Universiteit
Leiden
The Netherlands

Objective assessment of students' interpreting performance: an experimental study

Yenkimaleki, M.; Heuven, V.J.J.P. van

Citation
Yenkimaleki, M., & Heuven, V. J. J. P. van. (2023). Objective assessment of students' interpreting performance: an experimental study. *Teaching English Language*, *17*(1), 227-267. doi:10.22132/TEL.2022.164846

Version:	Publisher's Version
License:	Creative Commons CC BY-NC 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3621149

Note: To cite this publication please use the final published version (if applicable).

Teaching English Language Journal
ISSN: 2538-5488 – E-ISSN: 2538-547X – http://tel.journal.org
© 2023 – Published by Teaching English Language and Literature Society of Iran

Please cite this paper as follows:
Yenkimaleki, M., & Heuven, V. J. van (2023). Objective assessment of students' interpreting performance: An experimental study. *Teaching English Language, 17*(1), 227-265. https://doi.org/10.22132/TEL.2022.164846

Research Paper

Objective Assessment of Students' Interpreting Performance: An Experimental Study

Mahmood Yenkimaleki[1]
Nahavand Higher Education Complex, Bu-Ali Sina University, Iran
Vincent J. van Heuven
Leiden University, The Netherlands
University of Pannonia, Hungary

Abstract

The traditional metric of interpreting quality is a score given by human professional judges focusing on the interpreters' performance. However, there is a poor agreement on what constitutes an acceptable interpretation. This study investigates the objective assessment of interpreter trainees' performance. Two groups of 15 student interpreters were formed. Participants were assigned to groups at random, but with equal division between genders (seven males in each group). The control group was taught interpreting skills by the routine curriculum, while the experimental group spent part of the time instead on theoretical explanation and practical exercises emphasizing prosodic differences between Persian and English. Three raters assessed the quality of the interpreter trainees' performance in a post-test. Then the interpreting performance of the students was assessed objectively through Praat software. The results show that the intersubjective ratings of the students' interpreting performance can be adequately predicted from objective measures through multiple linear regression. These results have implications for designers of curricula for training interpreters, and material producers in interpreting education.

[1] Corresponding author: m.yenkimaleki@basu.ac.ir

Keywords: Assessment in Interpreting Performance, Objective Correlates, Interpreting Performance, Prosody Teaching

Received: November 11, 2022
Accepted: January 18, 2023

1. Introduction

The present study investigates an experimental approach to the objective assessment of interpreter trainees' performance. The interpreter trainees were taught English prosody explicitly. Pronunciation is the crucial element of speech, encompassing the properties of speech beyond individual sounds. The appropriate use of prosodic features such as stress and intonation reported to be more crucial for intelligibility than the accurate production of individual segments, or sounds (Kuronen & Tergujeff, 2018). Teaching segments is the primary need for second language learners but numerous studies suggest the more important role of prosody in second language speech perception and production (e.g., Yenkimaleki & Van Heuven, 2018, 2019; Kuronen & Tergujeff, 2018). The key to building interpreting expertise lies in improving the efficiency of the interpreter's perception and production skills in the L2 (e.g., Yenkimaleki & Van Heuven, 2022). Prosodic feature awareness training can be helpful for interpreters both in speech production and speech recognition (Yenkimaleki & Van Heuven, 2019b).

This study attempts to relate the intersubjective expert judgments to objective measures that can be expected to correlate with the judgments. If such correlates can be found, the expert judgment can be predicted by some combination of objective correlates. If the prediction is sufficiently accurate, expert judgments could be dispensed with in the future and be replaced by objective measurements. We have investigated the relationships between the expert judgments of the quality of the participants' interpreting performance

on the one hand and objective correlates of their performance on the other. Over the past decade, testing and assessing spoken-language interpreting has garnered an increasing amount of attention from stakeholders in interpreter education, professional certification, and interpreting research. This is because in these fields assessment results provide a critical evidential basis for high-stakes decision making, such as the selection, certification, and confirmation/refutation of a research hypothesis. However, few studies have addressed the systematic objective assessment for interpreting performance. Therefore, the present study is set up to examine this aspect in detail so that to shed light more on the objectivity in interpreting performance assessment.

