UNIVERSITÄT LEIPZIG

Institute for Test Research and Development
Professor Dr. Erwin Tschirner
Dr. Olaf Bärenfänger
Irmgard Wanner, M.A.
☎ 0341 9737 570
🖷 0341 9737 547

Universität Leipzig, Herder-Institut
Beethovenstraße 15, 04107 Leipzig

Leipzig, May 19, 2012

Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc)

Technical Report 2012-US-PUB-1

Prepared for:

Language Testing International
White Plains, NY

Prepared by:
Institute for Test Research and Development

Dr. Erwin Tschirner
Gerhard-Helbig-Professor of German as a Foreign Language

Dr. Olaf Bärenfänger
Director, Language Learning Centre

Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc)

Olaf Bärenfänger, Erwin Tschirner

Introduction

The ACTFL Oral Proficiency Interview (OPI) is a valid and reliable interactive direct assessment of oral proficiency conducted face-to-face or by telephone between a certified ACTFL OPI tester and the test taker. The tester follows a standardized protocol, eliciting speech samples for specific purposes and rating them according to tasks and functions performed, the linguistic quantity and quality of the speech samples elicited, their text type, their context and content. A standard OPI takes between 20 and 30 minutes.

The ACTFL Oral Proficiency Interview by computer (OPIc) is an internet delivered, semi-direct, individualized assessment of functional speaking skills, based on the OPI. The OPIc has many of the same features as the OPI with the exception of the presence of a live tester. A carefully designed computer program selects prompts that are asked of the test taker by an avatar figure to elicit a ratable sample of speech. Both the OPI and the OPIc are recorded and rated according to the ACTFL-Proficiency Guidelines—Speaking. Correlations between the OPI and the OPIc are significant and strong indicating that they both are of comparable reliability and that rating outcomes are highly consistent across human- and computer-administered interviews (Surface et al.: 2009).

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001) (CEFR) is used throughout Europe and increasingly in other parts of the world for educational standards, for curriculum and textbook development, and for assessment purposes. All major European test publishers adhere to the levels established by the CEFR and rate functional language skills according to these levels. Because of the history of the development of the CEFR and because both the CEFR and the ACTFL Guidelines were developed to describe and measure the same construct, “functional language proficiency”, both systems are comparable and there should be correspondences between the levels defined by either system. Establishing correspondences between tests and the scales they refer to is a process of linking.

To link any test to the CEFR, the Council of Europe developed a comprehensive manual (Figueras et al.: 2009). The linking process allows test takers to be categorized in terms of the proficiency levels of the CEFR. The process of linking individual speech samples empirically to the scales of the CEFR is called *standard setting* or *benchmarking*. After having trained raters on the scale system with calibrated items, these raters are asked to link test items to individual CEFR scales.

This is a report of a study conducted with six expert raters of CEFR oral proficiency tests in German at the University of Leipzig, Germany, to assess evidence of validity of assigning CEFR ratings to German speech samples that were elicited and rated following the official ACTFL pro-

tocol for the OPI and OPIc. This study follows the benchmarking protocol established by the Council of Europe (Figueras et al.: 2009) to link the ACTFL OPI and OPIc to the CEFR.

The study took place at the University of Leipzig on September 16-17, 2011. The whole benchmarking procedure consisted of familiarization, calibration, and benchmarking and lasted 14 hours and 30 minutes (including breaks).

The raters used in this study were very experienced tester trainers and testers for *The European Language Certificates* (telc). *Telc* is a member of the Association of Language Testers in Europe (ALTE) and it administers and rates foreign language proficiency tests in nine European languages including German, English, Spanish, French, Portuguese, Italian, Russian, Czech, and Turkish according to the CEFR. Because interrater reliability across languages is high between *telc* tester trainers, this study has implications not only for German, the language investigated in this study, but also for the other languages that are part of the *telc* suite of languages.

