Inter-rater correlation in native speaker German beginners course oral examinations

Rudolf Reinelt
Ehime University

Reference data:
Reinelt, R. (2008). Inter-rater correlation in native speaker German beginners course oral examinations.
In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

The aim of this study is to demonstrate that in the difficult situation of second foreign language teaching in Japan, taking oral examinations for a beginner's course in German as an example, native speaker exchange students can be employed as raters with at least sufficient inter-rater correlation. After an introduction to the author's speaking-focused course, the subjects, the raters and the procedures used in the oral examination are introduced. For the latter, a list of criteria and a scorecard for the exchange student raters was developed, and the scoring methodology is demonstrated. The inter-rater correlation and other results across all the author's courses are presented and briefly discussed. Views to future improvements conclude the paper.

この論文の目的は、第二外国語としてのドイツ語の初級クラスにおける口頭試験において、ドイツ語母語話者である留学生を採点者として雇い、十分なinter-rater　correlation評価の一致性を得ることが出来ることの可能性を提示することである。まず、筆者が行っている会話中心の授業の概要を示し、学習者、評価者、口頭試験の特徴について記述する。さらにその試験の評価の基準、および筆者によって考案された採点者のためのスコアーカード、そして実際のスコアーの例を提示する。その後、採点者間の評価の信頼性について考察を加え、筆者の担当するすべての授業における調査結果を報告し、その要点を手短に議論するとともに将来に向けての展望を記して本論の結論とする。

In a questionnaire administered in the very first class of the entire author's German as 2FL courses in the 2007 summer term 2007, the following numbers of responses are shown in table 1:

Table 1. Course introductory student questionnaire

設問と答え	Questions and answers	No of answers
この授業で習いたいものは何ですか	Item 1: What is it that you would like to learn in this course?	Total 196
ドイツ語会話全般 (日常会話)使えるドイツ語	All of German "conversation" (Everyday "conversation") Useful German	110
会話全体	Total mentioning of "conversation" (kaiwa)	137
これを達成したら満足します	Item 2 I would be satisfied if I reached ….	Total no. of answers 182
会話全体:	Total mentioning of "conversation"	102

Obviously, a considerable percentage of the answers contained "conversation", the ability to be taught only in a course focused on speaking.

In order to be fair, courses attempting at satisfying such wishes have to evaluate their students in an oral examination where, ideally, native speakers of the target language rate the learners' progress. In order to demonstrate how this can be done even in introductory second foreign language (2FL) courses in Japan, the author, using his German beginners course as an example, presents to the reader how the course's oral examination was prepared so that two cooperating native German speaking exchange students could function as raters and how at least sufficient inter-rater correlation was established.

Also, for the oral examination to be meaningful at all, a certain advanced level of speaking (considerably beyond *Guten Tag*) had to be reached carefully. In order to facilitate an understanding of this precondition of the examination and any ratable results, this paper mentions details and relevant parts of how the course was conducted in some length.

For a number of years the author in his courses has tried to satisfy the above-mentioned requests. In his one-term course, the students were paired with varying partners in dialogic activities and practiced German dialogues followed by activities freely expanding on these to cover the main conversational structures such as greetings, supplying information about oneself and others in various ways using the most basic grammatical structures, and the most important serial words such as numbers and days of the weeks, etc.. Course objectives included techniques for maintaining conversation in German and information gathering about Germany (extensive information in Reinelt, 2007b). Other skills were treated by a Japanese partner teacher and, after a brief introduction, outsourced to the learning management system Blackboard (Reinelt, 2008a, in press), and the mail server Active mail (Reinelt, 2008b, forthcoming), both available at Ehime University.

With the course focusing on speaking, the need for an oral examination arose. In 2000, a holistic test was designed (Reinelt 2000) where the teacher checked whether all previously studied elements of a 5-part algorithm had been learned. This was done by assessing the performance of randomly paired students in a two-minute ad-hoc dialogue. While this approach ranks high on feasibility since it allows for testing and assessing a large number of students within

the 90 minutes available, its disadvantages in terms of objectivity and other test criteria for validity are obvious and it was felt that the method for assessment should be improved.

