichnet das Potenzial oder die Disposition einer Person zur Bewältigung bestimmter Aufgaben und ist somit ein nicht direkt beobachtbares Merkmal, das lediglich indirekt anhand der beobachteten Leistung (engl. *perfor-*

Lyle Bachman, einer der gegenwärtig einflussreichsten Forscher für Sprachstandsmessung, weist in diesem Zusammenhang darauf hin, dass Sprachtests über ein ungemein großes Potenzial verfügen, und warnt gleichzeitig, dass sie aber umgekehrt auch großen Schaden anrichten können, wenn sie den so genannten Gütekriterien der Testtheorie nicht ausreichend genügen:

> Language tests thus have the potential for helping us collect useful information that will benefit a wide variety of individuals. […] [T]o realize this potential, we need to be able to demonstrate that scores we obtain from language tests are reliable […]. If the language tests we use do not provide reliable information […], we risk making incorrect and unfair decisions […]. (Bachman 2004:3)

In Bezug auf Chinesisch als Fremdsprache lässt sich feststellen, dass *language proficiency tests* wie die vom *Hanban* von der VR China aus organisierte HSK (*Hanyu Shuiping Kaoshi* 汉语水平考试) oder der auf der Republik Taiwan vom *Steering Committee for the Test Of Proficiency* verantwortete TOP (*Huayuwen Nengli Ceyan* 華語文能力測驗) einen immer größeren Einfluss auf die Lernenden und Lehrenden des Chinesischen ausüben – und zwar sowohl innerhalb Chinas und Taiwans als auch außerhalb. Diesen Einfluss kann man an der enorm gestiegenen Zahl der weltweiten Testteilnehmer der HSK ablesen, die im Jahr 2006 bei knapp über 160.000 lag.[5] Die meisten Teilnehmer legen den Test in erster Linie ab, um einen Studienplatz an einer chinesischen Universität zu bekommen, oder um ihre Chancen auf dem Arbeitsmarkt zu erhöhen.[6]

Auch wenn in Deutschland HSK-Zertifikate in der Regel derzeit noch nicht über Arbeitsplätze entscheiden,[7] so nimmt das Thema trotzdem an Relevanz zu:

1) Die HSK wird seit Mai 2009 in Deutschland in mittlerweile sieben vom *Hanban* offiziell genehmigten Testzentren durchgeführt: Berlin, Düsseldorf, Erlangen/Nürnberg, Frankfurt, Hannover, Hamburg und Trier.[8] Der TOP soll laut mündlicher Auskunft der Taipeh-Vertretung in Berlin 2009 erstmalig in Deutschland abgehalten werden, und zwar in Berlin.

---

mance) bei vorgelegten Aufgaben gemessen werden kann. Ist die Fähigkeit zum automatisierten Handeln gemeint, spricht man von Fertigkeit oder Können. (Grotjahn 2003:8)

[5] An der ersten HSK 1989 in Peking nahmen nicht ganz 50 Kandidaten teil, 1995 waren es 12.310, 2006 schon 162.781. Die Zahl der TOP-Kandidaten lag 2008 bei 2.135. Vgl. Li (2007), Steering Committee for the Test Of Proficiency-Huayu (18.02.2009) und Guojia Hanban (26.07.2007).

[6] Diese Hauptmotive treffen für deutsche Testteilnehmer in der Regel bisher kaum zu. Gemäß eigener Befragung wollen deutsche Lerner vorrangig wissen, wie gut ihr allgemeiner Sprachstand im Chinesischen ist.

[7] In Deutschland liegt die Zahl der HSK-Teilnehmer pro Jahr bei ca. 150-250.

[8] Die Prüfungsorte in Deutschland sind: das HSK-Zentrum-Erlangen (seit 2004 am Lehrstuhl für Sinologie der Friedrich-Alexander-Universität Erlangen-Nürnberg), das Chinesische Zentrum Hannover e.V. sowie die Konfuzius-Institute in Berlin (seit Oktober 2007), Frankfurt (seit Oktober 2008), Hamburg (seit Mai 2008), Düsseldorf und Trier (beide seit Mai 2009). Zur frühen Prüfungshistorie der HSK in Deutschland siehe Kaden (2004:3).

2) Die Zahl der Chinesischlernenden in Deutschland steigt stetig. Ende 2008 wurde Chinesisch an 164 Schulen als reguläres Schulfach oder in Form von Arbeitsgemeinschaften angeboten, die Schülerzahl liegt bei etwa 3.200.[9] Vergleicht man sie mit der Zahl der Schüler, die in Frankreich Chinesisch lernen – nämlich 15.990 (Stand: 2006; Hoffmann/Guder 2007:188) –, sieht man, welches große Entwicklungspotenzial trotz der bisherigen enormen Steigerungsraten der letzten Jahre in Deutschland noch erwartet werden darf.

2. Trend zur Standardisierung: Mehr Vergleichbarkeit und Transparenz

Auch wenn formelle Sprachtests in Deutschland bzw. in Europa im Vergleich zu den USA bislang eine eher untergeordnete Rolle spielen, so ist die Tendenz zu mehr Standardisierung eindeutig. Diese liegt in erster Linie an der stetig wachsenden beruflichen Mobilität, die dazu führt, dass eine "den jeweiligen Lernkontext überschreitende Messung und Beurteilung fremdsprachlicher Leistungen" immer stärker verlangt wird (Grotjahn 2003:1). Im europäischen Kontext ist an dieser Stelle der Gemeinsame europäische Referenzrahmen (GER)[10] zu nennen, der versucht, Transparenz und Vergleichbarkeit bei der Beurteilung und Zertifizierung fremdsprachlicher Kompetenz zu erhöhen, wobei der GER auch auf außereuropäische Sprachen wie das Chinesische mittlerweile einen hohen Einfluss ausübt (Yang/Zhang 2007:107-112). Die Bemühung um das Setzen von Bildungsstandards in Deutschland äußert sich zurzeit außerdem besonders im schulischen Sektor: Die seit dem Jahr 2000 regelmäßig durchgeführten PISA-Studien[11] führten Ende 2001 zu dem so genannten PISA-Schock, womit die Diskussion um Qualitätsstandards für Bildungsprozesse in den Mittelpunkt des öffentlichen Interesses rückte.[12] Die Relevanz von Standards wirkt sich auch auf das Fach Chinesisch als Fremdsprache aus. Sun Dejin, seinerzeit Leiter des HSK-Testzentrums an der Beijing Language and Culture University (BLCU), sagt diesbezüglich, dass "Standards und Rahmenvorgaben für die gesamte

[9] Kultusministerkonferenz: *Chinesisch an Schulen in Deutschland* (2008:23).

[10] Europarat (2001).

[11] Die PISA-Studien sind Schulleistungsuntersuchungen in den OECD-Mitgliedsstaaten und einigen Partnerstaaten. Sie sollen alltags- und berufsrelevante Kenntnisse und Fähigkeiten 15-jähriger Schüler messen. PISA steht auf Englisch für *Program for International Student Assessment* (Programm zur internationalen Schülerbewertung).

[12] Seit 2008 besteht in ganz Deutschland das Zentralabitur (einige Länder prüfen noch nicht in allen Fächern zentral). Zu erwähnen ist auch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB; URL: http://www.iqb.hu-berlin.de/institut), welches 2004 als wissenschaftliche Einrichtung der Länder gegründet wurde und sich an der Humboldt-Universität zu Berlin befindet. Seine Hauptaufgabe ist die Weiterentwicklung, Operationalisierung, Normierung und Überprüfung von Bildungsstandards. Wie der Autor dieses Beitrags in einem Gespräch mit dem damaligen Leiter des IQB, Professor Olaf Köller im Oktober 2007 erfuhr, ist das Schulfach Chinesisch in naher Zukunft kein Forschungsschwerpunkt des IQB.

Didaktik und wissenschaftliche Disziplin des Chinesischen als Fremdsprache äußerst große Bedeutung und Nutzen haben."[13] Es gibt jedoch auch Vertreter der gegenläufigen Meinung, die vor möglichen negativen Folgen von Standards warnen, auf die an dieser Stelle aber nicht genauer eingegangen werden soll.[14]

Die HSK ist momentan in Deutschland das einzig verfügbare Instrumentarium für eine hochobjektive und reliable Messung der Chinesischkompetenz, die die erbrachte Leistung zudem in Bezug zu einer Normgruppe setzt. Jedoch wird sie von nicht wenigen Chinesischlernenden und Dozenten in Deutschland als "unobjektiv", "zu schriftzeichenlastig"[15] oder schlichtweg als "nicht valide" angesehen, wie der Autor dieses Aufsatzes schon des Öfteren feststellen konnte. Viele Lerner und Lehrende wissen zudem nicht genau, was die HSK misst oder wer an ihr teilnehmen kann.[16] Fragen nach der Qualität der HSK als auch nach der Aussagekraft der Ergebnisse von Testteilnehmern drängen sich daher immer mehr auf und bedürfen dringend eingehender Untersuchungen.

In diesem Aufsatz will ich einen Einblick in die Hauptaspekte der HSK-Forschung geben (Abschnitt 3), wobei ich zunächst die oben erwähnten Gütekriterien der Sprachstandsmessung darstelle. Diese sind das entscheidende Maß für die Qualität eines Sprachtests; sie werden mit Beispielen aus der HSK-Forschung erläutert. Im zweiten Teil dieses Aufsatzes werden dann erste Ergebnisse meiner Untersuchung zum Zusammenhang von HSK-Ergebnissen zur Lerndauer bzw. zum Unterrichtsaufwand präsentiert (siehe Abschnitt 4).

---

[13] “标准和大纲对于整个对外汉语教学事业和学科具有非常重要的意义和作用。” Sun (2007:129-138).

[14] Neben der Tendenz zu mehr Standardisierung wird auch eine Tendenz hin zu einer "die Subjektivität des Lerners und die Individualität des Lernprozesses fokussierenden Form der Beurteilung" eingefordert. Nach Grotjahns Ansicht müssen sich beide Alternativen allerdings nicht notwendigerweise gegenseitig ausschließen. Vgl. Grotjahn (2003).