2. Review of related literature

2.1 Teaching English language prosody

Prosody is the ensemble of properties of speech that cannot be derived from the mere sequence of phonemes that make up a spoken sentence. Prosody then includes such phenomena as lexical tone, stress at the word, and at the sentence level, boundary marking, and intonation. All these suprasegmental phenomena are characteristics of linguistic units larger than a single vowel or consonant, namely larger than a segment (Van Heuven & Sluijter, 1996; Nooteboom, 1997; Van Heuven, 2017, 2018, 2022). Although words are recognized mainly from the sequence of segments, word-level prosody assumes a critical role in the recognition process when the segmental quality is poor, as is typically the case in foreign-accented speech (e.g., van Heuven, 2008, 2022; Cutler, 2012; Yenkimaleki et al., 2022). Moreover, sentence prosody is often indispensable in the signaling of the speaker's intention (e.g., O'Neal, 2010). Prosody plays an important role in the decoding and encoding of meaning. Segmentation of continuous speech into syllables, words, and phrases, informing syntactic structure, and emphasizing content words and other salient information are some of the functions of

prosody that facilitate the processing of speech (Whalley & Hansen, 2006; Yenkimaleki et al., 2021). For successful decoding of input speech and encoding speech output in the non-native language, the L2 learner will benefit from an explicit comparison of the prosodic properties of his native language and those of the L2 (Yenkimaleki & Van Heuven, 2020, 2019b).

Many researchers have emphasized the importance of awareness and 'consciousness-raising' for second language learning (e.g., Schmidt, 2010; Yenkimaleki, 2018). Mainstream cognitive psychologists consider awareness a fundamental pre-condition to learning and even claim that that learning is impossible without conscious awareness (Dawson & Schell, 1987). In the field of foreign-language education, these views are reflected by, for instance, Bialystok (1978), who proposed a theoretical framework in which consciousness knowledge plays a key role. In a similar vein, Rutherford et al. (1985) asserted that drawing the learner's conscious attention to the formal properties of the foreign language can be advantageous to second language learning. These perspectives could be applied to prosody awareness training for interpreters in order to make interpreters have conscious knowledge of prosodic features in reducing the number of competing representations of the incoming structures they have to entertain in working memory while having interpretation performance.

Prosody awareness training is the most marginalized activity in the training of interpreters though prosody plays a critical role in communicating the message. The neglect of prosody awareness training for interpreters may be due to the (apparent) complexity of this issue and the misconception about what content should be taught and how this could be done (Yenkimaleki & Van Heuven, 2017, 2018, 2019a, b). The reason is that the practitioners in EFL contexts find it difficult to listen analytically to the students' pronunciation, identify errors and suggest remedies, or because they give

priority to other aspects of communicative competence such as the acquisition of vocabulary and morphosyntax. Jackson and O'Brien (2011) maintain that the relationships between prosody, second language speech production, and second language comprehension are understudied and need more investigation. Systematic studies should be done to learn how interpreters may exploit the relationships between prosody and meaning when decoding messages in the source language and encoding the same message in the target language.

2.2 Assessment in interpreting

Interpreting actualizes the facilitation of verbal communication between different languages. Pöchhacker (2004) stated that interpreting is an immediate form of translational activity, performed for the benefit of people who want to engage in communication across barriers of language and culture. Assessment is fundamental to ensure the quality of interpreting in both the professional practice and educational training. Assessment plays an important gatekeeping role to ensure that only suitably qualified interpreters are endorsed to enter the job market, such as the professional examinations in the field of the interpreting profession, and the final examinations at the end of a training course (Wu, 2010). In interpreter training programs, it is important that assessment procedure be the crucial part of the training process. Some researchers doing empirical work have paid attention to assessment to evaluate quality of interpretation performance (e.g., Kopczynski, 1994; Garzone, 2002; Gile, 2005). The issue is to design an assessment method that can be compatible with the overall aims of the interpreter training program, can effectively assess the learning objectives of the training program, and support the development of students' professional competencies (i.e., the interpreting skills and the relevant knowledge about the profession) (Wu, 2010).