Method

Participants

The participants were six raters, five of them female and one male. They were between 33 and 60 years of age with an average of 46 years (standard deviation: 10.25). All raters had a university degree in German Studies and extensive teaching experience at CEFR levels A1 to B2; five of the six raters also taught at the C1 level. All raters had extensive experience in rating tests ranging from 4 to 21 years with an average of 13.5 years (standard deviation: 7.66 years). Four raters had extensive testing and rating experience at CEFR levels A1-C1, the other two had extensive testing and rating experience at CEFR levels A1-B2. Four raters worked as tester trainers and two as testers for *The European Language Certificates* (telc). Two raters had taken part in benchmarking sessions more than fifteen times, one rater eight times and another rater three times. For two raters, it was their first benchmarking experience.

Design

The study uses the canonical benchmarking procedure specified by the Council of Europe in the *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR)* (Council of Europe, 2009). The *Manual* describes the benchmarking procedures to be used to link any given test or rating scale to the scale system of the CEFR. The benchmarking procedure consists of three phases: (1) familiarization of the raters with the CEFR scale system on the basis of speech samples with calibrated CEFR ratings; (2) calibration in which the raters rate speech samples with calibrated CEFR ratings individually in order to determine the reliability of their ratings; (3) benchmarking in the narrow sense in which raters judge speech samples with unknown CEFR ratings. The results of the benchmarking were subsequently compared to the official ACTFL ratings of the samples that had been established previously according to ACTFL standards. Reliability estimates were calculated both for the calibration and the benchmarking data.

Familiarization: During familiarization, raters first received background information on the goals and procedures of the study. Then, the characteristics of the CEFR levels for speaking were

discussed in detail. Subsequently raters rated six German speech samples from official *telc* speaking tests. Each speech sample represented one CEFR level. After determining the CEFR level individually, raters discussed their ratings until they were satisfied that they all agreed on the same rating. The CEFR rating all raters agreed on was identical with or close to the official *telc* rating.

At the end of the familiarization phase raters were asked to complete a biographical questionnaire. They answered questions on their university degrees, present employment, teaching experience, CEFR levels taught, testing experience, CEFR levels tested, and prior experience with benchmarking sessions. The raters were also allowed to comment on any aspect of the study. The complete familiarization phase lasted 1:30 hours.

Calibration: The goal of the calibration is to determine raters' rating accuracy and reliability. For this purpose, participants rated 12 randomly presented speech samples using both the CEFR global speaking scale and the detailed speaking scale that contains the following four criteria: linguistic range, accuracy, fluency, and coherence. Raters were also asked to indicate on a four point Likert scale (1) how easy it was for them to rate the speech sample; (2) how well they could understand the speaker; and (3) how confident they were of their rating. The Likert scale ranged from 1 = easy/high/confident to 4 = not easy/not high/not confident. The calibrated speech samples were taken from *Mündlich* (Bolton et al.: 2008). This publication contains speech samples together with the results of benchmarking sessions with a group of 28 language teaching and testing experts that is commonly used in benchmarking activities by the Goethe Institute and *telc*, both members of the Association of Language Testers in Europe (ALTE). By comparing the calibrated ratings to the individual ratings of each participant, it was possible to calculate reliability measures for each rater. As a threshold value, a Spearman's *rho* of higher than .85 had been established as criterion to include the rater in the study. Because all raters exceeded this coefficient, all raters were included into the final analysis of the benchmarking data. The calibration phase lasted two hours.

Benchmarking: During the benchmarking in the narrow sense, the raters individually rated 54 OPIc and OPI speech samples, six each on the levels Novice Mid (NM), Novice High (NH), Intermediate Low (IL), Intermediate Mid (IM), Intermediate High (IH), Advanced Low (AL), Advanced Mid (AM), Advanced High (AH), and Superior (S). The samples used were extracts from official OPIc and OPI. OPI and OPIc contain samples of what a candidate can do (level checks) and what they cannot do (probes). Probes usually lead to linguistic breakdown, i.e., the candidate cannot discuss the question at the level required or does not discuss it at all. Breakdown is often associated with a great deal of linguistic error. Because the benchmarking study was supposed to ascertain the highest CEFR level at which the candidates would function adequately, only responses to level checks were selected for the purpose of this study.