[15] Der Vorwurf, dass die HSK zu viel Wert auf chinesische Schriftzeichen lege, wird nach Beobachtung des Autors zumeist von HSK-Kandidaten, die nicht aus dem ostasiatischen Raum kommen, erhoben. Dieses Problem wurde vor einigen Jahren ebenfalls schon von chinesischen Wissenschaftlern thematisiert. Vgl. u. a. Jing (2004:25).

[16] Aus einer Broschüre der VHS Düsseldorf zum "Turbokurs Chinesisch für 10-12jährige": "Nach den vorgesehenen 10 Doppelstunden beherrschen die Schülerinnen und Schüler 200 Wörter und 120 Zeichen. […] Nach dem Besuch von Aufbaukursen kann bei entsprechender Eignung der international anerkannte *Chinese Proficiency Test (HSK – Hanyu Shuiping Kaoshi)* abgelegt werden." (Volkshochschule Landeshauptstadt Düsseldorf 2008: 66; Hervorhebung im Original fett.)

3. Theoretische Grundlagen der Sprachstandsmessung

Bewertungsverfahren müssen bestimmten Gütekriterien[17] genügen. Diese stellen den Kern für die Qualität eines Bewertungsverfahrens dar. Um beurteilen zu können, ob ein Test "gut" ist oder nicht, muss man ihn auf seine Gütekriterien hin überprüfen. In diesem Abschnitt sollen die wichtigsten Gütekriterien aus Perspektive der klassischen Testtheorie[18] beschrieben und kurz mit Beispielen bezogen auf die chinesische HSK-Testforschung erläutert werden, was neben einem Einblick in dieses Forschungsfeld der genauen Verortung der eigenen Untersuchung (Abschnitt 4) innerhalb der Sprachstandsmessung dient.

3.1. Objektivität

Lienert/Raatz (1994:7) definieren Objektivität als "den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind." Um eine zufrieden stellende Objektivität zu erzielen, ist der Grad der Standardisierung der Durchführung, Auswertung und Interpretation eines Tests von zentraler Bedeutung. Kommen unterschiedliche Tester bei den gleichen Kandidaten zu den gleichen Ergebnissen, dann ist ein Test vollständig objektiv. (Ingenkamp 1997) Weiterhin wird zwischen Durchführungsobjektivität (Grad der Unabhängigkeit der Ergebnisse von der Durchführung), Auswertungsobjektivität (Auswertung der registrierten Reaktionen der Prüflinge) und Interpretationsobjektivität (Grad der Unabhängigkeit der Interpretation der Testergebnisse von der Person des interpretierenden Testbenutzers) unterschieden. (Grotjahn 2003:19)

Eine komplette *Chu-zhongdeng*-HSK-Prüfung[19] besteht aus 170 Items, von denen 154 Multiple-Choice-Items sind, bei denen jeweils nur eine Antwort richtig ist. Im Prüfungsteil Lückentext (*zonghe tiankong* 综合填空) müssen in mehreren Textabschnitten Lücken mit Schriftzeichen gefüllt werden. Auch hier ist die Lösung eindeutig. Die HSK verfügt somit über eine absolute Auswertungsobjektivität. Die Durchführungsobjektivität würde ich ebenfalls relativ hoch ansetzen, da diese sehr detailliert geregelt ist und in der Praxis – zumindest nach eigenen Erfahrungen an insgesamt vier verschiedenen Testorten – kaum Unterschiede in der Durchführung erkennen ließen. Die Zeitdauer für die Bearbeitung ist exakt vorgegeben und wird von den Testadministratoren in der Regel penibel

[17] Die Termini der Sprachstandsmessung und der Testtheorie kommen größtenteils aus dem Englischen. Eine Übersichtstabelle am Ende dieses Textes zeigt die gängigen deutschen Begriffe sowie ihre englischen und chinesischen Entsprechungen.

[18] Es gibt weiterhin die probabilistische Testtheorie. Siehe dazu Rost (1996), Lienert/ Raatz (1994:5).

[19] 初、中等汉语水平考试, the Elementary and Intermediate Chinese Proficiency Test.

eingehalten.[20] Der Sitzplatz darf nicht frei gewählt werden. Der Grad der Interpretationsobjektivität ist ebenfalls sehr hoch, allerdings beeinträchtigt die Intransparenz des Niveaustufensystems[21] die praktische Beurteilung von Testergebnissen, was nicht nur Kandidaten Probleme bereitet, sondern selbst Testadministratoren.[22]

3.2. Reliabilität

Die Reliabilität oder Zuverlässigkeit bezieht sich auf die Exaktheit, mit der ein Test misst. Dies schließt die Reproduzierbarkeit von Merkmalen gleicher Ausprägung ein. Die Exaktheit ist jedoch unabhängig davon, ob der Test wirklich die Eigenschaft misst, die gemessen werden soll. Ein Sprachtest könnte z. B. hoch reliabel sein, das heißt, dass ein Prüfling auf einem (idealerweise) konstanten Sprachniveau immer wieder ein nahezu gleiches Testergebnis erzielen würde, auch wenn der Test u. U. vorrangig etwas völlig Anderes misst. Werden Testergebnisse nur sehr ungenau reproduziert, ist dies ein Hinweis auf eine niedrige bzw. unbefriedigende Reliabilität (unter der Prämisse, dass sich die zu messende Eigenschaft in der Zwischenzeit, z. B. durch Sprachunterricht, nicht verändert hat).[23] Die Reliabilität ist abhängig von der Stichprobe und charakterisiert damit

---

[20] Die Bearbeitungsdauer für jeden Testteil ist minutengenau vorgegeben. Ein neuer Testteil darf nicht vor einem bereits bearbeiteten oder in Bearbeitung befindlichem begonnen werden und die Bearbeitung erfolgt zeitgleich mit allen anderen Kandidaten. Vorhergehende Testteile dürfen nach Beginn des nachfolgenden Testteils nicht mehr bearbeitet werden.

[21] Kaden (2004) übersetzt den Begriff *dengji fenshu* 等级分数 mit "Rangziffer". In der Testtheorie bezeichnet ein "Rang" jedoch zumeist den Platz, den ein Individuum, das in einer Gruppe hinsichtlich eines bestimmten Kriteriums eingeordnet wird, einnimmt. Ich halte es für sinnvoller, *dengji fenshu* mit "Niveaustufe" oder – wenn aus dem Kontext ersichtlich – "Stufe" zu übersetzen. Beide Bezeichnungen werden auch im GER verwendet. Die offiziellen HSK-Begriffe sind leider häufiger etwas missverständlich gewählt. So werden die verschiedenen *Test-* bzw. *Prüfungsformate* als *kaoshi dengji* 考试等级 bezeichnet (was man selbstverständlich auch – wie Kaden – als "Prüfungsstufe" übersetzen kann).

[22] Sun et al. (2007:128). Häufig wird übersehen, dass die HSK-Stufe 3 von zwei Testformaten abgedeckt wird. Vgl. Universität Trier (2009).

[23] Die sich auf zeitliche Stabilität der Testergebnisse beziehende Reliabilität wird als Retest- oder Testwiederholungsreliabilität bezeichnet. Verfügt man über zwei parallele Formen eines Tests, die dieselbe Eigenschaft messen (und sich in der Güte ihrer Messung nur unwesentlich unterscheiden), können an ein und derselben Stichprobe beide Tests eingesetzt werden. Die Korrelation der Ergebnisse beider Tests ist die Paralleltestreliabilität. Zur Messung der Reliabilität im Sinne von Item-Konsistenz wird meistens der "Cronbachs Alpha-Koeffizient" benutzt, der Werte zwischen 0 (absolut unzuverlässig) und 1 (absolut zuverlässig) annehmen kann. Zur Differenzierung zwischen Individuen wird eine Reliabilität von mindestens 0,9 gefordert, für Gruppen reicht ein Wert ab 0,6. (Grotjahn 2003:20f.; Lienert/Raatz 1994:189-192)

nur die durchschnittliche Zuverlässigkeit eines Tests bezogen auf eine bestimmte Population (einer Zielgruppe aus der Stichprobe).

Zur Reliabilität der HSK gibt es zahlreiche Untersuchungen, wie etwa von Chai (2002), der die Paralleltestreliabilität (siehe Fußnote 23) der *Chu-zhongdeng*-HSK gemessen hat. Dafür ließ er 152 Studenten der BLCU an zwei unterschiedlichen HSK im Abstand von zwei Wochen teilnehmen, woraus er eine Paralleltestreliabilität von 0,88 bzw. 0,90 berechnete (Chai 2002:65-69), was man dahingehend interpretieren kann, dass zwei verschiedene *Chu-zhongdeng*-HSK-Prüfungen zu etwa 90 % das gleiche Konstrukt messen – ein sehr zufriedenstellender Wert für einen *language proficiency test*.[24] Tian Qingyuan hat die Reliabilität bei Aufsatzbewertungen der *Gaodeng*-HSK-Prüfung (Oberstufe) untersucht (anhand eines Korpus von 406 Aufsätzen), Nie Dan behandelt die Retest-Reliabilität, Zhang Kai widmet in seinem Standardwerk zur HSK-Forschung der Reliabilität ein gesamtes Kapitel.[25]

3.3. Validität

Das entscheidende Gütekriterium eines Tests ist die Validität oder Gültigkeit. Sie bezieht sich auf das Ausmaß, in dem der Test das erfasst, was er erfassen soll, sowie auf die Adäquatheit der Gültigkeit der Entscheidungen. Die Validität muss stets in Abhängigkeit von der spezifischen Verwendung eines Tests gesehen werden und Aussagen wie "der Test ist valide" ohne Nennung weiterer Angaben sollten deswegen sehr argwöhnisch betrachtet werden. Es wird zwischen Inhaltsvalidität (Kontentvalidität), kriterienbezogener Validität (empirische Validität), Augenscheinvalidität (*face validity*) und Konstruktvalidität unterschieden. Jedoch sind diese Validitätsarten nicht immer klar voneinander abzugrenzen. Die HSK-Forschung hat die Relevanz der Validität ebenfalls erkannt und sich in den letzten Jahren immer mehr auf dieses auch als "wichtigstes" Gütekriterium bezeichnete Qualitätsmerkmal konzentriert. (Sun 2007:135)

3.3.1. Inhaltsvalidität

Die Inhaltsvalidität gibt an, inwieweit Testaufgaben (*items/shiti* 试题) dazu geeignet sind, bestimmte Aspekte eines Lernstoffs oder auch bestimmte Verhaltensweisen zu erfassen (inhaltliche Repräsentativität). In der Regel wird die Inhaltsvalidität durch Expertenurteile ermittelt, bei informellen Tests (s. Abschnitt

[24] Nach Lado (1961)gilt ein Wert ≥ 0,9 als ideal.