As professional tests (Sprachnachweis .n. d.) are too difficult for many learners of English in Japan (Smith & Nederend, 1998), as are most of the professional German speaking tests available (Sprachnachweise, n.d.). Furthermore, paying for them is beyond the means of (teachers at) former national universities in Japan, as is lengthy tester training. Japanese colleagues were not asked to participate in this first try, because they might have thought such a request intrusive, or were instructing their own classes and were not available.

Under such circumstances we have to make do with what is feasible, and try to develop a new method under these limited conditions. As Jeffrey (n.d.) and Smith & Nederend (1989) describe developing valid English oral examinations for language courses in Japan at length, we only have to discuss the specific differences to their approaches: First vs. second foreign language; longer vs. shorter learning time, unit requirements, etc..

Although the teacher scoring of oral examination, as in Reinelt (2000), is minimally acceptable according to Grotjahn (2006), objectivity can be improved if there are more raters. Such outside raters enable an assessment based on the "learners' performance in the test itself, and not on how they might be expected to perform based on performance in the classroom" (Jeffrey, n. d., p.14), if based on only one teacher, and other teacher biases as mentioned by Smith & Nederend (1989). Also, if there are two or more raters, the "inter-rater reliability (is) used as a measure of the consistency between the examiners while applying the test criteria" (Jeffrey, n. d., p. 2). More on details of inter-rater correlation can be found in Uebersax (2006) and Uebersax (2007), but the technical details are beyond the scope of this paper. Since not exactly the same measure was used, we here rather use the term inter-rater correlation.

The study

This part first introduces all involved parties and what they had to do, before coming to scoring and inter-rater correlation.

The learners

The learners in this study were all first year students at Ehime University with various majors. They took this 2FL (=Second Foreign Language) beginner's course in German as partial fulfillment of their general studies requirements. Previous FL learning experiences were limited to 6 years of English and some rudimentary Ancient Chinese.

The raters

The evaluators in the term-final oral examination (Reinelt 2007a) were the author, who had taught the courses, and, in turns, two exchange students (25) and (23) from Freiburg university, the sister city of Matsuyama, where Ehime University is located. They were majoring in computer information science and psychology, respectively. Neither was professionally engaged in language teaching. Both

had had 6 years of English instruction at the "gymnasium", the combined JHS and HS in Germany, and used this in their everyday life in Japan. Knowledge of other foreign languages such as French was limited, as was their Japanese. Due to limitations on time (the exchange students had to attend their own courses) and money (the author paid the students after the scoring out of his own pocket), no norming was possible, and their experience of long time gymnasium FL instruction in Germany was deemed "norming" enough. Both would then rely on their own ample (school) testee experiences in estimating the behavior of the testees in this study according to their own understanding of the criteria given below.

The oral examinations

The basic criterion

The basic criterion was defined, in accordance with Jeffrey (n. d., p.4) as: "will a native speaker of German, who is sincerely open to communicating with Japanese, be able to understand what the learner is trying to say, even though he or she is mostly unaccustomed with Japanese mannerisms and speech patterns?" As the learners in this study had only had a 15 week course in contrast to Jeffrey (n.d.)'s 6 to 8 years, the test goal was simply to be as practical as possible and to make optimal use of resources at the same time.

Criteria and scorecard

In order to provide the exchange students with a manageable list of criteria for evaluation and after sifting through a considerable number of speaking test descriptions, we ended up with criteria very similar to those mentioned in Jeffrey (n. d.). No outside or new criteria such as social relationships etc. were controlled; only those abilities practiced in class could be tested due to the limited content learnable in one term of 15 contact hours. Criteria should not be too detailed because simultaneously observing them makes scoring more difficult and cruder. Note also that the range of each of the criteria was limited by the little content the student had learned. We arrived at the following five criteria:

- Pronunciation (*Aussprache*),
- Correctness/ grammaticality (*Korrektheit, Grammatikalität*),
- Vocabulary (*Wortschatz*)
- Fluency (*Flüssigkeit*)
- Dialogicity (*Dialogizität*)