The OPIc contains 16 prompts. Responses are timed. The candidate has 1 minute to respond to Intermediate prompts, 2 minutes to respond to Advanced prompts, and 2:30 minutes to respond to Superior prompts. Between four and six responses to prompts were selected to be included in a sample depending on the level and depending on how much a particular candidate said in response to a particular prompt. The goal was to provide raters with approximately 5 minutes of speech at the levels Novice and Intermediate and 8 minutes of speech at the levels Advanced and Superior. At the Novice and Intermediate levels, 5-6 responses were selected for

a total of approximately 5 minutes of speech and at the Advanced levels, 4-6 responses were selected for a total of approximately 8 minutes of speech. Because the German OPIc had only been recently introduced, there were not enough candidates at the two highest levels to yield appropriate samples. Therefore two AH samples and all six S samples were taken from ACTFL Oral Proficiency Interviews. From each interview, approximately eight minutes of speech were selected that were most representative of the final ACTFL rating.

In order to avoid serial effects, raters were divided in two groups. The speech samples for each group were administered in random order, in different random orders for each group. For the rating procedure, raters used the same forms as in the calibration phase. These forms included the global rating, the detailed ratings, and the qualitative questions with respect to raters' rating confidence etc. (see above). For each test candidate, there was one form that also included the written OPIc prompts for the responses of a particular candidate or the provision that the question would be asked by the OPI interviewer. Raters heard the speech samples over headphones and were advised to listen to each sample only once. On average, they had 10 minutes for one speech sample. The complete activity lasted eight hours and took place over two days.

After the benchmarking activity, there was a focus group discussion in which the participants discussed any aspect of the rating procedure that they cared to comment on. A research assistant took notes during the discussion. In addition, the session was recorded in MP3 format. The focus group discussion lasted 30 minutes.

Statistical Analyses

In order to determine the *reliability* of the CEFR ratings, the following analyses were carried out.

Kolmogorov Smirnov Test: This is a non-parametric test for ordinal data that is commonly used to compare the distribution of ratings in two different samples. If the Kolmogorov-Smirnov Test yields a significant result for a rater, he has a preference for a specific category. If, on the contrary, the Kolmogorov-Smirnov Test does not have a significant result, the rater makes a balanced use of the category. Accordingly, a negative Kolmogorov-Smirnov test is an indicator for reliable ratings.

Spearman's *rho*: This measure assesses the extent to which a relationship between two variables may be described as a monotonic function. It is commonly computed for analyses involving ordinal data. Because the final ratings of the two tests cannot be assumed to be interval data, the calculation of *rho* is appropriate here. *Rho* is calculated by ordering the data by rank and by subsequently correlating the two rank orders. *Rho* is therefore also called a rank-order correlation. Its results are interpreted in a similar way as Pearson's correlation.

Kendall's *tau*: The relationship between final ratings was also analyzed using Kendall's *tau*, a measure of agreement. Like Spearman's *rho*, *tau* is calculated on the basis of rank orders. Whereas *rho* is based on the proportion of variability accounted for, *tau* is a measure of agreement. Thus, *tau* expresses the difference between the probability that participants are rated in the same order and the probability that participants are rated in different orders. Similarly to *rho*, a *tau* value of 1 stands for a perfect correspondence between the two tests and 0 for a non-existing correspondence.

Goodman Kruskal's *gamma*: This is another measure of agreement. Unlike Kendall's *tau*, *gamma* ignores bindings, i.e., cases when two participants are assigned the same rank (e.g., because they both received an A2 rating). Considering the fact that the final rating only consists of six values in the case of the CEFR (A1, A2, B1, B2, C1, C2) and nine values in the case of ACTFL (NM, NH, IL, IM, IH, AL, AM, AH, S), a computation method that is insensitive to bindings provides helpful insights. As a consequence of the calculation method, *gamma* tends to be higher than Kendall's *tau*.

Kendall's *W*: Kendall's *W* is sometimes also called coefficient of concordance. Kendall's *W* is the single measure of interrater agreement between more than two raters on the ordinal data level. A Kendall's *W* of 1 indicates complete agreement between all raters, a Kendall's *W* of 0 complete disagreement. This measure may be interpreted similarly to Cronbach's *alpha*.