[25] Tian (2007:65-69), Nie D. (2006), Sun (2007:134), Zhang (2006). Bei Zhang finden sich zahlreiche Aufsätze zur HSK-Forschung, geordnet in die Kapitel Testdesign, Punktvergabesystem, Reliabilität, Validität und Fairness, Testthema-Design und Probleme des Test-Itempools.

3.5) entscheiden Lehrer, die mit dem Lernstoff und der Lerngruppe hinreichend vertraut sind. Da Experten jedoch erheblich in der Einschätzung der Gültigkeit einer Aufgabe differieren können, ist die Inhaltsvalidität ein nicht unproblematisches Kriterium. Selbst wenn jedoch die Inhalte eines Tests den zu messenden Bereich angemessen repräsentieren, ist damit keineswegs sicher gestellt, dass der Test einen validen Rückschluss auf das zu messende Merkmal erlaubt. Beispiel: Aufgaben eines Hörverstehenstests können genau die Inhalte abfragen, die dem Lehrplan zugrunde liegen. Stellen sie aber einen zu hohen Anspruch an das Arbeitsgedächtnis, weil sie z. B. zu lang sind, wird die tatsächliche Hörverstehensleistung massiv unterschätzt, es sei denn, das Arbeitsgedächtnis ist ein wichtiger Bestandteil des zu messenden Konstrukts. Kontent- bzw. inhaltsvalide Tests müssen daher nicht nur die Inhalte der Items festlegen, sondern auch die Form bzw. das Itemformat.

Beispiele aus der HSK-Forschung: Jing Chengs Kritik, dass die HSK z. T. den Pekinger Dialekt (*Beijing fangyan* 北京方言) überprüft (Jing 2004:22-32), oder Arbeiten zum HSK-Wortschatz, welchen man als relativ kontentvalide ansehen kann. Da Jun hat – zumindest bei hoch- und mittelfrequenten Wörtern des HSK-Wortschatzes – herausgefunden, dass dieser sich zu einem großen Teil mit dem Wortschatz deckt, der in journalistischen Texten verwendet wird (Da 2007:251-278). Die ersten beiden Stufen (ca. 3.000 Wörter) reichen für etwa 80 % alltäglichen Lesematerials; nimmt man die dritte Stufe (ca. 5.000 Wörter) hinzu, werden ca. 90 % abgedeckt. Allerdings wird immer wieder gefordert, die Liste häufiger zu aktualisieren (vgl. Nie H. 2007:89). Zur Inhaltsvalidität gehört ebenso die Problematik der Definition von *Putonghua*.[26]

3.3.2. Kriterienbezogene Validität (Empirische Validität)

Für Aussagen zur kriterienbezogenen oder empirischen Validität wird geprüft, inwieweit die Testergebnisse mit den Werten eines unabhängigen Außenkriteriums, das ebenfalls das erfassende Merkmal misst, übereinstimmen. Außenkriterien können z. B. ein anderer Test oder Schulnoten sein. Der Grad der Übereinstimmung zwischen Test und Kriterium wird mithilfe eines Korrelationskoeffizienten (nach Pearson) gemessen, der Werte zwischen –1 und +1 annehmen kann. Beispiel: Eine Korrelation von 0,5 zwischen dem Testteil "Leseverstehen" der HSK und dem Testteil "Leseverstehen" des TOP würde bedeuten, dass beide Tests nur zu 25 % [27] die gleiche Eigenschaft messen, und zugleich wüsste man,

[26] Beispiel aus dem *Intermediate-TOP* (*zhongdeng* 中等): In einer Hörverstehensaufgabe fällt der Satz: "我把皮夾在計程車忘了。" *Pijia* 皮夾 (Brieftasche) wird auf Taiwan synonym für *qianbao* 钱包 verwendet, *jichengche* 計程車 für *chuzu qiche* 出租汽车. Dieser Satz ist für Chinesischlerner ohne genügend Taiwanerfahrung wahrscheinlich unverständlich. Die Definition des Konstrukts *Putonghua* hat daher großen Einfluss auf die Inhaltsvalidität.

[27] Der Korrelationskoeffizient von 0,5 muss quadriert und mit 100 multipliziert werden.

dass man nur sehr ungenau von den Ergebnissen des einen Testteils auf die des anderen schließen kann. Die weiter unten beschriebene empirische Untersuchung ist genau hier zu verorten, da sie ein unabhängiges Außenkriterium, in diesem Fall die Variable Lerndauer, in Bezug zum erzielten HSK-Ergebnis setzt (vgl. dazu Abschnitt 4).

3.3.3. Augenscheinvalidität (*face validity*)

Die Augenscheinvalidität beschreibt die Gültigkeit, die der Test in den Augen der Getesteten und Testabnehmer hat. Die Augenscheinvalidität ist nicht unwichtig für die Akzeptanz eines Testverfahrens. Neuere Testverfahren, wie der 1981 erstmals vorgestellte C-Test[28], haben häufig eine geringe Augenscheinvalidität für die anvisierte Personengruppe (Lerner, Lehrende, Testdurchführende etc.). Der (praktische) Wert eines Tests kann somit deutlich beeinträchtigt werden, was konkret bedeutet, dass Testteilnehmer den Test nicht hinreichend ernst nehmen und nicht ihr wahres Leistungspotenzial zeigen.

In der Tat scheint es so, dass die Augenscheinvalidität der HSK von westlichen Lernern als eher gering eingeschätzt wird und dies einer der Gründe dafür ist, warum viele Chinesischlernende gar nicht erst an einer HSK teilnehmen.

3.3.4. Konstruktvalidität

Die Konstruktvalidität fragt danach, inwieweit Testergebnisse valide Indikatoren für die zugrunde liegenden theoretischen Konstrukte sind bzw. inwieweit das Verhalten der Kandidaten bei der Bearbeitung des Tests Rückschlüsse auf die nicht direkt beobachtbaren Fähigkeiten zulässt. Die Konstruktvalidität bezieht sich damit zunächst einmal auf die Operationalisierung des zu messenden Konstrukts in Form von Testaufgaben und auf die durch die Testleistungen gezogenen Schlussfolgerungen über die Fähigkeiten und Fertigkeiten des Testteilnehmers. Beispiel: Inwieweit lässt der HSK-Leseverständnisteil Rückschlüsse auf das theoretische Konstrukt "Lesekompetenz" zu. Die Konstruktvalidität wird mittlerweile von einigen Autoren als ein den anderen Validitätsarten übergeordnetes Konzept gesehen. (Messick 1989) Zudem wird diese Validität auf den Gebrauch von Testergebnissen (z. B. in Form von Entscheidungen über einzelne Kandidaten), Wertimplikationen und soziale Konsequenzen (*washback effect*) ausgedehnt. Ein extremer *washback effect* der HSK kann in Südkorea beobachtet werden, wo das HSK-Ergebnis in vielen Fällen Einfluss auf die Stellenvergabe bei Firmen hat oder für Stipendienentscheidungen mit herangezogen wird, was den Chinesischunterricht in Südkorea massiv beeinflusst. (Nie H. 2007:87)

[28] Bei einem C-Test wird in mehreren kurzen Texten (aus 60 bis 80 Wörtern bestehend) beginnend mit dem zweiten Wort des zweiten Satzes in jedem Text bei jedem zweiten Wort die zweite Hälfte getilgt. (Grotjahn 2003:57)

Häufige Ursachen für Invalidität hinsichtlich der Konstruktvalidität eines Tests sind u. a. Unterrepräsentation eines zu messenden Konstrukts, d. h. der Test ist zu eng gefasst und lässt wichtige Dimensionen des Konstrukts unberücksichtigt. So kann man bei der *Chu-zhongdeng*-HSK sagen, dass Aussagen über produktive Kompetenzen wie Sprechen oder das Verfassen von Texten anhand der Testergebnisse nur in sehr begrenztem Umfang möglich sind.[29] Eine weitere Ursache für Invalidität in dieser Hinsicht wird als konstruktirrelevante Varianz bezeichnet (vgl. Abschnitt 3.4.). Sie liegt dann vor, wenn Merkmale, die nichts mit der zu messenden Fähigkeit zu tun haben, eine Aufgabe für bestimmte Personen(-gruppen) systematisch erleichtert oder erschwert. Hierzu gibt es ebenfalls zahlreiche Beispiele aus der HSK. So sollte in einem Item der *Chu-zhongdeng*-HSK-Prüfung die Lücke in dem Satz "在中国的医院看病，首先需要□号。" gefüllt werden. Allerdings braucht man zur Lösung dieser Aufgabe das Hintergrundwissen, dass in chinesischen Krankenhäusern eine Nummer gezogen werden muss, wenn man zum Arzt geht. Die Lösung, das Schriftzeichen *gua* 挂, gehört zwar zur häufigsten Wortklasse der HSK (*jia* 甲), der Aufgabenkontext erschwert das Item jedoch erheblich. Ren (2002a:66) klassifiziert die eben beschriebene Aufgabe als invalide, sagt aber gleichzeitig, dass die HSK nur über äußerst wenige Aufgaben mit solchen konstruktirrelevanten Varianzen verfüge. Letztere werden auch als so genannte Bias (Verzerrungen) bezeichnet. Weiteres Beispiel für Konstruktvalidierung: Jing kritisiert, dass beim Hörverstehen Lesekompetenz mitgetestet wird, da die Antwortvorgaben in Schriftzeichen vorliegen statt in Hanyu Pinyin. Legt man die Hypothese zugrunde, dass die Chinesischkompetenz mit zunehmender Lerndauer ansteigt, kann die Untersuchung des Autors des vorliegenden Artikels auch der Konstruktvalidierung zugeordnet werden.

