Der RR-Sprechtest und Test-DaF Ⅱ

Rudolf Reinelt

Überblick

Verwendete Abkürzungen TD= TestDaF, RS= RR-Sprechtest

3．1．2．Weitere wichtige Kriterien

Einige weitere Kriterien sollen nicht außer Acht bleiben:

Finanzielles

TD hat eine feste wirtschaftliche Ausrichtung: Möglichst genau die Zielgruppe erfassen und die Betreffenden testen und dann weiterschicken（oder deren Weitergehen ermöglichen）. Dieser Prozess ist mit Einnahmen（für die Bereitstellung, Beurteilung und Benachrichtigung）und Ausgaben（Erstellung, Beurteilung, Benachrichtigung）verbunden und wirtschaftlich umsetzbar. Der RS wird am Semesterende durchgeführt, ist also Teil des Unterrichts, für den allerdings im Rahmen der Studiengebühren bei der Universität bezahlt wird. Weil es bei RS institutionell nicht um die Eröffung von Perspektiven geht, sondern nur um die Vermittlung, dass die Lerner/Studenten selbst einmal etwas geschafft haben, muss man natürlich möglichst viele Teilnehmer durchbringen, die bis zu dieser Prüfung durchgehalten und nicht zwischendurch aufgegeben haben.
Während TD also Geld kostet und einbringt, ist RS für die Lerner umsonst. Für die Beurteiler gibt es von der Ehime Universität ewa 900 Yen und vom Verfasser den Rest bis 1000 Yen. Bis vor kurzem musste der Verfasser alle Ausgaben selbst tragen. Für die Muttersprachler und die Aufsicht gibt es ebenfalls 1000 yen. Dies ist in jedem Fall eigentlich eine unglaublich billige Summe, aber der Verfasser dankt natürlich allen für die Mitarbeit zu diesem Preis. Für den Verfasser ist diese（Test）entwicklung ein horrendes Verlustgeschäft, zumal die meisten anderen Lehrer ohne derartige Prüfungen auskommen und trotzdem schneller befördert werden und entsprechend noch mehr Geld verdienen.

Praktikabiliät und Machbarkeit

Machbarkeitsgrenzen sollten ebenfalls nicht unerwähnt bleiben. Besonders unter

nicht-idealen Bedingungen wie bei RS kann ein Test immer nur so gut sein, wie es die Dominanz der Machbarkeit（Reinelt 2005）zulässt. Dazu gehören bei RS wie wahrscheinlich nicht nur in nicht-geschäftlichen Kontexten Probleme der folgenden Art:

- finanzielle Probleme der Entwicklung
- Zeitprobleme der Tester und Testnehmer
- Probleme mit der Kraft für die Entwicklung
- Resourcen und technische Probleme

and vielleicht erst viel später

- Motivationsprobleme, oder
- Testangst der Testteilnehmer.

In anderen Arbeiten bin ich auf diese Thematiken eingegangen und Lösungen sind nur bis zu einem bestimmten Grad möglich, so dass an dieser Stelle keine weiteren Ausführungen nötig sind.

Rückwirkung auf den Unterricht

Dieses normalerweise mit *Washback* bezeichnete Phänomen ist ein sehr interessantes Thema, besonders da in Japan früher sehr gut funktionierende *Senpai*（ältere Studenten（helfen）- *Kouhai*（jüngeren Studenten）Beziehungen bestanden. Wenn diese Netzwerke heute nicht mehr so funktionieren, kann dies den Einsatz dieses Tests erschweren oder erleichtern, was allerdings noch zu untersuchen wäre. So können ältere Studenten schon einmal im voraus sagen（wenn nicht gar üben）, was in dem Test kommen kann, oder wie er abläuft.

In letzter Zeit ist die enge Zusammenarbeit von Studenten innerhalb einer Klasse auffällig. Diese war lange aufgrund der vertikalen Struktur（s. letzter Abschnitt）nicht so einfach durchführbar. Jetzt scheint es aber immer mehr Teilnehmer an selbstorganisierten Treffen zu geben. Dabei wird anscheinend genau das Wesentliche am Fremdsprachenlernen, der schnelle sprachliche Austausch in der Fremdsprache, geübt. Jedenfalls

lassen dies die guten Gesprächsergebnisse in den letzten Jahren vermuten.