For the non-linguistic exchange students the criteria were formulated in German so that they sounded familiar to their own (school) FL learning experiences in Germany in simple, everyday descriptions as questions (as in Jeffrey, n.d., p. 2), such as in Fig. 1:

Gegenseitigkeit (Dialogizität): Spricht nur eine Person (für die positiv, für den Partner negativ?). Sprechen die beiden einander an mit Fragen? Stellt immer nur eine(r) die Fragen? Verweigert eine(r) das Gespräch > Abbruch nach 10 Sekunden Nichtssagen (Time Out, den Lernern vorher bekannt). Sehen die Partner in verschiedene Richtungen oder nach unten? (Dialogicity: Is only one person speaking (advantageous for that person, disadvantageous for the other person)? Is only one posing the questions? Does one testee refuse to talk, leading to a time out after 10 sec. silence (time out introduced beforehand). Do the partners look at each other or in different directions or down?)

Figure1. A sample criterion description: dialogicity

Note that the criteria cannot easily be kept completely discreet and separate from each other, an issue beyond the scope of this study.

The need for familiarity also determined the 6 point range scale usually used in Germany, even in official government institutions such as Stiftung Warentest (n.d.).

Table 2. The 1 to 6- evaluation band for every criterion

1 very good
2 good
3 satisfactory
4 barely satisfactory
5 deficient
6 not sufficient at all
5 and 6 mean “not acceptable” and were collapsed to 5 in this study

Since in the oral examination students were tested in pairs, a scorecard for scoring two students simultaneously was developed. With the bandwidth for one student extending to the left and for the other to the right of the criterion given in the middle column, the actual scorecard is shown in figure two:

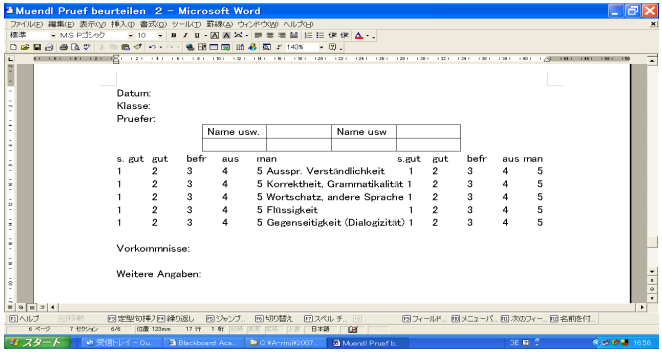

Figure 2. Left-right scorecard for two students in 5 criteria and 5 grades

Learner preparations

During instruction in the term in class, the upcoming criteria in the test were mentioned when the dialogues were practiced, but no extra time was reserved for any special training (Smith & Nederend, 1998). Important hints were presented in the Blackboard file, made visible to the students about three weeks before the test itself. For example, the following regulations were mentioned:

- There will be random partners
- 10 seconds of silence lead to time out
- Helping each other only in German
- Facing each other, etc.

In the lesson immediately preceding the oral examination, all speaking activities introduced so far were reviewed as one long series of communicative activities. Students would then practice these with a randomly attributed partner without any other material, in one long talk usually lasting about 7 minutes.

The oral examination itself

There were no specified content requirements in this test. Due to the limited amount coverable in the short term, the students just had to make use of what they had learnt.

In contrast to Jeffrey (n. d.) and Smith & Nederend (1998), the tests had no particular required parts (Smith & Nederend, 1998) (functions, role play, visual stimuli, prepared play), since there were only about four minutes available and it was up to the students what they did in this time, as long as the rules in the preceding paragraph, announced before in Blackboard, were followed (German only, no pauses longer than 10 seconds, no strange answers. Sudden topic changes had to be allowed due to the limited number of content items). The raters did not intrude (as in Smith and Nederend, 1998).