3.4. Fairness

Ein Test oder Item gilt in der Testtheorie dann als fair, wenn er/es bestimmte Gruppen von Kandidaten nicht systematisch aufgrund von Faktoren, die mit dem zu messenden Merkmal in keiner inhaltlichen Beziehung stehen (aufgrund konstruktirrelevanter Varianz, vgl. Abschn. 3.3.4), benachteiligt. Hat z. B. bei einem Hörverstehenstest die Sitzposition des Kandidaten (Entfernung zur Tonquelle) einen Einfluss auf die Hörverstehensleistung, ist der Hörverstehenstest als "unfair" einzustufen. Weitere Gründe, die zu unfairen Aufgaben führen, sind beispielsweise: systematische Benachteiligung aufgrund von Geschlecht oder Muttersprache, Verzerrung aufgrund von unterschiedlichem Hintergrundwis-

[29] U. a. Sun et al. (2007:128), Jing (2004:22-32).

sen[30] oder differentielle Effekte (sich unterschiedlich auswirkende Effekte) bestimmter Aufgabenformate[31], wie etwa ein zu kleiner Schrifttyp. Eine systematische Benachteiligung von Probanden wird, wie oben erwähnt, Bias genannt.

Beispiel der HSK-Forschung: Jing Cheng (2004:25, 29) ist der Ansicht, dass die HSK zu viel Wert auf Schriftzeichen lege, was ihrer Ansicht nach japanische Testteilnehmer (und teilweise koreanische) beim Teil Hörverstehen systematisch bevorteile, da die Antworten nur in Schriftzeichen vorliegen und somit Lesekompetenz die Performance in diesem Testteil beeinflusst.

3.5. Normierung/Standardisierung

Die Normierung ist die Eichung eines Tests für eine bestimmte Zielpopulation. Sie wird auch als Standardisierung bezeichnet (nicht zu verwechseln mit der Standardisierung in Abschnitt 3.1). Zielpopulationen können beispielsweise alle Schüler einer bestimmten Klassenstufe oder alle Studienanfänger in einem akademischen Fach sein etc. Die Eichung erfolgt mit Hilfe statistischer Verfahren anhand einer repräsentativen Stichprobe (Eichstichprobe) aus der Zielpopulation (Eichpopulation), woraus dann die Normwerte abgeleitet werden. Liegen Normwerte vor, kann der individuelle Testwert eines Testteilnehmers relativ zu den Leistungen der Zielpopulation betrachtet werden und nicht nur relativ zu den Leistungen anderer Kandidaten. Die Normierung ist für die Diagnose interindividueller Unterschiede innerhalb einer Gruppe unwichtig. Zudem ist das Gütekriterium Normierung unabhängig von den anderen Gütekriterien.

Der Begriff "standardisierter Test" wird von verschiedenen Autoren unterschiedlich aufgefasst. Viele halten die Normierung für das entscheidende Merkmal eines standardisierten Tests, andere Autoren die angemessene Standardisierung von Durchführung, Auswertung und Interpretation. Zusätzlich wird häufig eingefordert, dass ein standardisierter Test die Hauptgütekriterien Objektivität, Reliabilität und Validität in hinreichendem Maße erfüllt. Standardisierte Tests werden oft auch als formelle Tests[32] bezeichnet.

Die HSK ist eine normierte Prüfung (*biaozhunhua kaoshi* 标准化考试), was man u. a. daran sehen kann, dass das Ergebnis eines Prüflings in Bezug zu einer

[30] Beispielsweise wurde untersucht, inwieweit Kandidaten, die inner- und außerhalb Chinas an der HSK teilnehmen, unterschiedlich abschneiden. Hier wurden keine nennenswerten Unterschiede festgestellt. (Ren 2002b:69-74)

[31] Ob eine systematische Benachteiligung vorliegt, hängt von der Definition der Konstruktvalidität ab. Differentielle Effekte bezeichnet man als *differential item functioning*. Ob bei Items Testfairness vorliegt, wird häufig mit einer "DIF-Analyse" untersucht.

[32] Formelle Tests sind das Ergebnis langwieriger Bemühungen von Testspezialisten. Informelle Tests sind meist weit weniger aufwändige Produkte von Unterrichtspraktikern. Ihnen fehlt die Normierung an einer repräsentativen Stichprobe, was für Tests im Unterricht jedoch keinen Nachteil bedeutet. Für informelle Tests sind die Gütekriterien Situationsvalidität und Inhaltsvalidität entscheidend.

Normgruppe gesetzt wird. Dies kann man auch dem HSK-Ergebnis-Ausdruck (*chengjidan* 成绩单) entnehmen, der angibt, in welchem Prozentbereich zur Normgruppe sich der Prüfling befindet. Angaben zur Zusammensetzung der Normgruppe konnte der Autor dieses Beitrages bislang noch nicht ausfindig machen.

3.6. Authentizität

Beim Gütekriterium der Authentizität kann hinsichtlich der Authentizität der Vorgaben (z. B. Texte in einem Leseverstehenstext), der Authentizität der Testsituation und der Authentizität der Items unterschieden werden. Authentisch kann bedeuten, dass es sich um genuine und nicht spezifisch für den Test konzipierte Aufgaben handelt. Authentizität kann sich auch auf den Grad der Übereinstimmung zwischen den Merkmalen einer Testaufgabe und den Merkmalen einer zielsprachlichen Aufgabe (in der Realität) beziehen. Authentizität ist u. a. nach Bachman/Palmer (1996) ein wichtiges Gütekriterium, da authentische Aufgaben zum einen Generalisierungen im Hinblick auf die Fähigkeit zur Lösung zielsprachlicher Probleme außerhalb der Testsituation erlauben, und zum anderen hat die Authentizität einen sehr großen Einfluss darauf, für wie relevant der Testteilnehmer die Aufgaben hält.

Sun (2007:135) gibt an, dass Authentizität ein zentrales Prinzip der HSK sei, auf das man gleich von Beginn der HSK-Entwicklung an Wert gelegt habe. Leider spezifiziert er diese Aussage nicht weiter. Vermutlich bezieht er sich vorrangig auf die Inhalte der HSK und nicht auf die Item-Formate.

3.7. Nützlichkeit

Bachman/Palmer (1996:25-26) legen ein Gesamtkonzept für das Qualitätsmerkmal eines Tests vor: die Nützlichkeit. Sie setzt sich aus sechs komplementären Eigenschaften zusammen. Daraus ergibt sich:

Nützlichkeit = Reliabilität + Konstruktvalidität + Authentizität + Interaktivität + Rückwirkungseffekt + Praktikabilität

In diesem Modell treten nun Interaktivität (engl. *interactiveness*), Rückwirkungseffekt (engl. *impact*) und Praktikabilität zu den bisher oben genannten Kriterien hinzu. Interaktivität gibt das Ausmaß und die Art der Wechselwirkung zwischen Testaufgaben und den bezogen auf das zu messende Konstrukt relevanten kognitiven Merkmalen des Testteilnehmers an. Bei einem (realen) mündlichen Interview etwa weist das Merkmal Interaktivität einen höheren Grad auf als bei einem simulierten mündlichen Interview. Bei Ersterem kann der Kandidat den Gesprächsverlauf mitbestimmen, bei Letzterem muss der Kandidat auf die Stimuli, die in fester Reihenfolge über einen Tonträger präsentiert werden,

reagieren. Mit Rückwirkungseffekt meinen Bachman/Palmer den Einfluss des Tests auf eine Mikro- (einzelne Testkandidaten) und Makroebene (Erziehungssystem, jeweilige Gesellschaft). Damit fallen der Gebrauch der Testergebnisse in Form von Entscheidungen oder auch Wertimplikationen und soziale Konsequenzen (*washback*) bei Bachman/Palmer unter *impact*. Messick (1989) rechnet diese Aspekte der Validität zu. So wurde der negative Einfluss der *Chu-zhongdeng*-HSK auf den Unterricht von diversen Autoren kritisiert ("teaching for the test"; Niu 2003:45f.; Wang 2004:96-98).

Hervorzuheben sind auch die so genannten drei Prinzipien für die Sicherung der Nützlichkeit eines Tests: 1) Die Gesamtnützlichkeit eines Tests ist zu maximieren, nicht einzelne Komponenten der Nützlichkeit. 2) Die Komponenten dürfen nicht unabhängig betrachtet werden, sondern nur hinsichtlich ihrer kombinierten Wirkung. 3) Der Grad der Nützlichkeit bzw. die "richtige" Ausbalancierung der Teilkomponenten kann nicht allgemein festgelegt werden, sondern muss in Abhängigkeit von der spezifischen Testsituation bestimmt werden.

Grotjahn (2003:31) hält das Gesamtkriterium der Nützlichkeit für "ein sehr sinnvolles Qualitätskriterium, […] das auch die Rückwirkung des Tests auf den Unterricht und die Praktikabilität des Verfahrens berücksichtigt." Das Konzept der Nützlichkeit ist in der Sprachstandsmessung mittlerweile weit verbreitet und akzeptiert.