Stellung des Tests

Während es für den RS zumindest in Japan im Augenblick keinen Vergleich gibt, muss TestDAF mit einer ganzen Reihe anderer Tests konkurrieren und auch nicht alle Universitäten erkennnen TestDaF an. Dies macht die Konkurrenzsituation schwieriger, aber auch reichhaltiger, weil es immer neue Entwicklungen gibt. Dieser letzte Punkt gilt nicht so einfach für den RS, der schon fast am Ende der Entwicklung angekommen zu sein schien. Allerdings haben in den letzten zwei Jahren mehr Studenten ihren Willen erklärt, weiter Deutsch zu lernen. Dies heisst dann aber auch, dass eine Nachfolge zu dem bisherigen Test erarbeitet werden muss. Dieser gilt dann nicht mehr nur für den allgemeinbildenden Bereich am Anfang des Studiums, sondern muss idealerweise auf ein Weiterlernen in sogennnten *Hattenkamoku*（Weiterentwicklungskursen）im zweiten, dritten und vierten Studienjahr, die auch auf dem Abschlusszeugnis aufgeführt werden, und – wenn eben möglich – in Richtung auf eine（inter）nationale Anerkennung hin erstellt werden.

3.2. Praxis: Durchführung

3.2.1. Einleitung: High stakes vs. low stakes

Während TD einfach ein *High-Stakes Test*（HST）ist, ist die Situation beim RRSprechtest etwas anders. Beim letzteren handelt es sich erst einmal um einen *low stakes test*, allerdings ist es doch etwas komplizierter. Einerseits brauchen die Studenten gar keine zweite Fremdsprache an der Ehime Universität. Andererseits würde ein Durchfallen einen Verlust von Punkten für ein ganzes Jahr bedeuten. Dies können sich die meisten Studenten zwar leisten, weshalb es auch besonders im Sommersemester eine hohe Aussteigerrate gibt: wer aber bis zum Wintersemester mitmacht, hat meist schon soviel investiert, dass dieser Test dann besonders wichtig für die Person wird, er also dann

Merkmale von *high stakes* annimmt, besonders wenn man den möglichen Aufwand und Punkteverlust bei einem Durchfallen berücksichtigt.

3．2．2．Praktische Durchführung

3．2．2．1．Einleitung

Bei der konkreten Durchführung der beiden Tests gibt es auffallende Gemeinsamkeiten und Gegensätze, von denen hier nur die im Teil Sprechen erwähnt werden können. Dies geschieht vor dem Hintergrund, dass neben dem Leistungsziel in jedem Fall Fairness und Gleichheit die obersten Ziele beider Tests sind.

3．2．2．2．Format und Ablauf

TD（Daad 2004, p.27,）benutzt ein semi-direktes Testformat, ist weniger authentisch und testet hauptsächlich die Produktion. TD wird an einem Testzentrum durchgeführt. Die Aufgaben kommen von einer Kassette, können aber auch im Aufgabenbuch mitgelesen werden. Die Antworten sind auf Kassette zu sprechen und die Auswertung erfolgt zentral in Deutschland. Auffallend ist, dass es immer nur sehr wenig Zeit zum Überlegen gibt, eine bis drei Minuten, ebenso für die Antwort eine bis zwei Minuten. Die Aufgabentypen gehen von *Telefongespräch*（!）(1)，über *Informationen geben* bzw. *erfragen* oder *um etwas bitten* bzw. *jemanden überzeug*en（3），*Graphiken beschreiben*（2）bis *Stellungnehmen zu etwa*s（3）. Dabei müssen die Testnehmer sich selbst in verschiedenen Rollen spielen. Dies setzt voraus, dass die Lerner schon vorher mit den möglichen Rollen in Deutschland und deren Ausgestaltung vertraut sind（Freundin, Frau > Freund um Rat fragen, mit Professor sprechen, Stellung nehmen），d.h. auch welche Möglichkeiten jeweils bestehen, z.B. in der Beziehungsstruktur. Im Modellset（Zimmermann 2013）finden sich zum Beispiel:

- Sprachregister（+ Elemente, die da möglich sind）;
- Freundin: Ratschlag für den Beruf
- Argumentieren
- Stadtvorstellung

Bei TD erfolgt die Bewertung nach Gesamteindruck, Umsetzung der Aufgabe und sprachlicher Realisierung（Zimmermann 2013b: 28）.

Bei dieser Form der Durchführung fehlt natürlich die Spontaneität im wichtigsten Teil, der Reaktion auf die Reaktion des Partners, eine Fertigkeit, die nur schwer zu üben ist und in westlichen Ländern unerlässliche Vorausetzung für das spontane, ja unvorbereitbare Sprechen einer Sprache ist. Dieser Teil ist beim RS wichtig.

RS testet direkt in authentischer Interaktion und dabei gibt es die Möglichkeit auf die Leistung des Testnehmers einzugehen, allerdings ist die Standardisierung entsprechend schwieriger. Dasselbe gilt für die Durchführung mit einer großen Zahl von Prüflingen. Dazu kommen noch Zeit und Kostenintensivität（Zimmermann 2013a: 6）.

Andererseits sind japanische Studenten monologisches Sprechen, z.B. Deklamieren, gewöhnt, haben aber besondere Probleme mit einem schnellen Ablauf beim Sprecherwechsel, besonders bei nicht-präfigurierter Rede.