Therefore, it is important to understand how this practice of judging-by-impression may affect the validity and reliability of the interpreting examinations. At a live examination panel, the examiners perceive and judge many components in a simultaneous interpreting performance, such as the fidelity of the target-language speech, the quality of the interpreter's linguistic output, the quality of his or her voice, the prosodic characteristics of his or her delivery, the quality of his or her terminological usage (Gile, 1995). Given the complexity of the task itself when performing simultaneous interpreting (SI) and of the high cognitive demand on the examiners, the judgement as a result is usually made in a holistic and subjective manner, which has raised concerns about the consistency of the judgment process of the examiners (Wu, 2010). Serious concerns have been raised about how consistently professionals in the field of interpreting can exercise their judgement when it comes to assessing interpreting performances (Sawyer, 2004). Performance assessment has long been criticized as unreliable and in need of systematic study (Campbell & Hale, 2003; Etemadi & Abbasian, 2023) and the concerns about the problematic role of professional judgment are mainly due to its subjective nature (Messick, 1989; Khodashenas et al., 2023).

The traditional metric of interpreting quality assessment is human scoring which is a professional evaluation by interpreting judges focusing on interpreters' performance in the booth such as fluency and adequacy of their translations, on-site response and interpreting skills. However, there is poor agreement on what constitutes an acceptable interpretation. Some judges regard an interpretation as unacceptable if a single word choice is suboptimal. At the other end of the scale, there are judges who will accept any translation that conveys the approximate meaning of the sentence, irrespective of how many grammatical or stylistic mistakes it contains. Without specifying more

closely what is meant by acceptable, it is difficult to compare evaluations (Zhang, 2016).

We address the usability of objective assessment procedure for the quality of interpretation performance. Motivated by the above issues, we set up this experiment for systematic measurement of interpreter trainees' performance so that to relate the intersubjective expert judgments to objective measures that can be expected to correlate with the judgments. If such correlates can be found, the expert judgment can be predicted by some combination of objective correlates. If the prediction is sufficiently accurate, expert judgments could be dispensed with in the future and be replaced by objective measurements.

3. Method

3.1 Participants

The participants were 30 second-year students who were chosen randomly out of 64 BA students of English Translation and Interpreting at the University of Applied Sciences, Tehran, Iran. They were then divided into two classes of 15 (7 males in each group) one of which served as the experimental group and the other as the control group. All participants, aged between 20 and 22, were native speakers of Persian and took part in all training sessions during the semester.

3.2 Ethical issues

We received approval from the ethics committee of the Dept. of English Language for the present study. All the participants agreed to take part in the research project on the basis of informed consent.

3.3 Procedure

The interpreter trainees who participated in the study were divided into two groups at random. One group was designated as the experimental group while the other was considered the control group. At the beginning of the

program, all participants took a TOEFL test in order to establish that they were homogeneous. The test battery was the standard Longman's TOEFL English proficiency test, with separate modules testing the learner's (i) Listening comprehension, (ii) Reading comprehension, and (iii) Structure and writing skills. Then, the control group and experimental group took a pretest on interpreting so that their level of expertise in interpreting was assessed prior to receiving any type of training. In the next stage, the control group received instruction and practice about the techniques of interpreting, different aspects of interpreting, and types of interpreting. The experimental group received not only the same instruction as provided to the control group (in less time, however) but also awareness training on prosodic features (stress at word and at sentence level) of English and their effect on their performance. The prosody awareness training targeted the differences between the stress systems of English and Farsi, at the word and sentence level, in a cognitive way. Theoretical explanation was given and immediately put into practice in exercises. Both experimental and control groups received exercises in interpreting by presenting authentic audio extracts. The experimental group received awareness training of prosodic features of English for 20 min each session and the control group received additional practice in consecutive interpreting through practical exercises. To receive feedback regarding the progress of teaching and to detect weaknesses in methodology, different types of formative test were administered in all the sessions. At the end of the program, a post-test with the same structure but with different items compared with the pretest was administered to both groups in order to establish whether the treatment (explicit teaching of English prosody) of the experimental group had been effective or not. Six authentic audio extracts of IRIB (Islamic Republic of Iran Broadcasting) news cast in Standard Persian were chosen as test materials for the interpretation task. Through random sampling four of these were selected for the pretest. Each fragment lasted 30s. The other two fragments were used as