4. Untersuchung zur Relation von Lerndauer bzw. Unterrichtsaufwand zu HSK-Testergebnissen

4.1. Zielsetzung und Hintergrund der Untersuchung

Viele Lernende des Chinesischen interessiert nach einem gewissen Zeitraum des Chinesischstudiums, wo sie sich mit ihren Sprachkenntnissen einzuordnen haben. Die HSK ist, wie eingangs erläutert, eines der wenigen Instrumentarien, das deutschen Lernern für die Messung der Chinesischkompetenz zur Verfügung steht. Um an dem geeigneten Testformat (*Threshold/Rumenji* 入门级; *Basic/Jichu* 基础; *Elementary-Intermediate/Chu-zhongdeng* 初、中等; *Advanced/Gaodeng* 高等)[33] teilzunehmen, gibt das *Hanban* Richtwerte in der Maßzahl "Unterrichtseinheiten" (*xueshi* 学时) an, die grob anzeigen, mit wie viel investiertem Unterrichtsaufwand man in etwa an welcher Teststufe als Kandidat – voraussichtlich mit Erfolg – teilnehmen kann. Allerdings sind diese Richtwerte nicht weiter spezifiziert oder relativiert. Sie geben keinen Hinweis darauf, ob sich die Stundenzahl des Chinesischunterrichts auf den Unterricht inner- oder außerhalb Chinas bezieht, oder darauf, welchen muttersprachlichen Hintergrund der Lernende hat. Die Lerneinheit ist hinsichtlich des zeitlichen Umfangs zudem

[33] Der *Threshold* kann in Deutschland bis jetzt noch nicht abgelegt werden.

nicht definiert.[34] So setzen offizielle Zahlen des *Hanban* für die Stufe 6 (*Intermediate Certificate C*/*Zhongdeng C* 中等 C), der Stufe, ab der man alle Fächer ohne Ausnahme in Bachelor-Studiengängen (*benke* 本科) studieren darf, 1.200 bis 1.600 Stunden Chinesischunterricht voraus. Für die HSK-(Niveau-)Stufe 3 (*Basic Level A*/*Jichu A* 基础 A bzw. *Elementary C*/*Chudeng C* 初等 C) werden 400 bis 800 Unterrichtseinheiten veranschlagt.[35] Vorrangiges Ziel dieser Untersuchung ist es daher herauszufinden, wie viele Lerneinheiten Chinesischunterricht mit welcher HSK-Stufe bzw. erzielten Punktzahl bezogen auf Testteilnehmer deutscher Muttersprache (ohne muttersprachliche Vorkenntnisse des Chinesischen) korrespondieren.

4.2. Zentrale Fragestellung und weiterführende Fragen

Aus dem oben formulierten Forschungsansatz ergibt sich die folgende zentrale Fragestellung:

> Mit welchem Unterrichtsaufwand erzielt ein Chinesischlerner deutscher Muttersprache (ohne muttersprachliche Vorkenntnisse des Chinesischen) bei der *Chuzhongdeng*-HSK (bzw. *Jichu*-HSK) welches Ergebnis?

Daran schließt sich die Frage an, inwieweit die Angaben des *Hanban* für deutsche Lerner zutreffen oder nicht. Die Definition für Unterrichtsaufwand lautet: Unterrichtsaufwand beinhaltet in dieser Untersuchung den institutionalisierten Chinesischunterricht, den die/der Befragte in seinem Chinesischlernprozess bis zum Zeitpunkt der Befragung erhalten hat. Institutionen können sein: (Privat-) Schulen (inner-/außerhalb Chinas[36]), Hoch- und Fachhochschulen (inner-/außerhalb Chinas). Der Unterrichtsaufwand wird in der Einheit "Unterrichtseinheit" (UE) gemessen. Eine UE bezieht sich auf eine "Unterrichtsstunde", also in Deutschland in der Regel 45 Minuten bzw. in der VR China 50 Minuten. Weiterführende Fragen der Untersuchung sind:

---

[34] In der Regel muss ein Testteilnehmer mit deutscher Muttersprache einen wesentlich höheren Aufwand betreiben, um dieselbe Niveaustufe (*dengji fenshu* 等级分数) zu erreichen wie ein Kandidat mit japanischer Muttersprache. Das Handbuch zur HSK für Prüflinge (Zhongguo Hanyu Shuiping Kaoshi Weiyuanhui 2006) empfiehlt 100 bis 800 reguläre Unterrichtsstunden (UE) modernes Chinesisch für die Teilnahme an der Elementarstufen-HSK (*Jichu* 基础) und 400 bis 2000 UE für die Teilnahme an der Mittelstufen-HSK (*Chu-zhongdeng* 初、中等).

[35] Xie (1995:74). HSK 2005, zit. nach Guder (2007:21). Die Werte für die jeweiligen Niveaustufen (*dengji fenshu* 等级分数) konnten nach eigener Recherche (2008) in den aktuellen offiziellen Verlautbarungen des *Hanban* und des HSK-Zentrums der BLCU nicht mehr ausfindig gemacht werden, was bedeuten könnte, dass die alten Zahlen revidiert oder der Öffentlichkeit nicht mehr zugänglich gemacht werden sollen.

[36] China einschließlich der VR sowie Hongkong und Taiwan.

- Wer nimmt in Deutschland an der HSK teil? (Grundgesamtheit)
- Als wie leicht oder schwierig beurteilen HSK-Teilnehmer in Deutschland die einzelnen HSK-Testteile[37]?
- Wie angemessen bzw. geeignet finden die Kandidaten die Messung ihrer Chinesisch-Kompetenz durch die HSK?
- Was würden HSK-Teilnehmer an der HSK verändert haben wollen?
- Warum nehmen deutsche Chinesischlerner an der HSK teil?

4.3. Grundgesamtheit und Stichprobe

Die Grundgesamtheit sind alle Prüflinge der HSK in Deutschland, die im Oktober 2007 und im Mai 2008 an der HSK teilnahmen.[38] Die Kandidaten der HSK wurden nach den Prüfungen gefragt, ob sie bereit wären, an einer Befragung über die HSK teilzunehmen. Im Oktober 2007 wurden Kandidaten in Berlin und Hannover befragt, im Mai 2008 wurde die Befragung auf Erlangen und Hamburg ausgedehnt. In einer zweiten Befragung wurden die Befragungsteilnehmer nach Erhalt ihrer HSK-Ergebnisse darum gebeten, ihr Resultat mitzuteilen. Insgesamt nahmen 95 Probanden[39] an den Befragungen unmittelbar nach den Prüfungen teil, was in etwa der Hälfte aller Testteilnehmer entspricht, 44 gaben ihr Prüfungsergebnis an.[40] Die Stichprobe ist damit hinsichtlich der Gesamtverteilung aller Prüflinge nicht repräsentativ, da Kandidaten, die u. U. ein schlechtes Prüfungsergebnis befürchteten, eine Teilnahme an der Befragung ablehnten. Aufgrund der Hauptfragestellung dieser Untersuchung nach dem Zusammenhang von Lerndauer/Unterrichtsaufwand zum HSK-Testergebnis kann diese Problematik jedoch zunächst vernachlässigt werden. Allerdings sollte die genaue Zusammensetzung der Grundgesamtheit noch untersucht werden, damit die gewonnenen Daten nachgeglättet werden können. Unter die 44 Teilnehmer, die ihr Prüfungsergebnis angaben, flossen neun Teilnehmer mit muttersprachlichen Vorkenntnissen in Chinesisch ein. Die hier veröffentlichten Ergebnisse beziehen

[37] Gemeint sind hier Hörverständnis (*tingli* 听力), Grammatik (*yufa* 语法), Leseverständnis (*yuedu* 阅读) und Lückentest (*zonghe tiankong* 综合填空). Kaden (2004) spricht bei den einzelnen Testteilen von "Komplexen".

[38] Bis jetzt konnte ich keine repräsentative Zusammensetzung der Grundgesamtheit ermitteln. Anfragen bei mehreren Testadministratoren liegen vor.

[39] Die Befragung wurde im Oktober 2008 fortgesetzt. Allein an dieser Befragungsrunde nahmen weitere 46 HSK-Prüflinge teil, 20 Befragte gaben ihr HSK-Ergebnis an.

[40] Kandidaten, die die HSK in Deutschland ablegen, erhalten ihr Ergebnis zwei bis drei Monate später (mittlerweile kann das Ergebnis zusätzlich schon eher im Internet eingesehen werden). Bei der ersten Befragung wurden die Probanden gebeten, ihre E-Mail-Adresse zu hinterlassen (was ca. 95 % taten), so dass die Befragten nach Erhalt ihrer Ergebnisse angeschrieben werden konnten. Testergebnisse der Befragungsteilnehmer letztendlich zu erhalten, ist jedoch nur mit viel Geduld und einem gewissen Maß an Hartnäckigkeit möglich.

sich somit auf 35 Kandidaten und müssen wegen der Gesamtzahl der Probanden vorsichtig interpretiert werden.

4.4. Untersuchungsmethode und Operationalisierung

Die Untersuchungsmethode bestand in einer zweistufigen Befragung. In einem ersten Schritt füllten Kandidaten freiwillig nach der HSK-Prüfung einen Fragebogen aus, auf dem sie u. a. darum gebeten wurden, ihre E-mail-Adresse anzugeben. Über die Adresse wurden sie in einem zweiten Schritt nach Erhalt ihrer Prüfungsergebnisse kontaktiert, um so ihr Ergebnis mitzuteilen. Die Zuordnung von Prüfungsergebnis zu Fragebogen erfolgte über die E-mail-Adresse.

Zur Operationalisierung der Konstrukte Lerndauer und Unterrichtsaufwand: Die reine Lerndauer wurde über die Angabe der Anzahl der Jahre bzw. des Jahres, in dem die/der Befragte begann, Chinesisch zu lernen, erfasst. Diese Daten sollten relativ verlässlich sein.[41] Problematischer erweist sich die Variable Unterrichtsaufwand. Hier wurde eine Reihe von Annahmen über den institutionalisierten Chinesischunterricht formuliert. Im Folgenden wird kurz dargestellt, wie viele Unterrichtseinheiten (UE) Chinesischunterricht einem Schul- bzw. Hochschuljahr in Deutschland oder China im Mittel ungefähr entsprechen. Die Befragten wurden im Fragebogen zusätzlich direkt gefragt, wie viel Chinesischunterricht insgesamt sie bis zum Zeitpunkt der Befragung erhalten hatten. Wie zu erwarten war, machten hier nur wenige Befragungsteilnehmer Angaben. Daher musste die Variable Unterrichtsaufwand zusätzlich indirekt gemessen werden.

1) Chinesisch an Sekundarschulen in Deutschland. Überlegung:

3 UE /Woche	→	35 Wochen x 3 UE	→	105 UE (≅ 1 Schuljahr) .

Eine Unterrichtswoche Chinesisch kann mit durchschnittlich etwa drei Unterrichtseinheiten veranschlagt werden. Bei (gemittelt) 35 Wochen Unterricht pro Schuljahr ergibt das 105 UE Chinesischunterricht pro Schuljahr.