Bei RS ist in der Prüfungssituation im ersten Semester ein zufallig zugeteilter Student aus derselben oder einer Parallel-Klasse und im zweiten Semester ein Muttersprachler oder gewohnheitsmäßiger Benutzer der Zielsprache als Testpartner in Person oder über Skype anwesend. Proband und Testpartner kennen sich noch nicht und es gibt für den Muttersprachler eigentlich auch keine Vorgaben außer der Bitte um etwas Geduld. Die Studenten haben bei der Gesprächsdurchführung auch darauf zu achten, dass Inhalte aus dem Sommer- bzw. Wintersemester in hinreichender Menge vorkommen（haben, es gibt, Perfekt, Kleidung, deutsche Stadt, usw. s. Reinelt 2012a）. Für beide ist dies in jedem Fall ein Anfangsgespräch. Dadurch ist die Anzahl und Auswahl an möglichen Themen am Anfang（Vorstellung, einige Daten, Vorlieben, einige Aktivitäten...）begrenzt, kann aber fast unbegrenzt erweitert werden. Dadurch ergibt sich eine Sprechzeit von mindestens 2 bis 3 Minuten bis zu 4 bis 5 Minuten, je nach zur Verfügung stehender Zeit. Die Initiative ergreifen die Probanden, die dies als Teil des Unterrichts vorher gelernt haben. Beispiele für solche Dyaden finden sich in vielen Arbeiten des Autors.

Um die Gespräche in Gang zu halten, haben die Probanden einige wichtige Gesprächs-

regeln schon im Unterricht vorher gelernt, und deren Befolgung bzw. Überwindung geübt. Dies gilt z.B. für die Regel, dass 10 Sekunden Schweigen zum Gesprächsabbruch führen. Die Gleichheit der Tests wird also durch den einheitlichen Gesprächstyp Anfangsgespräch und die normalerweise beschränkten Möglichkeiten gewährleistet.

Allerdings müssen bei der Durchführung noch einige Punkte beachtet werden, deren Einhaltung aber nicht immer gewährleistet werden kann:

− Unterschiedliche Testversionen sind gleich schwer, z.B. wenn nicht viel Zeit zur Verfügung steht, und die Gespräche deshalb kürzer sein müssen.

− Die Testkonditionen sind nicht für alle Kandidaten gleich, wenn einige Studenten warten müssen, weil ein Test vorher länger dauert.

Dazu kommen natürlich noch unwägbare Besonderheiten der Testsituation, wie Wetter（Kälte）usw.

3. 2. 2. 3. Zusammenfassung zur Durchführung

Um eine für alle gleiche Durchführung zu garantieren und so gegen alle möglichen Einsprüche gewappnet zu sein, muss TD soweit wie möglich maschinell durchgeführt werden. Dazu werden alle Unterlagen von allen Teilnehmern gesammelt und zentral ausgewertet. Abgesehen von logistischen Problemen beim Versand und bei der Durchführung ist dies sicher die beste Vorgehensweise. Dabei entsteht eine Zeitverzögerung, die sich TD auch leisten kann, die bei RS allerdings nicht möglich ist. Dieses Vorgehen von TD erfordert allerdings einen hohen Personalaufwand im Empfangsland Deutschland.

Beim RS ist es genau umgekehrt: In ihm werden alle und nur die Teile angesprochen, die gerade nicht maschinell durchführbar sind. Aufgrund dieser Besonderheit hat RS als Sprechtest auch einen wichigen Stellenwert. Deshalb wird er auch durch Tests für die anderen Fertigkeiten ergänzt. So werden alle anderen Teile getrennt und möglichst maschinell bewertet（s. aber die Schwierigkeiten dazu bei Reinelt 2012b und Eckes 2004）. Die Notenabgabezeit für die Beurteiler beträgt immer nur ein paar Tage, was auch Auswirkungen auf die Beurteilung（s. u.）hat.

3．2．3．Vergleich

Die folgende Gegenüberstellung erwähnt noch einmal die wichtigsten Punkte bei der Durchführung. TD wird zentral durchgeführt, RS ist immer lokal an der Stelle, wo Prüfling und Testpartner（auch virtuell）zusammenkommen. TD folgt verbindlichen Vorgaben auch in der Vorbereitung der Gespräche, während bei RS nur eine Steuerung durch den Gesprächstyp erfolgt, aber Kommunikationsmanagement möglich（*noch einmal bitte*）ist. Bei TD sind Hören und Sprechen getrennt. Bei RS gibt es dafür keine Zeit, aber Sprechen beinhaltet, idealerweise, ja immer auch hören und verstehen, was der Partner sagt.