the post-test. The procedure used in the pretest and the post-test was the same. Students were seated in sound-proofed half-open cubicles which attenuated ambient noise well enough to yield clean recordings. They listened to the source texts being played to them over a loudspeaker at a comfortable listening level. After every 30-s fragment, they were allowed one minute to record an interpretation of the source text in English. Recordings were made directly onto a digital computer through individual, table-mounted microphones. As part of the one-minute intervals, and also earlier while listening to the stimulus text, participants were allowed to make written notes (as is not uncommon in consecutive interpreting). The participants' performance, both in the pretest and in the post-test, was scored applying the criteria adapted from Sawyer (2004) in subjective assessment by three experts in the field of interpreter training. In the second stage we assessed students' performance objectively (see data analysis section for details).

3.4 Data analysis

We made an attempt to relate the intersubjective expert judgments to objective measures that can be expected to correlate with the judgments. If such correlations can be found, the expert judgment can be predicted by some combination of objective correlates. If the prediction is sufficiently accurate, expert judgments could be dispensed with in the future and be replaced by objective measurements.

Table 1

Ten evaluation criteria subdivided into three domains used in the quality judgment of interpreting performance. For each criterion the maximum number of point that can be awarded is specified. Maximum overall score = 100. After Sawyer (2004).

Meaning		Language use		Presentation	
Accuracy	20	Grammar	7	Pace	10
Omissions	15	Expression	7	Voice	10
Overall coherence	10	Word choice	7		
		Terminology	7		
		Accent	7		

In the set of evaluation criteria not all scales can be grounded in experimental measures. We did not try, for instance, to come up with

objective measures that might predict overall coherence of the interpretation into Persian relative to the original English text, nor will we attempt to define an objective measure for 'Expression'. However, omissions – i.e., failure to translate an important word or concept – can be counted, and the number of grammatical anomalies can be determined by analyzing transcripts of the interpretations. We also tried to establish correlates of at least some of the prosodic evaluation criteria such as accent and pace. Especially 'Pace' (or fluency) would seem to be amenable to objective testing. At least two correlates of pace will be considered, viz. speaking rate and articulation rate. Speaking rate is traditionally defined as the number of linguistic units, i.e., words or syllables, produced per unit time (per minute or per second). Here the total speaking time includes all pauses, whether silent or filled (*eh*, *ehm*). Articulation rate is computed the same way as is speaking rate but the total time does not include pauses and hesitations. Defined this way, obviously, speaking rate and articulation rate are strongly correlated. When trying to predict judgments on a rating scale from objective measures it is better to work with independent predictors, (i.e., predictors that do not or only weakly correlate with each other). It seemed to us that a feasible way to disentangle speaking rate and articulation rate would be to use articulation rate only and supplement this parameter with a more direct measure of the incidence of pauses and hesitations. This latter aspect can be adequately captured by computing the percentage of the total speaking time that is taken up by pauses. We call this latter parameter '%-pause'.

We note that it may not necessarily be the case that pace (fluency) is monotonically related to either %-pause or to articulation rate. It would seem more likely that the relationship between the judgment and the acoustic measure will be U-shaped, i.e., judgments may well be most favorable for values in the middle of the range, when the speaker does not insert a great

many pauses (indicative of difficulties in producing the interpretation) nor speaks with very few pauses (which would create a burden on the listener). Similarly, articulation rates in the middle of the range are expected to receive the most favorable judgments.