2) Chinesisch an Hochschulen in Deutschland. Überlegung:

8 UE /Woche	→	30 Wochen x 8 UE	→	240 UE (≅ 1 Studienjahr) .

Eine Unterrichtswoche Chinesisch kann mit durchschnittlich etwa acht Unterrichtseinheiten veranschlagt werden. Bei (gemittelt) 30 Wochen Unterricht pro Studienjahr ergibt das 240 UE Chinesischunterricht.[42]

[41] Einschränkungen lägen vor, wenn die Befragten systematisch gelogen hätten oder wenn sie sich nur schlecht oder ungenau erinnert hätten.

[42] Laut eines Referats mit dem Titel "The relations between Chinese language learning and Sinology in German universities", gehalten von Franziska Trempler an der FU Berlin im

3) Chinesisch an Universitäten in China/auf Taiwan. Überlegung:

20 UE /Woche	→	45 Wochen x 20 UE	→	900 UE (≅ 1 Studienjahr) .

Eine Unterrichtswoche Chinesisch kann mit durchschnittlich etwa zwanzig Unterrichtseinheiten veranschlagt werden. Bei (gemittelt) 45 Wochen Unterricht pro Studienjahr ergibt das 900 UE Chinesischunterricht pro Studienjahr.

4) Schüleraustausch. Überlegung: Für einen Schüleraustausch in China wurden 350 UE veranschlagt, womit ein Auslandsjahr an einer chinesischen Schule nur ein Drittel von der Gewichtung erfährt, das einem (zumindest relativ vergleichbaren) Studienjahr in China entspricht. Die Zahl wurde bewusst niedrig angelegt, da die UE in diesem Fall nur äußerst schwierig einzuschätzen sind. Mit 350 UE soll die tatsächliche Anzahl der UE eher unterschätzt werden.

Es ist völlig klar, dass die hier vorgestellten Berechnungsgrundlagen eine Reihe von Problemen in sich bergen. Die Werte sind daher so gewählt, dass sie im Zweifel eher unter- als überschätzen. Es muss darum betont werden, dass die bisherigen Ergebnisse bezogen auf den Unterrichtsaufwand mit großer Vorsicht zu betrachten sind und weiterer empirischer Forschung bedürfen. Bei der Relation von Lerndauer zu HSK-Ergebnis sollten die Ergebnisse hingegen relativ stichhaltig ein, müssen allerdings hinsichtlich der Anzahl der Probanden ebenfalls weiter ausgebaut werden.

4.5. Bisherige Ergebnisse

Abbildung 1 zeigt das Verhältnis von HSK-Ergebnis (erreichte Niveaustufe) zur vorangegangenen Dauer des Chinesischlernens in Jahren. Die Ergebnisse von 35 Befragungsteilnehmern, die keine muttersprachlichen Vorkenntnisse in Chinesisch haben und ihr HSK-Ergebnis mitteilten, sind darin erfasst. Die HSK-Stufe 3 wurde von zwei Befragten nach 1,5 bzw. 2 Jahren erreicht, von zwei weiteren nach 4 bzw. 5,5 Jahren. Lässt man den Ausreißer (12 Jahre) unberücksichtigt, geht der Trend im Mittel dahin, Stufe 3 nach etwa drei Jahren Lerndauer zu erreichen. Bis auf drei Ausreißer schaffte es kein Testteilnehmer, innerhalb von fünf Jahren über die Stufe 5 hinaus zu kommen. Mit anderen Worten: Die Stufe 6 (oder höher) weist eine Lerndauer von 7,5 bis 10,5 Jahren auf. Die beiden Lerner mit der längsten Lerndauer gaben 16 bzw. 23 Jahre an. Bei ihnen handelt es sich um Selbstlerner. Die Korrelation zwischen Lerndauer und HSK-Ergebnis beträgt 0,36 (nach Pearson).

Kurs "Aspects of Teaching Chinese as a Foreign Language" (Kursleiter: Prof. Hsin Shih-chang; SS 2008), entspricht das arithmetische Mittel der UE für Chinesisch an deutschen Universitäten in Bachelor-Studiengängen 578 UE (bei drei Studienjahren; Stand: 2008).

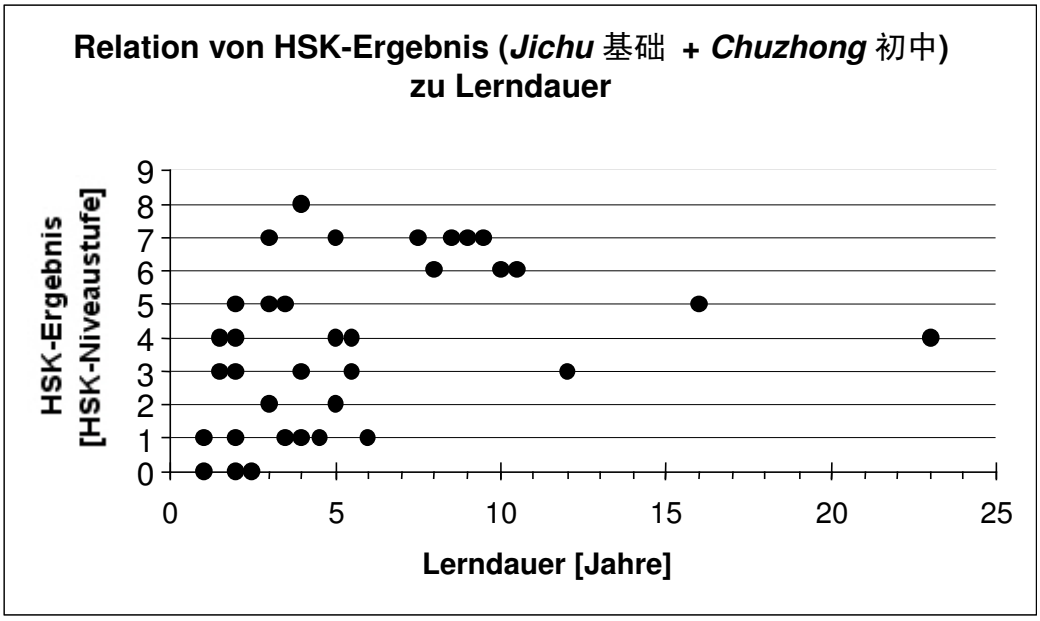

Abbildung 1: Verhältnis von HSK-Ergebnis zu Lerndauer

Abbildung 2 stellt die erreichte HSK-Niveaustufe in Abhängigkeit von der Zahl der Unterrichtseinheiten dar. Interessant ist, dass hier HSK-Stufe und Unterrichtsaufwand viel stärker miteinander korrelieren als in Abbildung 1. Der Korrelationskoeffizient liegt bei 0,79 (nach Pearson). Damit ist der Zusammenhang zwischen Unterrichtsaufwand und HSK-Ergebnis viel eindeutiger als der zwischen Lerndauer in Jahren und HSK-Ergebnis, da in diesem Fall "Langzeitlerner" das Ergebnis nicht so sehr verzerren.

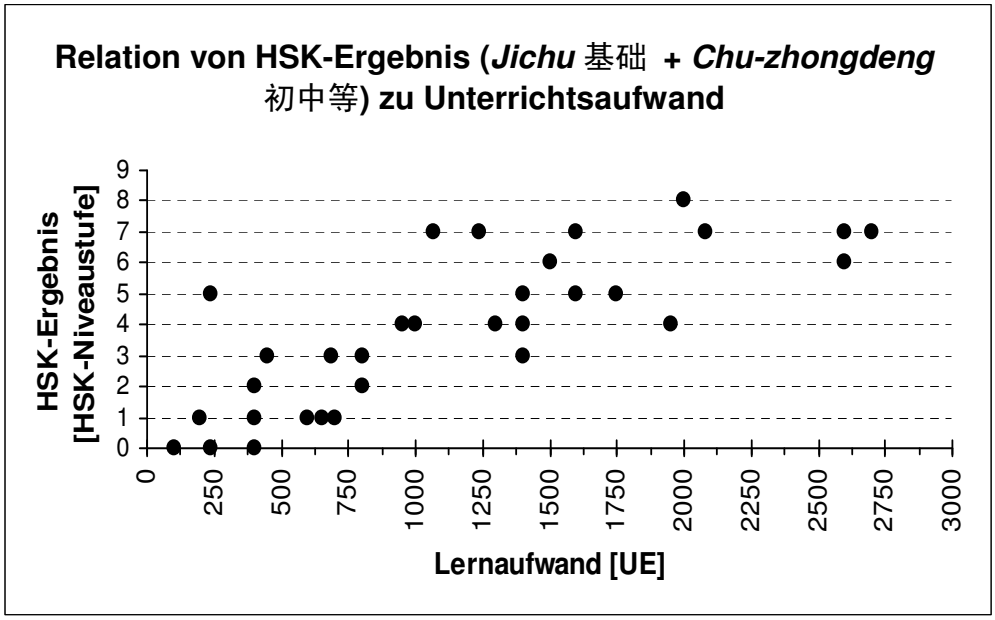

Abbildung 2: Verhältnis von HSK-Ergebnis zu Unterrichtsaufwand

Bis etwa 2000 UE kann man sogar einen linearen Zusammenhang erkennen. Bei drei Befragten konnten aus den im Fragebogen gemachten Angaben keine Rückschlüsse auf die Menge des erhaltenen institutionalisierten Chinesischunterrichts gezogen werden, was u. a. daran liegen könnte, dass die Befragten kaum oder u. U. gar keinen institutionalisierten Chinesischunterricht erhalten hatten.[43]