3．3．Beurteilung

3．3．1．Einleitung

Schließlich gibt es auch einige Gemeinsamkeiten und Unterschiede in der Weise, wie die Beurteilung erfolgt. Dabei ist einerseits die Ausgangslage schon etwas unterschiedlich: Während es bei beiden Tests um die Beurteilung von mündlichen Produktionen geht, haben diese Produktionen unterschiedlichen Charakter. Zudem stehen die Produkte nicht gleich lange für die Beurteilung zur Verfügung. Schließlich werden verschiedene Typen von Kriterien angelegt. Im folgenden gehe ich nur kurz auf die Punkte ein, die für die Verbesserung des RS sinnvoll sein könnten.

3．3．2．Mündliche Produkte beurteilen

Beim TD gibt es 10 Aufgaben, die alle ein bis zwei Minuten dauern, so dass die Beurteiler deutsche Produkte von etwa 15 bis 20 Minuten Gesamtlänge von jedem Prüfling als Aufnahme vorliegen haben und diese natürlich so oft wie gewünscht abspielen können. Bei RS gibt es mehrere Beurteiler, die mit verschiedenen Kriterientypen urteilen.

1） Zum einen beurteilen alle Gesprächspartner ihren Studenten unmittelbar nachdem jede einzelne Testdyade zu Ende ist. Der Gesprächspartner hat noch die Erfahrung des Gesprächs selbst präsent.

2） Alle anderen Gesprächspartner und der nicht als solcher fungierende Kursleiter beurteilen jedes Gespräch.

Da es in dieser Situation mehrere Beurteiler gibt, erscheint es sinnvoll, je einen der beiden wichtigsten Beurteilungstypen zu benutzen（Yamada 2012）:

2a） Zur Beurteilung nach Kriterien wurde eine Vielzahl von Kriterienlisten konsultiert und eine Liste von 7 und dann 5 Kriterien erstellt, die für Deutschlerner wichtig sein könnten hinsichtlich Varianz, Relevanz usw., aber zugeich auch von nicht-professionellen Beurteilern gehandhabt werden können. Dementsprechend beurteilen die Prüfer nach den Kriterien Aussprache, Wortschatz, Grammatik, Flüssigkeit und Dialogizität und jeweils nach den aus Deutschland bekannten Notenstufen 1（sehr gut） bis 5（mangelhaft）（6 wird normalerweise nicht gegeben）. Wir können hier nicht auf alle detailliert vergleichend eingehen, stellen aber unten ein Beispiel vor.

Im Gegensatz zu RS sieht TD nach, ob beim Wortschatz Fehler stören und das verwendete Vokabular situationsangemessen ist. Da RS es mit Anfängern zu tun hat, ist die Angemessenheit weniger ein Problem, allerdings versuchen einige Probanden mit einem Englisch oder Japanisch durchsetzten Wortschatz weiterzukommen, was dann auch unterbrochen werden kann. Dies ist ein wesentlicher Unterschied zu TD, das nachsieht, ob alle Punkte und somit die Sprechhandlung（en） erfüllt sind. Bei RS gibt es keine solche Grenze oder Mindestanforderung, ausser im Wintersemester den oben erwähnten Wintersemesterinhalt.

Wennn nötig, kann man die einzelnen Kriterien noch anders gewichten. Die Punkte werden automatisch auf die 100 Punkte Skala der Ehime Universität umgerechnet.

2b） Weil es allerdings sein kann, dass eine Prüfung, besonders bei Gesprächen, nach allen Einzelkriterien problematisch war, aber die Beteiligten eine gute Zeit hatten, sollte man, wenn eben möglich, auch eine generelle, holistische Beurteilung einbeziehen. Diese nimmt der Kursleiter vor, der zugleich das Kommen, Gehen und Warten, und den Zeitablauf der Prüfung organisiert. Da er auch die Prüfung sehen kann, aber keine Zeit fuer detaillierte Beurteilungen bleibt, beurteilt er holistisch auf der 100 Punkteskala der

Ehime Universität.

Alle Beurteiler geben ihre Punkte unmittelbar nach jedem einzelnen Gespräch – und vor Beginn der nächsten Dyade - um Vergessens- und Überlagerungseffekte zu vermeiden, d.h. RS akzeptiert nur *on-the-spot* Ergebnisse. Allerdings werden aus Beweisgründen für die Verwaltung alle Prüfungen aufgenommen, so dass im Notfall auch eine Videonachkontrolle möglich ist.

3. 3. 3. Weiteres

TD kann es sich leisten, viele qualitätssichernde Maßnahmen vorzunehmen, die bei RS gerade nicht möglich sind.