Students knew that they would be tested in pairs, but of course, not who the partner would be. Smith & Nederend (1998) mention saving time and a reduced anxiety over interacting with the teacher as advantages of using pairs. Naturalness, however, was not easy to keep in the test situation (see however Reinelt, 2007b). In this study, the test was more important than a good conversation environment, which may have produced different results again. For educational purposes, all pairs were videotaped.

While the usual length in standardized tests is 10-to 30 minutes, this was both too long and logistically impossible in this study. The test had to be performed, scored and rated within one and the same class, i.e. a course time of 90 minutes. This meant scoring, on the scorecard as well as the author's holistic scoring, had to take place immediately after each pair was finished.

On the test day, the course gathered in the classroom. As the written examination had to take place simultaneously, students were then given the topic "Mein Deutsch" (my German - what I have learned in this course -) to write about without any material at hand. Three students were then called to a separate room, where the exchange student and the author waited. The students were again randomly paired with classmates already waiting and asked to sit facing each other halfway across a table corner, so that they would face each other, but also the examiners and the camera could see them. They were then given the start sign and a stopwatch was started in order to keep the limit of two to four minutes and to control for silent periods. Students started immediately, usually with personal introductions, and continued for two to four minutes, sometimes even longer, especially if there were problems. Then they were given the stop sign. They thanked each other and left. The author then randomly paired two of the waiting students and called them in.

Scoring

Two scores

Every student was given two scores, a holistic one by the author and one on the scorecard by the exchange student. In this part, we take A3 in table 3 below as an example.

- The author gave one holistic score of his impression of each student's productions on the 1-100 scale used in Japan. Note that at Ehime University, as at many other universities, the passing bracket starts from 60, giving a barely passing until 65, a good passing between 65 and 79, a very good between 80 and 89, and excellent between 90 and 100 points reached.
- The exchange students used the scorecard and evaluated each student according to the criteria as described above. These scoring results were then weighed and converted to the Japanese 1-100 scale.

As two types of scores were employed, the criterion referenced scorecard and a holistic score by the author, this procedure followed that of Smith and Nederend's (1998) dual method of scoring.

Weighing of criteria and conversion into the Japanese scale

Weighing of the criteria was possible with the criterion-referenced part (Jeffrey, n. d.) and introduced in the conversion of the data, not at the point of scoring. Due to the course's focus on speaking and conversation, the following ratios (A3) were used:

For the example in A3 this lead to:

- the sum of C2 to G2 as SUM in H2 (=100)
- and at H3 etc. for each student the SUMPRODUCT of A3 to G3 and the weighs divided by H2 (H3=2.325).
- at I3, the weighed rates H3 etc. were converted into the Japanese scale of 100 with a range of 40 (from 60) and allowing for three levels (good, normal, weak) within each grade (and rounded where necessary): 100-((H3-1)/3)*40 (I3 =82).

With the exchange student's points in I3 and the author's holistic scoring in J3, we now have two sets of data for each student. We are fully aware of the differences of how these numbers were arrived at, but they are relevant data in the face of the 100 point grading system, and we only have to make sure that the two grades do not differ too widely.

Table 3. Scores for two students

A1 Stud. nr.	B Name (anonym)	C Pronunciation	D Correctness	E Vocabulary	F Fluency	G Dialogicity	H	I Hen	J RR Oral Test score
A2		10%	15%	25%	35%	15%	100%		
A3	KeNi	2	2.5	3	2	2	2.325	82	84
A4	SaMo	2	2	2	2	1	1.85	89	88

Inter-rater correlation

Inter-rater reliability, the "degree of agreement among raters" (Inter-rater reliability, n. d.) "gives a score of how much homogeneity, or consensus, there is in the rating given by judges" (Inter-rater reliability, n. d.). For this correlation testing, the intra-class correlation coefficient (ICC), an "improvement over Pearson's r and Spearman's rho" (Inter-rater reliability, n.d.) is used. As every case is rated by all raters, and as the raters are not accidentally chosen, and as we want to know the average measure, ICC(3,k) applies (Intra-Klassen-Korrelation). For this, university of Ulm offers a convenient calculator (Intraclass Correlation, n.d.) at <http://sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc.html>..