In order to determine *correspondences* between CEFR and ACTFL ratings, raw percentages of agreement were cross-tabulated for ACTFL and average CEFR ratings. To determine the most likely CEFR level, if a test takers ACTFL level is known, an ordinal regression model was estimated. In addition, the reliability estimates described above were calculated.

Data analysis

Rater Reliability and Agreement of the Calibration Data

A Kolmogorov Smirnov Test was performed to determine whether raters had a preference for a specific category. Because the Kolmogorov Smirnov Test did not yield any significant results for any of the raters, it may be assumed that all raters made a balanced use of the categories.

For each rater, correlation measures of his or her ratings with the official CEFR rating of the calibration samples were calculated using Spearman's *rho*, Kendall's *tau* and Goodman Kruskal's *gamma*. The results of these analyses are reported in Table 1. As can be seen, all raters show very high rating reliability and agreement.

Table 1
Correlation Measures Between Raters and the Calibrated CEFR Rating

	Spearman's *rho*	Kendall's *tau*	Goodman Kruskal's *gamma*
Rater 1	.879	.794	.920
Rater 2	.883	.791	.882
Rater 3	.906	.814	.889
Rater 4	.971	.933	.966
Rater 5	.961	.908	1.0
Rater 6	.968	.923	1.0

Note: All correlation measures are significant at the $p < .01$ level.

The excellent quality of rater reliability is corroborated by Kendall's *W*. Kendall's *W* is a measure of rater reliability for more than two raters. Kendall's *W* coefficient of concordance is .934 with $p < .001$.

All reliability measures show a very strong relationship between raters' ratings and the calibrated CEFR rating. Because all raters exceeded a Spearman's *rho* correlation of .85, the data of all six raters were included in the final analysis.

Correspondences between CEFR and ACTFL Ratings

Table 2 shows ACTFL ratings cross-tabulated with CEFR ratings. There were six raters and six samples at all ACTFL levels from NM to S for a total of 36 CEFR rating possibilities for each ACTFL level. All raters, for example, rated all NM samples A1 for a total of 35 ratings. (One sample was not rated by one rater.) All NH samples were rated A1 28 times; they were rated A2 8 times. IL samples were rated A1 13 times; they were rated A2 21 times; and they were rated B1 2 times. IM samples were rated A2 10 times; they were rated B1 25 times and B2 1 time. IH samples were rated B1 28 times; they were rated B2 7 times. AL samples were rated B1 13 times; they were rated B2 21 times and C1 2 times. AM samples were rated B2 15 times; they were rated C1 14 times and C2 4 times. AH samples were rated B2 3 times; they were rated C1 11 times and C2 19 times. Superior samples were rated C1 3 times and C2 33 times.

Table 2
ACTFL and CEFR ratings (all ratings)

		CEFR Rating						
		A1	A2	B1	B2	C1	C2	Total
ACTFL Rating	NM	35						35
	NH	28	8					36
	IL	13	21	2				36
	IM		10	25	1			36
	IH			28	7			35
	AL			13	21	2		36
	AM				15	14	5	34
	AH				3	11	19	33
	S					3	33	36
Total		76	39	68	47	30	57	317

Table 3 shows ACTFL ratings cross-tabulated with the mode of the CEFR ratings, i.e., with the most frequently assigned CEFR rating. For this purpose, the six verbal ratings were converted into numeric values and the most frequent value was identified. CEFR ratings were converted into numeric values as follows: A1 = 1, A2 = 2, B1 = 3, B2 = 4, C1 = 5, C2 = 6.

Table 3
ACTFL and CEFR ratings (mode of CEFR ratings)

		CEFR Rating						
		A1	A2	B1	B2	C1	C2	Total
ACTFL Rating	NM	6						6
	NH	5	1					6
	IL	2	4					6
	IM			6				6
	IH			6				6
	AL			2	4			6
	AM				3	3		6
	AH					3	3	6
	S					1	5	6
Total		13	5	14	7	7	8	54

Table 3 shows that both NM and NH are most frequently associated with A1. IL is most frequently associated with A2 and both IM and IH are clearly associated with B1. AL is most frequently associated with B2 and AM is associated with both B2 and C1. AH is associated with both C1 and C2, and S is most frequently associated with C2.