TD schult die Beurteiler und stellt ihnen ein Handbuch zur Beurteilung zur Verfügung（Zimmermann 2012: .3）und monitort die Beurteiler regelmäßig. Bei RS ist dies nicht nur aufgrund fehlender geldlicher und zeitlicher Resourcen nicht möglich, eine ausführliche Schulung würde auch gerade das wichtigste Element, die Spontaneität, beeinflussen, die Prüfung vorbelasten und u.U. zur Verwendung prüfungstypischer Äußerungen（bzw. Fragen: *Haben Sie ein Hobby?*）führen. Deshalb werden die Sprechpartner/ Beurteiler gebeten, so zu beurteilen, wie sie es von ihrer Schulzeit kennen und welche Noten sie da geben würden. In dieser Hinsicht sind alle Beurteiler ebenfalls Experten.

TD kann ganze statistische Analysen zur internen Qualitätssicherung benutzen und damit Konsistenz und Strenge/Milde berücksichtigen. Mit den beschränkten Mitteln, die RS zur Verfügung stehen, können wir nur einen Strenge /Milde-Faktor errechnen, wenn ein Beurteiler auffällig zu streng wird. Dies ist allerdings bei produktiven Fertigkeiten einer der wichtigsten Faktoren（Eckes 2004b）und wird bei TD als Strengekoeffizient bei jeder Note auch zur Einzelbewertung hinzugefügt （Zimmermann 2013b: 3）.

TD kann solche Vergleichsbeurteilungen in den regelmäßigen Auswertungs- und Beurteilungsprozess einbauen. Bei RS sind zwar normalerweise, d.h. ohne zwingende Gründe, Mittelwert und Zweit- bzw. Drittkorrektur nicht möglich. Wir haben aber Vergleiche mit Nachbeurteilunegen per Video gemacht und sie führten zu sehr ähnlichen

Ergebnissen wie die Spontanbeurteilungen, d.h. das Instrument war wohl doch nicht so schlecht und die Annahme, dass die Beurteiler schnell aus ihrer eigenen Erfahrung urteilen können, wohl richtig.

3．3．4．Einordnung und Verwendung der Beurteilung

TD testet das obere Ende des Referenzrahmens und die Angabe im Zeugnis weist die TD Prüfungsleistungen auf dem Zeugnis getrennt aus mit Kann-Beschreibungen zu allen Teilfertigkeiten je nach Anforderungsprofil.

RS testet dagegen das untere Ende des Referenzrahmens. Außerdem trägt im Universitätssystem die Beurteilung dieses Testteils nur einen kleinen Teil zur Scheinvergabe bei, da Sprechen eben nur eine von vier Fertigkeiten im Endtest ist. Der Kursleiter kann für seine Gesamtnote auch noch andere Kursleistungen miteinbeziehen, und jeder Kursleiter kann sowieso nur 50% der Notenpunkte zur Gesamtnote beitragen.

3．4．Beispiel Sprechen

Während TD seine Beispiele nur sehr spärlich herausgibt, kann man RS-Beispiele in vielen Publikationen des Autors finden. Ein auch nicht ganz vollständiges Beispiel für TD findet sich in Musterprüfung 5, allerdings nur mit einem von drei Stellungnahmeteilen.

Im folgenden geben wir ein transkribiertes Beispiel für RS. Alle Teile sind anonymisiert.

Tafel 1 : Beispiel RS WS 2010/11 S14 mit YG, Beurteiler RR und MK

	S14	RR		YG
		01 RR o.k. Start		
002	Guten Tag (0:14-8)		003	Guten Tag (0:18-3)
004	Eh, wie heißen Sie? (0:21-6)		005	Ich heiße Y G. (0:21-6) Wie heißen Sie?
006	Ah, ich heiße S14 (0:25-9) Saki		007	(kurze Pause) Ah, Hallo (0:30-6)
008	Ehm, wie/ woher kommen Sie? (0:35-6)		009	Ich komme aus China. (0:37-6)
010	Ich komme aus Mat/eh Matsuyam a (0:43-7)		011	Ah, Sie kommen aus Ja/aus Matsuyama (0:46-9)
012	(kurz) Ja! (0:48-4)		013	Wo wohnen Sie jetzt? (0:48-4)
014	Eh, in Kewa, Kewa ist Norden von Matsuyama (0:57-8)		015	Ah, schön. (01:00-2)
016	Haben Sie Familie? (01:06-3)		017	Ja, meine Eltern sind ja in Ja/er in China, ich bin in Deutschland. Haben Sie Familie/haben Sie Geschwister? (01:17-6)
018	Meine Vater heißt eh Kirao. Er ist Angestellter. Meine Mutter heißto Kumi. Sie isto Hausfrau. (01:32-8)		019 020	Haben Sie keine Geschwister? (01:36-8) Haben Sie Bruder.
020a	Nein.		021	Ich auch nicht (lacht leicht) (01:40-7)
022	Was haben Sie am Wochenende denn so alles gemacht? (01:48-7)		023	Ehm, am Wochenende habe ich zu Hause DVD geguckt. (01:51-7)
024	Oh, was? (01:56-0)		025	Was/