With the data for Mi6 pairs entered and the first 10 pairs shown in fig 3 below, the calculator returns 0.7743665 as result for ICC3k for the whole course.

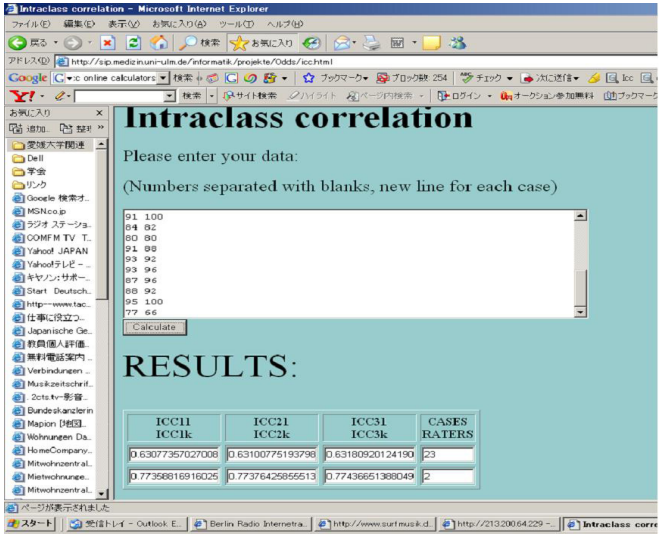

Figure 3. The Ulm university ICC calculator

Results and discussion

The following table gives the complete figures for one class.

Table 4. Complete score table for one class

Stud. nr.	Name (anonym)	Pronunciation	Correctness	Vocabulary	Fluency	Dialogicity		Exch. Stud	OralTest RR
		10%	15%	25%	35%	15%	100%		
94	KeNi	2	2.5	3	2	2	2.325	82	84
77	SaMo	2	2	2	2	1	1.85	89	88
52	YuSa	1	1	1	1	1	1	100	100
50	KaNa	1.5	1.5	1	1	1	1.125	98	96
53	AkSo	2.5	2	2	2.5	1	2.075	86	82
90	ShUe	2	2	2	2	1	1.85	89	92
42	KaKo	2.5	3	2.5	2.5	1	2.35	82	96
50	TaJo	1.5	1.5	1	1	1	1.125	98	100
16	HiMa	2	1.5	1.5	2	2	1.8	89	80
93	MiHy	2	1.5	2	2	2	1.925	88	88
13	KaJo	1.5	2	2	2	2	1.95	87	82
	YuYa						0	113	
71	KiKa	2	2	2	1	1	1.5	93	96
64	HiMo	1.5	1	1.5	1	1	2.525	80	96
44	ChiNa	1.5	1	2	2	1	1.65	91	100
62	KeSa	2	2	2	2.5	2	2.175	84	82
56	MaFu	2	2.5	3	2.5	2	2.5	80	80
04	MaMa	1.5	1.5	1.5	2	1.5	1.675	91	88
84	MiMi	1.5	2	1.5	1.5	1	1.5	93	92
07	MiFu	1.5	2	1.5	1.5	1	1.5	93	96
74	TaNa	2	2.5	2	2	1.5	2	87	96
01	SaUma	1.5	2	2	2	1.5	1.875	88	92
73	SuHo	1	1.5	1.5	1.5	1	1.375	95	100
18	YuYa	2	2.5	3	3	2.5	2.75	77	66

ICC3k: 0,77

Table 5: All inter-rater correlation results

The ICC3k data between raters in all classes were as follows:

Weekday/ lesson	class maj	ExSt	Auth	Nr-of-pairs	ICC3k
Mo4	Science	MirSe	RR	23 Pairs	0.48
Di 3	LawLitPaed	HenSki	RR	39 Pairs	0.58
Mi 4	Science	HenSki	RR	32 Pairs	0.83
Mi 6	Ev Gen Polit	HenSki	RR	23 Pairs	0.77
Fr 3	LawLitPaed	HenSki	RR	17 Pairs	0.90

Except for one class with under 0.50 degree of agreement, all other cases showed either moderate 0.5<-0.75 or substantial (0.75<-0.90) agreement between the scorers. Usually a rate in the high eighties is considered wishful, and with the amateurs, a somewhat lower rate in the high seventies can still be accepted. With the striking exception of Mo4 where almost total disagreement can be stated, and Di3, the very first scoring, the results are encouraging at least, although not outright overwhelming.