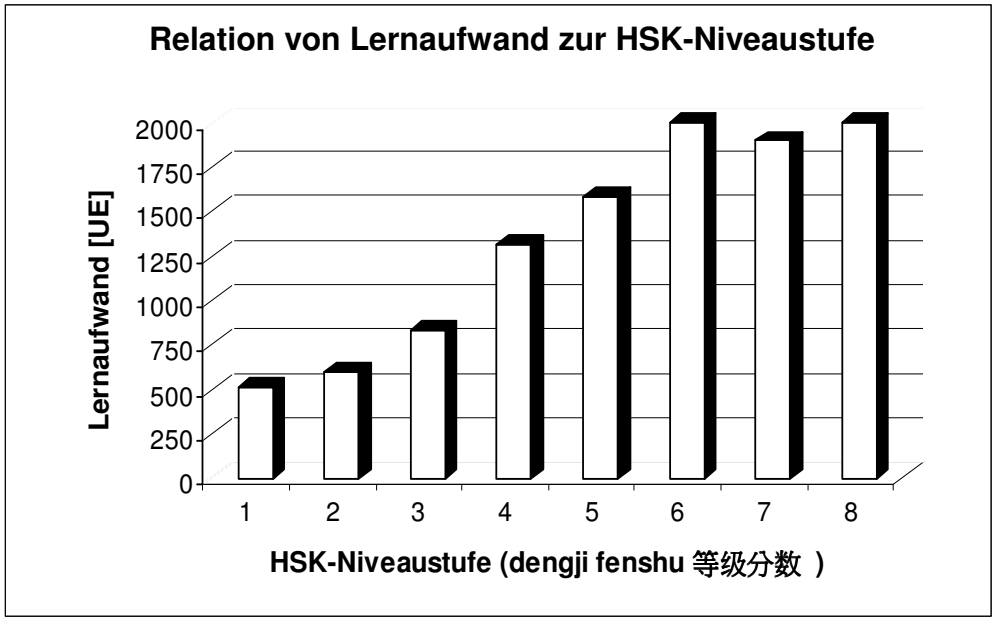

Abbildung 3: Verhältnis von Lernaufwand zur erreichten HSK-Stufe

Für Abbildung 3 wurde aus den Unterrichtseinheiten (UE) aller Kandidaten innerhalb einer HSK-Stufe das arithmetische Mittel der UE errechnet und in einem Säulendiagramm dargestellt. Stufe 6 fällt leicht aus dem Rahmen, beruht aber auch auf den Daten von nur zwei Befragten. Stufe 8 basiert sogar auf der Angabe nur eines Probanden. Um die *Jichu*-HSK zu bestehen, sind demzufolge 515 UE nötig, laut *Hanban* sollen schon 100 UE genügen. Die Stufe 3 erreicht man in etwa mit einem Unterrichtsaufwand von 835 UE. Zum Studium aller Fächer sind den Daten der vorliegenden Untersuchung zufolge bereits 2.000 UE nötig (*Hanban*-Angabe: 1.200-1.600 UE). Aus den in den Abbildungen 1 bis 3 zusammengefassten Daten ergibt sich die folgende Tabelle (Tabelle 1). Sie stellt die HSK-Testformate,[44] die zum Erreichen jeweils notwendige Anzahl von Unterrichtseinheiten lt. *Hanban*-Angaben sowie die Ergebnisse der vorliegenden Untersuchung und die Kompetenzstufen des Gemeinsamen europäischen Refe-

[43] Zwei Befragte vermerkten auf ihrem Fragebogen "Selbststudium".

[44] Nicht enthalten sind der *Business Chinese Test (BCT/Shangwu Hanyu Kaoshi* 商务汉语考试; auch "Wirtschafts-HSK"), der *Youth Chinese Test (YCT/Zhong-Xiaoxuesheng Hanyu Kaoshi* 中小学生汉语考试) und die reformierte HSK – *Gaijinban* (改进版). Zu diesen Tests siehe Guojia Hanban (o. J.). Der *YCT* ist für Schüler gedacht, die außerhalb Chinas Chinesisch lernen und keine chinesischen Muttersprachler sind.

renzrahmens (GER) gegenüber. Die Gegenüberstellung zum GER ist jedoch nicht unproblematisch, da dieser explizit auch produktive sprachliche Kompetenzen enthält, welche in den HSK-Testformaten *Jichu* und *Chu-zhongdeng* aber nicht direkt überprüft werden, was insbesondere für mündliche kommunikative Fertigkeiten und Schreibkompetenz gilt.

Testformat *kaoshi dengji* 考试等级 UE lt. *Hanban*	HSK-Stufe *dengji fenshu* 等级分数	Stufe *dengji* 等级	Untersu-chungs-ergebnis [UE]	GER-Stufe
Advanced *Gaodeng* 高等 > 3000	11	高等 A		C 2
	10	高等 B		
	9	高等 C		C 1[45]
Elementary-Intermediate *Chu-zhongdeng* 初、中等 400-2000 (Stufe 6: ≈ 1.500)	8	中等 A	≈ 2000	
	7	中等 B	≈ 1900	B 2
	6	中等 C	≈ 2000	
	5	初等 A	≈ 1580	B 1
	4	初等 B	≈ 1320	
	3	初等 C	≈ 835	
Basic *Jichu* 基础 100-800		基础 A		A 2
	2	基础 B	≈ 600	
	1	基础 C	≈ 515	A 1
Threshold *Rumenji* 入门级 80-300	[0]		≈ 250	

Tabelle 1: Gegenüberstellung von HSK-Testformaten, offiziell notwendigen UE, Niveaustufen, tatsächlich benötigten UE deutscher Lerner und GER-Stufen

5. Zusammenfassung der Ergebnisse

Die vorliegende Untersuchung weist mehrere Probleme auf:

1. Sie berücksichtigt nicht die Qualität des Unterrichts. Allerdings ist das Konstrukt "Qualität des Unterrichts" institutionen- bzw. länderübergreifend nicht objektiv zu erfassen. Zudem kann man argumentieren, dass es sich bei den Befragten um Lernende handelt, die über mehrere Jahre hindurch Chinesisch gelernt haben, und sich unterschiedliche Qualität des Unterrichts im Mittel teilweise wieder ausgleicht.

[45] Dass die GER-Stufe C1 mit den HSK-Stufen 8 und 9 zusammenfällt, wird auch durch den Aufsatz von Tao/Chen (2002) gestützt. Die Gegenüberstellung der GER-Kompetenzstufen zu den HSK-Stufen stammt von Guder (2007:21). Sie sollte mit Vorbehalt betrachtet werden, da sie noch weiterer Forschung bedarf.

2. Ausmaß und Qualität des Selbststudiums wurden nicht gemessen. Hier gilt ebenfalls: Die Variable "Selbststudium" ist bei einem Lerner, der seit mindestens etwa zwei Jahren Chinesisch lernt, in seiner zeitlichen Dimension kaum zu messen, hinsichtlich der Qualität oder Intensität vermutlich gar nicht.

3. Für Ausreißer in den Punkteverteilungen gilt: Hier ist eine Verzerrung durch eine nicht erfasste Variable nicht auszuschließen, wie z. B. ob Befragte mit chinesischen Muttersprachlern zusammenleben oder in häufigem Kontakt stehen, oder etwa über fortgeschrittene Japanischkenntnisse verfügen.

Auch wenn die in diesem Aufsatz vorgelegten Zahlen gerade wegen der relativ geringen Zahl an Befragten mit großem Vorbehalt betrachtet werden müssen,[46] so lassen sich jedoch erste Tendenzen ablesen:

1. Der Einstieg für die Niveaustufe 1 der HSK (*Basic C*/*Jichu C* 基础 C) ist für Lernende deutscher Muttersprache wohl höher anzusetzen als die angegebenen Zahlen des *Hanban* vermuten lassen. Eine Teilnahme ist vermutlich erst ab etwa 500 UE sinnvoll (man vgl. die UE der Kandidaten, die die *Jichu*-Prüfung nicht bestanden haben, in Tabelle 1 als Stufe "[0]" klassifiziert).

2. Für die Stufe 6 (*Intermediate C*/*Zhongdeng C* 中等 C), die zum Studium aller Fächer in der VR China berechtigt (in Bachelor-Studiengängen), benötigen deutsche Lerner wahrscheinlich ca. 2.000 UE und nicht "nur" 1.200 bis 1.600.

3. Ein durchschnittlicher deutscher Lerner braucht mit hoher Wahrscheinlichkeit in der Regel etwa fünf Jahre Chinesischstudium, um die Niveaustufe 5 (oder höher) zu erreichen, was jedoch nicht ausschließt, dass hoch motivierte und fleißige Lerner diese Stufe auch schon nach ca. drei Jahren erreichen können.

Abschließend zur Kritik an der HSK: 27 % aller Befragten gaben an, dass sie es gut fänden, wenn die HSK einen Sprechteil beinhalten würde, wobei dieser Wert auch so interpretiert werden kann, dass etwa drei Viertel aller Getesteten einen mündlichen Testteil für nicht wichtig halten. Immerhin 11,7 % der Befragten würden sich wünschen, dass die HSK auch in den Testformaten *Jichu* und *Chu-zhongdeng* Schreibkompetenz überprüfte. Vereinzelt wurde gefordert bzw. kritisiert, dass es mehr und längere Pausen und bessere bzw. sinnvollere Multiple-Choice-Antworten geben sollte, und dass die Aussprache extra getestet werden könnte.

Literaturverzeichnis

Bachman, Lyle. 2004. *Statistical analyses for language assessment*. Cambridge

Bachman, L. F./Palmer, A. S. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford

Bausch, Karl-Richard/Christ, Herbert/Krumm, Hans-Jürgen (Hg.). 2007. *Handbuch Fremdsprachenunterricht*. Tübingen

[46] Die Befragung wurde im Oktober 2008 und Mai 2009 fortgesetzt.