027	Madagaskar?(02:02-8)		026	Kennen Sie einen Film: Madagaskar(02:00-6)
029	Ooh!(02:04-8)		028	Hm, Anime, Anime
031	Von/eh/am Freitag abend, eh, habe ich gejobbt. Am Samstag habe ich gefahren das Kino (0:02:26-8)		030	Ja. Was haben Sie am Wochenende gemacht? (02:08-8)
033	Ah... (02:43-3		032	Ah, Sie sind ins Kino gegangen. Was haben Sie geguckt?(02:31-2)
035	(lacht kurz)		034	Zu schwer(02:43-3)
			036	is o.k.(02:45-8)
038	ja, schön:02:54-5)		037	Ja, aber war interessant? (02:49-3) War es gut ?:02:50-6)
			039	war es schön,ja(02:54-3) gut, super(02:55-8)
041	Ah. ich. koche. Schoko/ Schokolade(03:11-3)		040	Was machen Sie heute abend?(03:00-3)
043	(lacht)		042	Ah, Valentin, ah!(03:16-3)
045	Ja!(03:21-8)		044	Ah, ja ja ja!(03:18-5) Machen Sie/machen Sie selber Schokolade? (03:20-9)
046	(kichert weiter)(03:25-6)		047	Sie müssen ja eine Tafel Schokolade an Herrn Rudolf (03:28-5)
		048 Ja! (03:29-4) Sehr gut, das habe ich gerade gedacht! (03:31-6)		

049	(kichert weiter) (03:31-6)	050 RR Das habe ich gerade gedacht. :03:33-8) o.k.	051	Ja! (03:36-7)
053	Und Sie? (03:41-8)		052	Ja, ist schon sehr interessant.
055	Ja		054	Essen Sie Schokolade gern?
057	(03:52-7 Und Sie, was machen Sie heute abend? (03:54-5)		056	Auch, o.k. gut. (03:45-0) (03:52-7)
			058	Heute abend gehe ich mit Freunden essen (03:58-2)
060	Ich (liebe gern) Milch gern (04:18-7)		059	Ein Freund, hm, in Freiburg kommt mich besuchen (04:05-1) Ja, eh, was essen Sie gern und was trinken Sie gern? (04:09-2)
062	Aach. Und Sie? (04:21-6)		061	Ah, Milch (04:18-7)
		064 RR Machen wir so weit. (04:27-4)	063	Ich trinke gern Tee. 3-6) gern. grünen Tee. (04:27-4)
066	Danke schön (0:04:32-3)		065	Danke schön (0:04:32-3)

Tafel 2: Beurteilungen fuer S14: Scores:YG, KM, RR

	Aussprache	Korrektheit	Wortschatz	Flüssigkeit	Dialogizität		MK	RR 口頭
	10%	15%	25%	35%	15%	100%		
S14	1	1	1	1	1	1	100	100
	Aussprache	Korrektheit	Wortschatz	Flüssigkeit	Dialogizität		YG	
	10%	15%	25%	35%	15%	100%		
S14	2	2	1	1	1	1.25	97	

Die Studentin in diesem Beispiel erhält unmittelbar nach ihrer Prüfung eine sehr gute Bewertung mit der Aussprachebeurteilung von YG als einzigem Ausreißer. Allerdings kann man bei der Kontrolle durch das Transkript feststellen, dass doch nicht nicht alles gut läuft in diesem Beispiel mit einer Studentin im WS 2010/2011 und YG, einer Chinesin, die GER C1 bestanden und in Deutschland ihr Studium abgeschlossen hat, und jetzt dort arbeitet. So initiert S14 viel, fragt nach Vergangenem und kann dies auch im Perfekt beantworten, und hält außerdem eine ausreichende Dialogizität aufrecht, so dass die beiden Male mit über 10 Sekunden Wartezeit nicht auffallen.

4. Ergebnisse

Die Ergebnisse dieses Vergleichs beziehen sich auf zwei Bereiche:

1. Unterschiede bzw. Gegensätze, wo diese gerade nötig sind.
2. Ähnlichkeiten, wo wenigstens der Versuch der Ähnlichkeit unternommen wurde, auch wenn natürlich keine Vergleichbarkeit erreicht werden kann.

4.1. Unterschiede und Gegensätze

Während sich Unterschiede auf allen Ebenen und in allen Bereichen leicht finden lassen, sind doch einige kennzeichnende Gegensätze feststellbar, die durch ihre Kategoriengleichheit Hinweise geben können. Dazu gehört die Stellung innerhalb der Fremdsprachenlerngeschichte der Lerner am Ende der Anfangsstufe（RS）vs. am Anfang der Endstufe（TD）. Ebenso gehört in diese Kategorie die Direktheit des schnellen Äußerungsaustausches mit dem Partner, bei der RS uneinholbar punktet, während TD durch die Anforderung der Verfügbarkeit für viele auf technische Maßnahmen angewiesen ist und damit den Verlust der Partnerpersönlichkeit in Kauf nehmen muss.