Certainly these were not surprisingly high rates, but they do still prove that at least for an initial increase in objectivity, this approach could be practicable, given all the other problems that usually intervene any objectification of such tests in the context considered in this paper. In the end, it would mean that the use of native speakers with such a criteria list can lead to at least minimally more objective results than if only the teachers score.

There are of course too many issues requiring improvement to mention here.

The native speaker exchange students also were surprised by how much the students produced, and how natural this looked, but they also stated the students' unadjusted, haphazard follow-up of topics/ questions, where there would have been a chance to put some individuality to the conversation event. This test would thus minimally fulfill Smith & Nederend's (1998) aim of such a test; i.e. to show and profile language functions.

Outlook for the future

This course made students speak in a second FL for a considerable time during practice, preparation and the test itself. However, only very few opinions were entered by the students on the course evaluation sheet. This lets any negative aspects stand out even stronger: The course did not result in a raise in the class satisfaction score. In the present university situation, with as many or more vacancies than students, this may, however, be a more important point to consider than objectivity or refining a test.

Other problematic issues should also be mentioned:

- limited natural(istic)ness.
- co-operation by other teachers;
- if no norming is possible, more raters would perhaps be better, although this would have to be tested;
- so far only learner pairs were tested, but the aim of any language learning is to be able to speak with speakers of the target tongue who usually do not speak the language of the first speaker. It would thus be better to test every student individually. But this may not

even be necessary (LeBlanc, 1997), especially when feasibility is an issue.

Despite these remaining issues, a few positive results can be taken as a starting point for further development and research:

- An extension of the test should be possible for a second term, e.g. by talking about topics after prompts. This is presently being tried in the winter term (German city, their partner, etc.; topics dealt with in class). This would probably facilitate scoring (Smith & Nederend, 1998). An example for this is Silva (2007).
- Forward effects (Smith & Nederend, 1998) are difficult to measure because 2FL courses usually last only one term, and many students actually could not take the second part in the winter term without sacrificing one unit point. If the fact that 40 students still did come can be considered a forward effect, this was indeed impressive.
- Even longer long-term effects can hardly be measured, but students who go to the target country, have an initial advantage, as do those taking German courses again later by being able to easily dig up their speaking knowledge.

We hope this paper, taking German as an example, has shown that even in the difficult situation in Japan it is possible to conduct oral examinations of 2FL courses, for example by making use of the human resources in the area, and that this can at least be worth a try, and if lucky, lead to coherent results. Improved repetitions of such examinations in even other languages may not only lead to increasing student satisfaction by showing them that they can indeed learn to speak a foreign language. The increased objectivity may also be used to convince the administration that foreign language courses indeed lead to accountable results. This may even lead to a more stable position for 2FL courses.

However, how the presented examination can be improved remains a promising task for the future.

Rudolf Reinelt has been teaching German on all levels of acquisition at Ehime University in Matsuyama since 1981. Recently, he made Blackboard, the digital LMS, and Active Mail in combination usable for second foreign language teaching. He can be reached at <reinelt@iec.ehime-u.ac.jp>.

References

Grotjahn, R. (2006). Sprachprüfungen als Instrument der Qualitätssicherung im Hochschulbetrieb (Language tests as quality assessment in higher education). *Dai 11kai kyojuhou seminar (11th didactic seminar) Kansai Gakuin university*, March 6-10.