As Table 4 shows, there is a strong correlation between ACTFL and CEFR ratings. All measures of correspondence and agreement are significant and very strong.

Table 4
Correlation and Agreement Measures Between CEFR and ACTFL Ratings

	Spearman's *rho*	Kendall's *tau*	Goodman Kruskal's *gamma*
ACTFL and CEFR rating	.966	.911	.968

Note: All correlation and agreement measures are significant at the $p < .01$ level.

The excellent quality of rater reliability is also corroborated by Kendall's *W*. Its value is .969 with $p < .001$.

To determine the most likely CEFR rating if a test taker's ACTFL OPI level is known, an ordinal regression model was estimated (for a description of this analytic methodology, see Agresti, 2002). Like ordinary least squares regression, ordinal regression estimates a prediction equation that maximizes the prediction of the outcome (CEFR rating, in this case) using specific predictors (ACTFL OPI level, in this case). The results indicated ACTFL OPI level was a statistically significant predictor of CEFR rating. The model with the ACTFL OPI level predictor fit the data significantly better ($\chi^2 = 156.4$, $df = 8$, $p < .001$) than a model with no predictors (i.e., intercept only).

Using the regression model coefficients, the probability of observing each CEFR rating based on the ACTFL OPI level was calculated (see Agresti, 2002). These model-based probabilities are presented in Table 5.

Table 5
Probability of Being at Each CEFR Level Based on ACTFL Level

	Predicted Probabilities from Ordinal Regression					
ACTFL Level	Probability of A1	Probability of A2	Probability of B1	Probability of B2	Probability of C1	Probability of C2
NM	.999	.001				
NH	.833	.166	.001			
IL	.333	.662	.005			
IM		.003	.995	.003		
IH		.003	.995	.003		
AL			.333	.662	.004	
AM			.002	.498	.498	.002
AH				.002	.498	.500
S					.166	.833

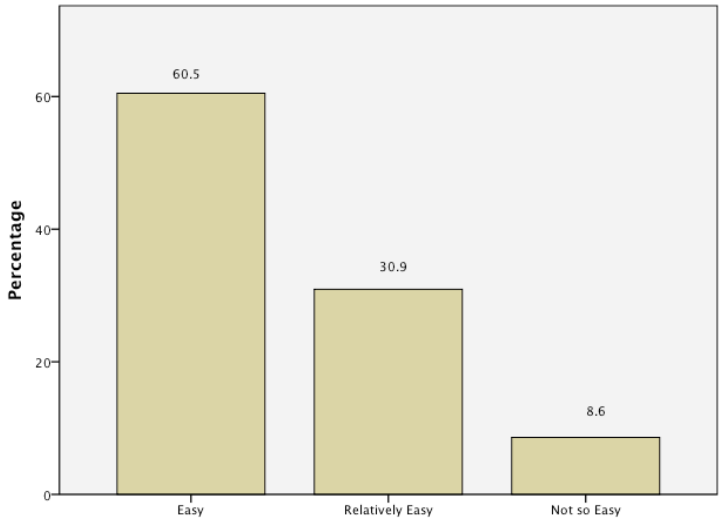
Note: Variance explained by the model (Pseudo R squared): Cox and Snell: .945; Nagelkerke: .976; McFadden: .839

As Table 5 shows, the probability of someone receiving an NM rating to be rated A1 is 99.9%, of someone receiving an NH rating to be rated at least A1, it is 100%. The probability of someone receiving an IL rating to be rated at least A2 is 66.7%, of someone receiving an IM rating to be rated at least B1, it is 99.8%. The probability of someone receiving an IH rating to be rated at least B1 is also 99.8%, of someone receiving an AL rating to be rated at least B2, it is 66.6%. The probability of someone receiving an AM rating to be rated at least B2 is 99.8%, 49.8% of test takers receiving an AM rating will be rated C1. The probability of someone receiving an AH rating to be rated at least C1 is 99.8%, 50% of test takers receiving an AH rating will be rated C2. Finally, the probability of someone receiving a S rating to be rated C2 is 83.3%.