Chai, Xingsan 柴省三. 2002. 关于 HSK (初中等) 平行信度的实证研究 (Forschungen zum Nachweis der Paralleltestreliabilität bei der Chu-zhongdeng-HSK). In: 《汉语学习》 2002 年 4 月第 2 期, 65-69

Da, Jun. 2007. "Reading news for information: How much vocabulary a CFL learner should know". In: Guder/Jiang/Wan (Hg.), 251-278

Europarat. 2001. *Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin/München

Grotjahn, Rüdiger. 2003. Leistungsmessung und Leistungsbewertung. Studienbrief. Weiterbildungs-Masterstudiengang "Deutschlandstudien". Studienschwerpunkt: Deutsche Sprache und ihre Vermittlung. Fernuniversität Hagen

Guder, Andreas. 2007. "The Chinese writing system as third dimension of foreign language learning". In: Guder/Jiang/Wan (Hg.), 17-32

Guder, A./Jiang, Xin 江新/Wan, Yexin 万业馨 (Hg.). 2007. 汉字的认知与教学—西方学习者汉字认知国际研讨会论文集. *The cognition, learning and teaching of Chinese characters.* Beijing

Guojia Hanban (Hg.). (o. J.). 商务汉语考试 (BCT) (Business Chinese Test). URL: http://www.hanban.edu.cn/hanyukaoshi_more.php?ithd=swhyksbct (04.02.2009)

Guojia Hanban (Hg.) (o. J.). 中小学生汉语考试(YCT)简介 (Kurzvorstellung des Youth Chinese Test). URL: http://www.hanban.edu.cn/hanyukaoshi_more.php?ithd=kaosjs (04.02.2009)

Guojia Hanban (考试处 Prüfungssektion) (Hg.). 26.07.2007. 汉语考试发展简介 (Kurzvorstellung der Entwicklung der Chinesischprüfung) URL: http://www.hanban.org/content.php?id=2627 (26.03.2009)

Hoffmann, Jana N./Guder, Andreas. 2007. "'Une langue émergente' – Chinesischunterricht in Frankreich". In: *CHUN* 22, 187-195

Ingenkamp, K. 1997. *Lehrbuch der pädagogischen Diagnostik: Studienausgabe*. Weinheim

Jing, Cheng 竟成. 2004. 关于 HSK 若干问题的思考 (Überlegungen zu einigen Problemen der HSK). In:《暨南大学化文学院学报》2004 年第 1 期, 22-32

Kaden, Klaus. 2004. Prüfung zum Nachweis chinesischer Sprachkenntnisse 汉语水平考试 HSK. Elementarstufe, Grund- und Mittelstufe, Oberstufe, Dokumente. Stand 2004, übersetzt von Klaus Kaden für den Fachverband Chinesisch e. V.

Kultusministerkonferenz (Hg.). 2008. Chinesisch an Schulen in Deutschland. Bonn

Kühn, Ingrid/Lehker, Marianne/Timmermann, Waltraud (Hg.). 2005. *Sprachtests in der Diskussion*. Frankfurt

Lado, R. 1961. *Language testing. The construction and use of language tests*. London

Li, Aihua 李爱华. 2007. 汉语水平考试成来华留学生软肋 (Die HSK wird für nach China kommende Austauschstudenten zu einem wunden Punkt). In: 《科学日报》 2007 年 2 月 6 日. URL: http://www.science.net.cn/htmlnews/20072601147437311 6.html?id=3116 (29.06.2009)

Lienert, G. A./Raatz, U. 1994. *Testaufbau und Testanalyse*. Weinheim

Messick, S. 1989. "Validity." In: Linn, R. L. (Hg.). 1989. *Educational measurement*. New York, 13-103

Nie, Dan 聂丹. 2006. HSK(初、中等)再测信度验证 (Überprüfung und Bestätigung der Retestreliabilität der Chu-zhongdeng-HSK). In: 《中国考试》(研究版) 第 5 期, 43-47

Nie, Hongying 聂鸿英. 2007. 多元文化融合催生全球汉语热情况下语言教学的新思路 —对韩国人的 HSK 教学的思考 (Neue Gedankengänge zur Sprachdidaktik in einer Situation, in der multikulturelle Verschmelzungen zu einem globalen Chinesischfieber antreiben – Überlegungen zur HSK-Didaktik für Koreaner). In: 《东疆学刊》第 24 卷第 3 期, 86-91

Niu, Jing 牛静. 2003. 浅谈 HSK 与汉语教学 (Kurze Diskussion über HSK und Chinesischdidaktik). In: 《新疆广播电视大学学报》2003 年第 3 期, 45-46

Ren, Jie 任杰. 2002a. 在 HSK 考试中如何保证试题的公平性 (Wie bei der HSK-Prüfung die Fairness der Test-Items gewährleistet wird). In: 《汉语学习》2002 年 6 月第 3 期, 66-70

Ren, Jie 任杰. 2002b. 中国境内外 HSK 成绩公平性的分析 (Analyse zur Fairness von HSK-Ergebnissen inner- und außerhalb Chinas). In: 《语言教学与研究》2002 年第 5 期, 69-74

Rost, Jürgen. 1996. *Lehrbuch Testtheorie, Testkonstruktion.* Bern

Steering Committee for the Test Of Proficiency-Huayu (SC-TOP) (Hrsg). 18.02.2009. Total number of TOP test-takers. URL: http://www.sc-top.org.tw/english/report.php (12.03.2009)

Sun, Dejin 孙德金. 2007. 汉语水平考试（HSK）的科学本质 (Der wissenschaftliche Charakter der HSK). In: 《世界汉语教学》2007 年第 4 期, 129-138

Sun, Dejin 孙德金/Zhang, Kai 张凯/Guo, Shujun 郭树军/Wang, Jimin 王佶旻. 2007. 汉语水平考试（HSK）改进方案；北京语言大学汉语水平考试中心 "HSK 改进工作" 项目组 (Design der Weiterentwicklung der HSK; die Projektgruppe "Weiterentwicklungsarbeit HSK" des HSK-Zentrums der BLCU). In: 《世界汉语教学》2007 年第 2 期, 126-135

Tao, Liming 陶黎铭/Chen, Hong 陈宏. 2002. 汉语水平考试 (HSK) 等级结构中的几个系统理论问题 (Einige systematische theoretische Probleme in der Struktur der HSK-Niveaustufen). In: 《汉语学习》第 2 期, 58-64

Tian, Qingyuan 田清源. 2007. HSK 主观考试评分的 Rasch 实验分析 (Rasch-Experiment-Analyse zur subjektiven Punktvergabe der HSK). In: 《心理学探新》第 1 期, 65-69

Universität Trier. 2009. HSK (Hanyu Shuiping Kaoshi). URL: http://www.uni-trier.de/index.php?id=26379 (21.05.2009)

Volkshochschule Landeshauptstadt Düsseldorf. 2008. Programm für das 2. Halbjahr 2008. Düsseldorf

Vollmer, Helmut J. 2007. "Leistungsmessung, Lernerfolgskontrolle, Selbstbeurteilung: Überblick." In: Bausch/Christ/Krumm (Hg.) 2007, 365-370

Wang, Ying 王瑛. 2004. 试论 HSK 测试与现行教学存在的矛盾及思考 (Diskussion zur HSK-Prüfung und momentan in der Lehre existenten Widersprüchen und Überlegungen). In: 《新疆教育学院学报》2004 年 9 月, 96-98

Xie, Xiaoqing 谢小庆. 1995. 汉语水平考试的分数体系 (Punktvergabesystem der HSK). In: 张凯 Zhang K. 2006, 66-82

Yang, Chengqing 杨承青/Zhang, Jinjun 张晋军. 2007. 汉语水平考试(HSK)改革设想 (Überlegungen und Vorschläge zur Reform der HSK). In: 《语言文字应用》2007 年 8 月第 3 期, 107-112

Zhang, Kai 张凯 . 2006.《汉语水平考试 (HSK) 研究》 (HSK-Forschung). Beijing

Zhongguo Hanyu Shuiping Kaoshi Weiyuanhui 中国汉语水平考试委员会 . 2006. 2006 HSK 中国汉语水平考试 —考生手册 (HSK – Handbuch für Prüflinge). Beijing

Übersichtstabelle zu Fachtermini der Sprachstandsmessung

Augenscheinvalidität	face validity	biaomian xiaodu	表面效度
Authentizität	authenticity	zhenshixing	真实性
Differentielle Effekte	differential item functioning (DIF)	xiangmu gongneng chayi	项目功能差异
Fairness	fairness	gongpingxing	公平性
Gütekriterium	quality criterion	zhiliang biaozhun	质量标准
Hanban	Chinese Language Council International	Guojia Hanyu Guoji Tuiguang Lingdao Xiaozu Bangongshi	国家汉语国际推广领导小组办公室
HSK	HSK/Chinese Proficiency Test	Hanyu Shuiping Kaoshi	汉语水平考试
HSK-Testzentrum	Chinese Proficiency Test Center	Hanyu Shuiping Kaoshi Zhongxin	汉语水平考试中心
Inhaltsvalidität/ Kontentvalidität	content validity	neirong xiaodu	内容效度

Konstruktvalidität	construct validity	jiegou xiaodu/ gouxiang xiaodu	结构效度 构想效度
Kriterienbezogene Validität/ Empirische Validität	criterion validity/ Empirical validity	xiaobiao xiaodu	效标效度
Normierung/ Standardisierung	standardization	biaozhunhua	标准化
Nützlichkeit	usefulness	shiyongxing	实用性
Objektivität	objectivity	keguanxing	客观性
Paralleltestreliabilität	parallel test reliability	pingxing xindu	平行信度
Reliabilität	reliability	xindu/kekaoxing	信度/可靠性
SC-TOP	Steering Committee for the Test Of Proficiency	Guojia Huayu Ceyan Tuidong Gongzuo Weiyuanhui	國家華語測驗推動工作委員會
TOP	TOP Test Of Proficiency	Huayuwen Nengli Ceyan	華語文能力測驗
Universität für Sprache und Kultur Beijing	Beijing Language and Culture University (BLCU)	Beijing Yuyan Daxue	北京语言大学
Validität	validity	xiaodu	效度
Washback	washback effect	houxiao zuoyong	后效作用

摘要

语言测试和语言能力分级的标准化是一个世界范围内的趋势。同时，关于标准和能力的定义的讨论，对对外汉语教学的影响越来越深入。单从目前每年参加汉语水平考试（HSK）的考生达到了大约 16 万人，即可见一斑。本文首先以“汉语水平考试研究”中的例句阐明了语言测试传统理论的质量标准。这些质量标准包括客观性、可靠性及效度，并通过公平性、真实性、实用性和标准化予以补充。文章的第二部分展示了对汉语水平考试德国考生的调查研究的阶段性结果。调查旨在发现学习时间的长短以及实际授课学时与汉语水平考试成绩之间的关系。调查研究的两个主要结果显示，德语为母语的 HSK 考生通过第一级（基础 C）的考试，基本上需要大约 500 个学时。而要通过代表中国所有本科院校入学资格的第六级（中等 C）的考试，甚至需要将近 2000 个学时。因此，要让学生通过这些级别的考试，德国教师明显需要比国家汉办所给出的学时数 (基础 C：100 学时，中等 C：1500 学时)更多的教学时间。