3.5 Objective measures used

We distinguish between counts of phenomena that can be established by analyzing written transcripts of the interpreter's performance (and comparing it to the original text), and measurement of acoustic properties, which, of course, cannot be done from a written transcript.

3.5.1 Count measures

Generally, the norm is that interpreters should have a complete transfer of the source text to their audience without any omission of ideas or changes of meaning. This issue has received a lot of attention in typology and error analysis of translation and interpreting performance. However, we know that in some cases, omission of some aspects in interpretation enhances the quality of interpreting and as a result, communication of message is done properly ('less is more'). Jones (2014: 139) pointed out that interpreters in some situations are not in position to render the exact and complete message. Therefore, in these situations the interpreters omit part of the source text in order to relay a coherent message to the audience. Therefore, in some cases, the interpreters intentionally omit part of the source language because they want to transfer the gist of the message so that the audience may perceive the message more easily. In interpreting, the important aspects and essentials are preferred over the completeness of the message.

It is an open question, in the present study, whether the judgment of accuracy and omissions is monotonically related to the number of words (or concepts) incorrectly translated or left out altogether. One hypothesis would be that the more accurate and complete the interpreting is, the better the

accuracy and omission judgments. We leave room, however, for a more sophisticated possibility in case the relationship between the objective counts and the global judgments should be U-shaped. In the latter case, keeping in all details would detract from the judged adequacy or optimality of the interpreting job.

The number of omissions was established by comparing an optimal translation of the original English texts into Persian with transcripts of the student's interpretation. The unit of measurement was the content word. We checked for every content word in the model translation whether it occurred in some adequate or at least acceptable form (identical, synonym or paraphrase) in the student's transcript. When the word or concept was not an acceptable stand-in for the original, it was counted as an inaccuracy or meaning error. When the word or concept was absent from the student's interpretation altogether it was scored as an omission. The total number of errors was then equal to the number of inaccuracies and the number of omissions added together.

3.5.2 Acoustic measures of pace

The sound recordings of each of the 30 speakers were segmented into interpausal units. An interpausal unit, or IPU, is defined as a stretch of speech not interrupted by a silent or filled pause (Koiso et al., 1998, Buhmann et al., 2002). In order to qualify as a pause, a silence in the spoken utterance must be longer than 100 ms. If shorter silences would also be considered, the occlusion phases of voiceless plosives would be counted as pauses, which would be undesirable.

The recordings were recoded from mp3-format to wav-format. Normally, lossy coding such as mp3 would be ill advised for the analysis of speech but in the present case, where only duration, fundamental frequency and intensity will be measured, measurements will be quite faithful. The segmentation of

the recordings was done semi-automatically with Praat speech processing software (Boersma & Weenink, 1996, 2015). As a first approximation the recordings for a given speaker were automatically split up into stretches of uninterrupted speech and pauses using the annotation module with automatic speech/silence detection. For male speakers the bottom pitch was set at 70 Hz, for females at 120 Hz. For all other parameters the default setting was used (both speech and silences should exceed 100 ms, silence threshold at −25 dB). The resulting annotation grids plus waveforms were inspected by ear and eye. The procedures laid down by Buhmann et al. (2002) were followed. Filled pauses, which are not detected as such by the algorithm, were set by hand, and misplaced segmentation boundaries were corrected when necessary. Each speaker produced three fragments. Time intervals preceding and following fragments were discarded. Only pauses within each of the three fragments were included in the computations. Filled pauses were separately labelled. A filled pause, by definition, is not coarticulated with whatever precedes it. As a result, a filled pause is always preceded by a short stretch of silence. It occurred regularly that a speaker fell silent for several hundreds of milliseconds, then produced an *eh* or *ehm* filled pause, which could or could not be followed fluently by the onset of the next fragment. In such cases two or even three pauses were distinguished, one of which was filled and the others were silent. As a result of this procedure the number of pauses found could be greater than the number of IPUs. In a number of cases the speaker lengthened a word-final vowel, which was clearly indicative of a hesitation. In such cases, we did not mark a pause; lengthened vowels lead to slower speaking rates. The occurrences of such lengthened vowels were also marked and counted.