Inter-rater reliability. (n.d.). [Online] Available: <en.wikipedia.org/wiki/Inter-rater reliability>

Intraclass Correlation. (n.d.) [Online] Available: <sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc.html>

Intra-Klassen-Korrelation. (n.d.). [Online] Available: <de.wikipedia.org/wiki/Intra-Klassen-Korrelation>

Jeffrey, D. (n.d.). *The Challenges of Creating a Valid and Reliable Speaking Test as Part of a Communicative English Program*. [Online] Available: <www.nuis.ac.jp/~hadley/publication/jeffrey/jeffrey-speakingtest.htm>

LeBlanc, L. B. (1997). Testing French Teacher Certification Candidates for Speaking Ability; An Alternative to the OPI. *The French review* Vol. 70, No.3. p. 383-394.

Reinelt, R. (2000). Mündliche Prüfungen im Unterricht DaF in Japan. *Der Deutschunterricht in Japan.* Nihon dokubungakkai Doitugo Kyouikubukai kaihou (日本独文学会ドイツ語教育部会会報) 2000, 118-123.

Reinelt, R. (2007a). Doitugo kotoshiken no doitugo bokokugo-washa ni yoru hyoka (in Jap.)ドイツ語口答試験のドイツ語母語話者による評価 „Muttersprachlerbeurteilung von Sprechprüfungen im Deutschunterricht"(native speaker evaluation of German oral examinations). 2007nendo nihondokubungakkai chugokushikoku shibugakkai Tokushima daigaku 2007年度日本独文学会中国四国支部学会. 徳島大学11月10日. *German Teachers Association Chugoku Shikoku Chapter 2007 regional meeting.* Tokushima, Nov. 10, paper. Accepted for Chugoku-Shikoku Doitsu Bungaku 47.

Reinelt, R. (2007b). ichinensei no doitugokotoshiken ni okeru shizensa 一年生のドイツ語口頭試験における自然さ "Real Communication aspects in first year German as 2FL oral examinations". daiju kai nihon komyunike-shon gakkkai (CAJ) chugokushikoku shibu taikai, hiroshima daigaku igakubu Deceber 15th; 第10回日本コミュニケーション学会（CAJ）中国四国支部大会, 広島大学医学部12月15日. *Communication Association of Japan Chugoku-Shikoku Chapter 10th Annual meeting, Hiroshima University*. Hiroshima, Dec. 15, paper.

Reinelt, R. (2008a in press). Kotoshiken kyouka no tameni mishugaikokugokyouiku ni okeru bubunteki na outsourcing no kanosei – Blackboard (BB)TM no shikou kara - 口頭技能強化のために未習外国語教育における部分的なOutsourcingの可能性 – Blackboard(BB)TMの試行から ー Enhancing speaking abilities in second foreign language teaching by partial outsourcing – a try with Blackboard – Ehime University Integrated Education Center.

Reinelt, R. (2008b forthcoming). *Productive e-mails in beginner's German*. JALT Pan SIG 2007 Conference, Sendai Tohoku Bunka Gakuen University, May. 12, 2007, paper.

Silva, C. (2007). Students Culture in the Classroom. In K. Bradford-Watts (Ed.), *JALT2006 Conference Proceedings*. Tokyo: JALT. [Online] Available: <www.jalt-publications.org/archive/proceedings/2006/E096.pdf>

Smith, A.F.V. & Nederend, W. (1998). Using Oral Interviews at a Junior College. *The Language Teacher.* [Online] Available: <www.jalt-publications.org/tlt/files/98/apr/smith.html>

Sprachnachweis. (n.d.) *sprachnachweis.de.* [Online] Available: <www.sprachnachweis.de/sprachnachweis/index.do>

Stiftung Warentest (consumer goods testing fundation) .(n.d.). *Stiftung Warentest.* [Online] Available: <de.wikipedia.org/wiki/Stiftung_Warentest> and <www.test.de/≥

Uebersax, J. (2006). *Intraclass Correlation and Related Methods.* [Online] Available: <ourworld.compuserve.com/homepages/jsuebersax/icc.htm>

Uebersax, J. (2007). *Statistical Methods for Rater Agreement.* [Online] Available: <ourworld.compuserve.com/homepages/jsuebersax/agree.htm>