4．2．Ähnlichkeiten für einen Vergleich

Ähnlichkeiten zwischen RS und TD finden sich in einigen Bereichen der Beurteilung（Einberechnung von Strenge）, der Stelle am Anfang des Universitätsbereichs（allerdings an anderen Polen）, in der Erfassung, d.h. beide erlauben als eine Art Sprachstandstest in ihrer Gesamtheit eine einheitliche Beurteilung der sprachlichen Leistungen und Leistungsfähigkeiten der Probanden: TD ist darauf ausgelegt, von vornherein alle Fertigkeiten zu testen. Auch RS hat für die anderen Fertigkeiten entsprechende, hier nicht vorgestellte, Tests, was in Japan eher selten ist.

Beide Tests benutzen offene Fragesysteme im Bereich Sprechen, denn nur diese sind hier sinnvoll.

Als ein Ergebnis dieser Arbeit kann man wohl feststellen, dass trotz aller Unterschiede es eben doch auch in Japan sinnvoll ist, zum einen sich - fast perfekte - Tests wie TD zum Vorbild zu machen, und zum anderen auch Universitätsstudenten zuzutrauen ist, sich mit Zielsprachenmuttersprachlern zu unterhalten, und dies bestätigt werden kann. Dies ist immerhin eine Leistung, die vor weniger als einem Jahrzehnt noch fast unvorstellbar erschien. RS startete mit den zwei übergeordneten Zielen: Fairness UND Leistung. ,

Allerdings ist auch auf der beschriebenen Stufe noch nicht alles ideal und der abschließende Teil geht noch darauf ein. Dabei waren für alle Prufungsteilnehmenden gleiche Bedingungen zu schaffen hinsichtlich

- Testformat und Schwierigkeit der Aufgaben
- Konditionen der Testdurchführung
- Beurteilungsmaßstäbe

und vieler anderer Faktoren, sowie aber auch die Festlegung und Forderung einer Leistung in einer Prüfung einer Fremdsprache auf einer unteren Ebene im allgemeinbildenden Bereich des Studiums an einer japanischen Universität. Vielleicht glückt nach vielen Revisionen die Erfüllung der Testgütekriterienkombination, die sich aus der folgenden Formel ergibt: Nützlichkeit = Reliabilität +（Konstrukt）validität + Authentizität + Interaktivität + Rückwirkungseffekt + Praktikabilität（Bachmannn & Palmer 1996）.

5. Verbesserungen, Gefahren und zukünftige Entwicklungen

Verbesserungen sind an allen Ecken und Enden nötig, andererseits gibt es sicher auch Teile. die beibehalten werden sollten, nicht zuletzt die Spontaneität, die ja gerade diesen Test von anderen bisherigen abhebt. Gleiches gilt für die Anwesenheit der Gesprächspartner, in Person oder über Skype. Ob diese durch fortgeschrittene Studenten teilweise ergänzt oder gar ersetzt werden können, muss noch untersucht werden（Reinelt 2013e）.
Eine（positive!）Gefahr für den Test entsteht, wenn Studenten im 2. Studienjahr keine weitere, verbesserte Prüfung vorgelegt（abverlangt）werden kann. Diese sollte aber weiterreichend sein und wohl auch Gespräche über das Fach der Studenten beinhalten, ist aber im zweiten Jahr noch nicht erstellt, und die Studenten integrieren Gesprächsteile über ihr Fachstudium nur ganz wenig in die Dyade!
Im Bereich der Beurteilung kann der Ausgleich der ermittelten Strenge bzw. Milde（Zimmermann 2013b）auch bei RS notwendig werden.
In den Bereichen Wortschatz und Grammatik muss noch genauer eruiert werden, welche Teile aufgenommen werden sollen und welche übergangen werden können. Schließlich werden in diesen Bereichen auch Fehler relevant, z.B. inwieweit sie stören, oder aber übergangen werden können.
Auf Seiten der Partner können Ungleichheit und Unwohlsein durch harte Dialektsprecher entstehen. Ebenso in den Bereich der Zukunft gehört die Überlegung, Studenten, die dann S2 zweimal gemacht haben, als Tester（Beurteiler und Partner）einzusetzen. Schließlich wäre es eine Idee, den Test mit offiziellen, international anerkannten Tests zu arrangieren.

6. Literatur

Bachmannn, L. & Palmer, A.（1996）*Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: OUP.

DAAD（2004） *Deutsch als Fremdsprache – Sprachprüfungen für den Hochschulzugang.* Bonn 2004.