The transcripts of the students' interpretations were automatically converted from the Arabic script to a Western transliteration. This

transliteration is close enough to a broad phonemic transcription of what was said to enable correct syllabification. Word boundaries were checked and corrected by hand. A list of word types was extracted from the transcripts. In each word in the list, syllable boundaries were inserted by hand. Syllable boundaries were then inserted automatically in the materials by applying a series of find-and-replace commands using the words and their hyphenation in the list of types. The number of syllables as well as the number of words was then counted automatically for each IPU and stored in the database.

For each speaker the following speech rate related measures were computed from the duration data and the syllable and word counts:

- Total articulation time: (i.e., the duration of all the IPUs added together)
- Total pause time: the duration of all the intervals, whether silent or filled, between IPUs added together
- Total filled pause time: the duration of all filled pauses (*eh*, *ehm*, *mm*, *mmm*) added together
- Number of IPUs
- Number of silent pauses
- Number of filled pauses
- Standard deviation of IPU duration
- Standard deviation of pause duration
- Speaking rate in words/s: (total articulation time + total pause time) / number of words
- Speaking rate in syll/s: (total articulation time + total pause time) / number of syllables
- Articulation rate in words/s: total articulation time / number of words
- Articulation rate in syll/s: total articulation time / number of syllables
- %pause: total pause duration / (total articulation time + total pause duration).

For more information and background on these measures we refer to Wu and Van Heuven (2017, 2021) and references provided there.

4. Results

4.1 Count measures

The number of inaccuracies and omissions were counted by comparing each individual student's written transcript with the ideal, model interpretation. Note that the model interpretation contained a rendition of all words and concepts that occurred in the English source text.

Table 2 presents the number of inaccurately translated words as well as the number of omissions, for the members of the control group and the experimental group separately. Moreover, the individuals in the two groups were matched pairwise on the basis of their performance on the overall TOEFL score obtained in the pre-test.

Table 2

Number of incorrectly translated words, omitted words and total number of word errors for individual subjects in control and experimental groups. Subjects are matched on TOEFL score in pre-test, and listed from best to poorest.

Control group				Experimental group			
Subject	Wrong word	Omission	Total	Subject	Wrong word	Omission	Total
C01	15	15	30	E01	5	13	18
C02	12	18	30	E02	6	30	36
C03	20	10	30	E03	22	8	30
C04	24	10	34	E04	15	8	23
C05	20	12	32	E05	19	8	27
C06	34	6	40	E06	22	8	30
C07	13	19	32	E07	10	17	27
C08	21	29	50	E08	26	7	33
C09	32	15	47	E09	22	10	32
C10	22	30	52	E10	25	15	40
C11	23	56	79	E11	19	12	31
C12	26	41	67	E12	18	22	40
C13	37	43	80	E13	32	32	64
C14	31	50	81	E14	23	27	50
C15	24	110	134	E15	39	40	79
Mean	23.6	30.9	54.5	Mean	20.2	17.1	37.3

Although the two groups did not differ from one another on the pre-test, there is a substantial difference in the number of word errors counted in the

transcripts of the subjects' interpreting performance in the post-test. The number of word errors is significantly smaller for the experimental group for both wrong words (24 versus 20) and for omissions (31 versus 17), $t(14) = 1.8$ ($p = .045$, one-tailed) and $t(14) = 2.6$ ($p = .016$, one-tailed), respectively. The effect is most clearly seen in the total number of word errors (55 versus 37), $t(14) = 4.0$ ($p < .001$).

The crucial question is if the experts' global judgment of the accuracy of the interpretations can be understood from the objective post-hoc error counts. To answer this question, we computed the correlation coefficients between the objective counts and the expert judgments. The correlations are shown in Table 3.

Table 3
Correlation matrix for objective error counts and global expert judgments. The lower triangle contains Pearson's R, the upper triangle shows the non-parametric Spearman's rho.