Perceptions of the Rating Process by the Raters

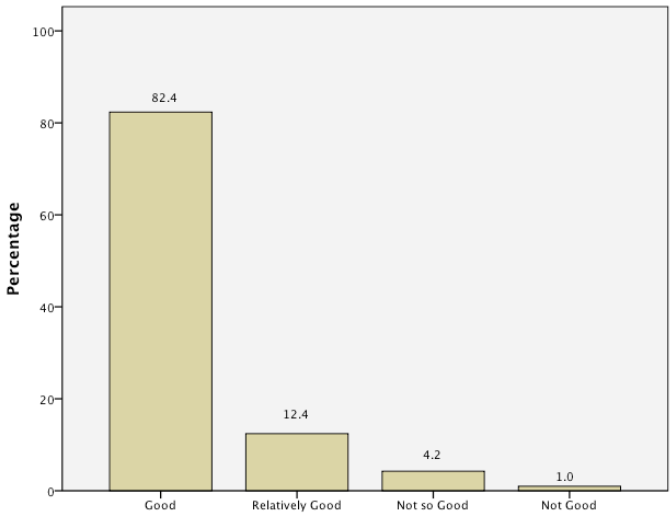
Following each rating, raters indicated on a four point Likert scale how easy it was for them to rate the sample; how well they could understand the candidate in terms of acoustic quality; and how confident they were of their rating. The verbal descriptors were transformed into numbers using the following algorithm: “easily/well/confident” = 1; “relatively easily/well/confident” = 2; not so easily/well/confident” = 3; “not easily/well/confident” = 4. Graph 1 depicts the frequency of answers for all raters regarding the ease of the rating process.

Graphic 1
Ease of Rating



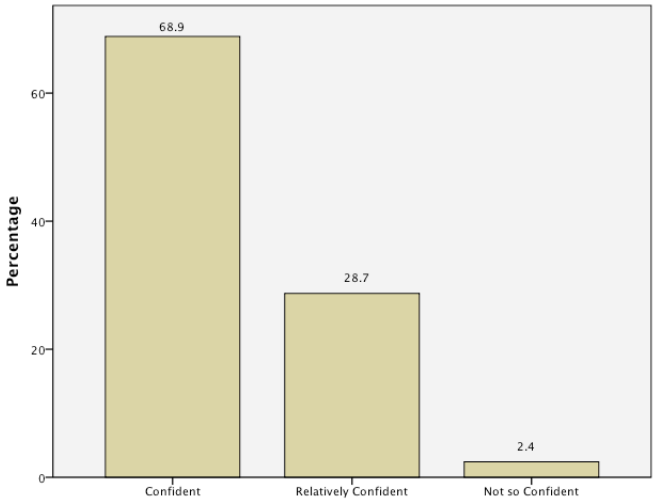
As Graphic 1 shows, 91.4% of the ratings were thought to be easy or relatively easy. Only 8.6% of the samples were considered to be not so easy to rate. No sample was considered to be not easy to rate. The ease of rating is also reflected in participants' ease of understanding the speech samples as displayed in Graphic 2.

Graphic 2
Understandability



Raters were able to understand 94.8% of all speech samples well or relatively well. Only 4.2% of the speech samples were considered to be not so well understandable and 1.0% were not well understandable. As Graphic 3 shows, both the ease of rating and the ease of understanding the samples also led raters to feel very confident of their ratings.

Graphic 3
Confidence



Raters felt confident or relatively confident about 97.6% of their ratings. They felt not so confident about only 2.4% of their ratings

Insights from the Focus Group

During the focus group discussion, participants were asked to comment on anything in particular that they had observed during the rating procedure, on any difficulties they may have had in the rating process, and to make suggestions for future benchmarking activities. Some participants mentioned having difficulty clearly differentiating between the levels A1 and A2, B1 and B2, and C1 and C2, respectively, at times. According to them, especially CEFR C1 and C2 descriptors appear to be underspecified. They also expressed a need to have clearer descriptors to distinguish B1 from B2 and to have more complex tasks at the B2 level. Because the tasks at the level that was associated with B2 were close to every-day life, some of the participants felt that they did not always have enough information to assign a B2 rating.