Eckes, T.（2004a）. Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In A. Wolff, T. Ostermann & C. Chlosta（Hrsg.）, *Integration durch Sprache （Materialien Deutsch als Fremdsprache*, Bd. 73, S. 485–518）. Regensburg: Fachverband Deutsch als Fremdsprache.

Eckes, T.（2004b）Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache“（TestDaF）. *Diagnostica*, *50,* 65–77.

g.a.s.t.（2012） *TestAS 2012*（Gesellschaft für Akademische Studienvorbereitung und Testentwicklung）（g.a.s.t.）: Bochum.

HST, *High-stakes testing*（2013） http://en.wikipedia.org/wiki/High-stakes_testing

GER（2013） *Gemeinsamer Europäischer Referenzrahmen* https://de.wikipedia.org/wiki/Gemeinsamer_Europ%C3%A4ischer_Referenzrahmen

Reinelt, R.（2005） *Einleitung: Herausforderung und Chance - Krisenbewältigung im Fach Deutsch als Fremdsprache in Japan* Daf-Westjapan, Ryukyu Universität, Okinawa, Japan 12–14. Dezember 2003 München: Iudicium, 2005, p 13–22（herausgegeben mit Petra Balmus und Guido Oebel）.

Reinelt, R.（2012a） Azoc: *How German made a difference*. In: Reinelt, R.（ed.）（2012） The OLE at JALT 2011. Compendium compiled for OLE by Rudolf Reinelt Research Laboratory, Ehime University, Matsuyama, Japan, p. 144–153. http://web.iess.ehime-u.ac.jp/JALT2012Compendium/26%20Rudolf%20Reinelt.pdf

Reinelt, R.（2012b） Integrating speaking, writing and grammar in 2FL German testing. In: Reinelt, R.（ed.）（2012） *Other Language Educators 2012 at JaltCALL and PanSIG* compiled for OLE by Rudolf Reinelt Research Laboratory, Ehime University, Matsuyama, Japan, p. 17–27. http://web.iess.ehime-u.ac.jp/JALT2012/2%20Rudolf%20Reinelt.pdf

Reinelt, R.（2013a） *Über Englisch hinaus mit Deutsch - im ersten Jahr im allgemeinbildenden Bereich* - DAAD Lektorenfachseminar 2013, Kyoto, February 9.

Reinelt, R.（2013b） *Students as raters（SR）in 2FL German oral exams* 4th Annual Shikoku JALT Conference, Kagawa University, Takamatsu, May 11.

Reinelt, R.（2013c） *Employing student learners as speaking partners in the 2FL German oral exam*. JALTCALL 2013, Shinshu University, Matsumoto, June 2.

Reinelt, R. 2013d） ***「グローバル化時代における教養教育としての初修ドイツ語学習ー授業からの例を用いてー」* 第61回中国・四国地区大学教育研究会**, Ehime University, Matsuyama, June 9.

Reinelt, R.（2013e） *Second foreign language instruction as communication education* **日本コミュニケーション学会第43回年次大会, 立教大学**, Tokyo, June 23.

TestDaF（o.J.）*Musterprüfung 5*. München: Hueber.
Yamada, T. 2012 学生の能力を高める評価法のコツ。松山：愛媛大学.
Zimmermann, S.（2013a）*Grundlagen des Testens und Bewertens – Das Beispiel TestDaF*. DAAD Lektorenseminar, Kyoto, February 10..
Zimmermann, S.（2013b）*Sprachkompetenz testen – Das Beispiel Test Deutsch als Fremdsprache（TestDaF）*. DAAD: Tokyo.

この論文では、当教室で開発した口答試験（RS）とドイツの大学入学条件に当たるTestDaF（TD）との比較の可能性を論じる。最近の大学の共通教育における第二外国語教育では、４技能を育てる必要がある。筆者は、練習・実践が教室でしか行えない口頭技能のために、期末口頭試験（RS）を開発した。そこで受講生は前学期末に同受講生と資料を見ずに２－３分程度話し、後学期末に初対面のドイツ人と前準備なしで２－４分会話を交わすという口頭試験を実施している。日本において、このようなレベルの試験としては比較できる対象が少ないため、ドイツ大学入学試験であるTestDaFとの比較で重要な改善点などを得られることになる。

内容として、前書きでこの論文の目的と構造を説明し、第一章ではRSの背景（第一部）、RSの開発とその後の展開（第二部）、現状（第三部）について論じる。第二章ではTestDaFを紹介し、第三章で比較を行う。第三章の第一部では理論的な基準を取り扱い、第二部では実践を考え、第三部では評価方法、第四部では具体的な例を紹介する。第四章では比較の結果と問題点を論じ、第五章ではRSとの改善点とこれからの課題に言及する。