	Objective error counts			Judgments	
	wrong words	word omissions	total errors	accuracy	omissions
wrong words		.218	.732	−.696	−.718
omissions	.255		.755	−.712	−.698
total errors	.555	.946		−.921	−.904
Accuracy	−.691	−.704	−.838		.973
omissions	−.694	−.718	−.851	.976	

Note: $r > .555$: $p < .001$

It can be observed, first of all, that the global accuracy and omission judgments are very strongly correlated ($r = .976$). This means that the judges did not differentiate between these two aspects of the interpreting performance. This seems understandable, given that leaving out words or concepts that occurred in the English source text from the interpretation is a form of inaccuracy. In the objective post-hoc error counts the numbers of incorrectly translated words and omitted words are not significantly correlated, so that these two types of error might have contributed to the

global judgments separately and independently. Note that the number of omission errors was much larger than the number of inaccurately translated words, which explains the much higher correlation between the former (r = .946) and the total error score than the latter (r = .555). Observe, finally, that the non-parametric *rho* coefficients tend to be somewhat better than their parametric counterparts *r*. This suggests that the relationships between the objective error counts and the global judgments are non-linear. We will come back to this issue presently.

The most important conclusion to be drawn from Table 3 is that the global accuracy and omission judgments (which are virtually identical) can be predicted with great precision from the objective error counts, especially when the total number of errors is used as the predictor, with *rho*-values in excess of .9. Clearly, then, the experts' global judgments have a high concurrent validity in that they lead to the same ranking of students as can be achieved on the basis of laborious error counts.

To conclude the analysis here, Figure 1 plots the mean of the global accuracy and omission judgments as a function of the total error number for each of the 30 students. The x-axis of the plot, however, is not linear but logarithmic. A preliminary check revealed that the percentage of the judgment scores accounted for by a logarithmic model was appreciably better than by a linear model, with R^2 = .824 and .720, respectively.

It can be seen in Figure 1 that accuracy judgments for the experimental group are better than those for the control group. We now understand that the difference between the two groups is related in a perfectly straightforward manner to the difference in numbers of incorrectly translated words and words omitted during interpretation from English into Persian. Moreover, the relationship works the same way for both groups of student interpreters. What we do not know is how this difference in performance can be

explained. The experimental group received ample explanation of prosodic differences between English and Persian, and did practical exercises emphasizing these prosodic differences, but this in and by itself does not explain why the accuracy of the translation of the contents should improve.

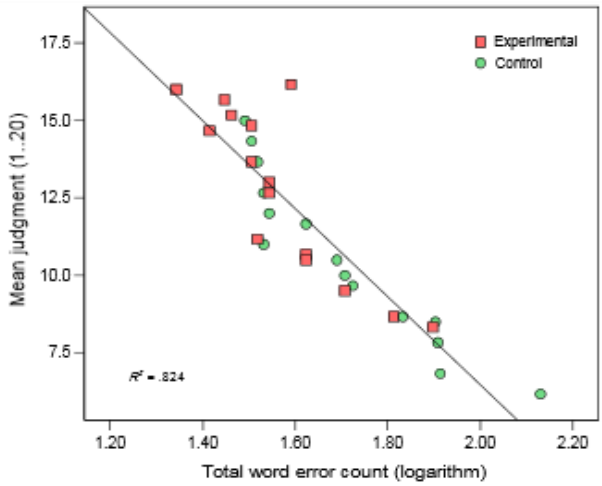

Figure 1. *Mean of global word accuracy and word omission judgments as a function of the logarithm of the total count of errors*

4.2 Acoustic measures

A total of 15 speech rate related parameters were computed. Some of these were measured from the acoustic signal, other were counted in written transcripts of the interpreting performance by the participants. Compound measures were derived by computing ratios or percentages based on raw measurements. For instance, articulation rate was defined as the Total articulation time divided by the Total number of syllables counted. Table 4 presents the summary statistics for these 15 parameters, for the experimental and control groups separately. Independent t-tests indicate that the small differences between the two groups never reach statistical significance for any of the 15 parameters, with p-values ranging between .187 and .950.