Assigning CEFR Ratings to ACTFL OPI and OPIc

The study shows that experienced CEFR raters are able to assign CEFR ratings to OPIc and OPI samples reliably. The study also suggests that there are clear correspondences between CEFR and ACTFL ratings. At several levels, e.g., NH, IM, IH, and Superior, ACTFL and CEFR levels are closely aligned, at other levels, e.g., AM and AH, ACTFL levels span two CEFR levels. Because OPI and OPIc are primarily used in high stakes testing, it is suggested that correspondences are set conservatively, i.e., that the lower of two levels is chosen in cases where the ACTFL level corresponds to two adjacent CEFR levels. For example, the probability of someone receiving an AM rating to be rated B2 is 49.8%, to be rated C1, it also I 49.8%. Therefore, it is recommended to

use the more conservative correspondence of AM being at least B2. (In fact, the probability of someone receiving an AM rating to be rated B2 is 99.6%.) The higher CEFR level is recommended only when the probability of someone reaching that level is at least 66%. See Table 6 shows the correspondences suggested by the present study.

Table 6
Assigning CEFR Ratings to OPIc and OPI Ratings

ACTFL OPIc and OPI	*CEFR*
Novice High	A1
Intermediate Low	A2
Intermediate Mid	B1
Intermediate High	B1
Advanced Low	B2
Advanced Mid	B2
Advanced High	C1
Superior	C2

Despite the fact that all NM samples were judged to be A1 by all raters, it is not recommended that a CEFR rating is given for an ACTFL NM rating. The majority of NM samples were judged to be barely A1 indicated by a minus rating raters felt necessary to add. Table 6 suggests to assign the CEFR rating A1 to NH; A2 to IL; B1 to both IM and IH; B2 to both AL and AM; C1 to AH; and C2 to Superior.

As the CEFR also allows for a finer differentiation at the levels A2, B1, and B2, dividing them into A2.1 and A2.2, B1.1 and B1.2, and B2.1 and B2.2, respectively, the following correspondences may be used instead (see Table 7)

Table 7
Assigning CEFR Ratings to OPIc and OPI Ratings

ACTFL OPIc and OPI	*CEFR*
Novice High	A1
Intermediate Low	A2
Intermediate Mid	B1.1
Intermediate High	B1.2
Advanced Low	B2.1
Advanced Mid	B2.2
Advanced High	C1
Superior	C2

Conclusion

Experienced CEFR raters are able to assign CEFR ratings to OPIc and OPI samples very reliably. All measures investigated in this study indicate a strong correspondence between CEFR and ACTFL ratings. There are clear correspondences between CEFR and ACTFL ratings at the levels Novice High, Intermediate Low, Intermediate Mid, Intermediate High, Advanced Low, and Superior. At the levels Advanced Mid and Advanced High, ACTFL ratings align with two CEFR levels. In high stakes testing, it is suggested to set correspondences conservatively and to use the lower of the two corresponding CEFR levels. The correspondences established in this study are the following: NH = A1, IL = A2, IM and IH = B1, AL and AM = B2, AH = C1, and S = C2. If a finer delineation is desired, IM may be associated with B1.1 and IH with B1.2, and similarly, AL with B2.1 and AM with B2.2.

Because the raters used in this study were very experienced tester trainers and testers for *The European Language Certificates* (telc), a member of the Association of Language Testers in Europe (ALTE), this study has implications not only for German, the language of the study, but also for the other languages that are part of the *telc* suite of languages. These include English, Spanish, French, Portuguese, Italian, Russian, Czech, and Turkish.

Bibliography

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.

Bolton, S., Perlmann-Balme, M., Steiner, S., Glaboniat, M., & Lorenz, H. (2008). *Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*. Berlin: Langenscheidt.

Council of Europe (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR).* Strasbourg: Language Policy Division. Available: http://www.coe.int/t/dg4/linguistic/manuel1_en.asp#P15_1111.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. Language Testing, 22, 261-279.