Table 4

Mean and standard deviation of 15 fluency-related acoustical correlates for control group and experimental group. The difference between the two means (Δ) and the t and p-values are given (df = 28 for each parameter).

Parameters	Control group		Exper. Group		All		Δ	*t*	*p*
	Mean	SD	Mean	SD	Mean	SD			
Total articulation time (s)	71.2	14.2	73.6	10.7	72.4	12.4	−2.4	−.5	.606
Total pause time (silent + filled)	21.9	10.5	21.7	9.0	21.8	9.6	.2	.1	.950
Total N words	220.3	33.8	231.5	17.3	225.9	27.0	−11.1	−1.1	.266
Total N syllables	444.5	72.9	470.9	36.6	457.7	58.3	−26.4	−1.3	.221
Percent pause (silent + filled)	22.9	6.0	22.2	5.9	22.6	5.9	.7	.3	.742
Speech rate (words/s)	2.4	.5	2.5	.4	2.5	.4	−.1	−.3	.756
Speech rate (syllables/s)	4.9	1.0	5.1	.8	5.0	.9	−.2	−.4	.657
Articulation rate (words/s)	3.1	.5	3.2	.4	3.2	.4	.0	−.3	.781
Articulation rate (syllables/s)	6.3	1.0	6.5	.8	6.4	.9	−.2	−.5	.640
SD IPU duration (s)	1.2	.3	1.6	1.3	1.4	1.0	−.5	−1.4	.187
SD pause duration (s)	.9	.4	1.0	.7	.9	.5	−.1	−.5	.644
SD N words in IPU	4.0	1.1	5.0	3.6	4.5	2.7	−1.0	−1.0	.311
SD N syllables in IPU	8.0	2.5	10.3	7.7	9.1	5.8	−2.4	−1.1	.270
N IPUs	33.9	11.4	33.1	11.4	33.5	11.2	.7	.2	.862
N pauses (silent + filled)	34.1	15.9	31.7	15.8	32.9	15.6	2.3	.4	.690

Figure 2 presents the relationship between percentage of pause and the judged pace of delivery, shown separately for the experimental and control groups.

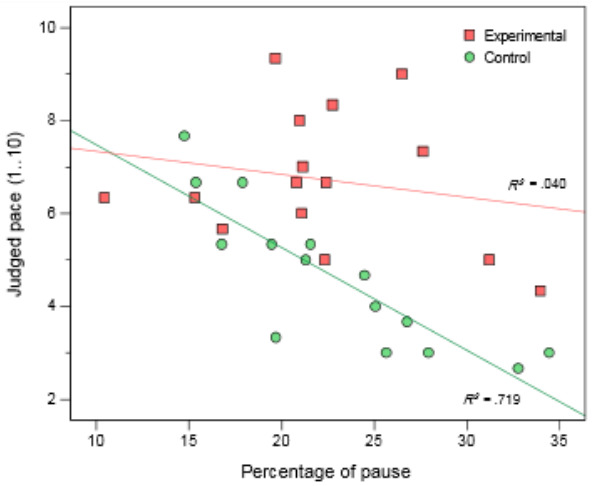

Figure 2*. Judged pace as a function of percentage of pause (silent and filled) in spoken text, shown separately for members of experimental and control groups*

The figure shows that the judged pace is better for the experimental group than for the control group. However, the figure also shows, quite clearly, that the relationship between percentage of pause and the judged pace of delivery is strong and linear as far as the control group is concerned. The greater the percentage of pausing, the poorer the judged pace, where the objective measure accounts for 72% of the variance the judged pace score, R^2 = .719. The relationship is much weaker, in fact almost absent, for the experimental group. It is not the case that the experimental group has no variability in percentage of pause: the distribution of this objective parameter is roughly the same for experimental and control group alike, with a spread between 10 and 35%. In order to shed light on this curious asymmetry, let us now examine the relationship between articulation rates (words/s) and judged

pace. The expectation is that a faster articulation rate should correlate with better pace judgments. The results are shown in Figure 3.

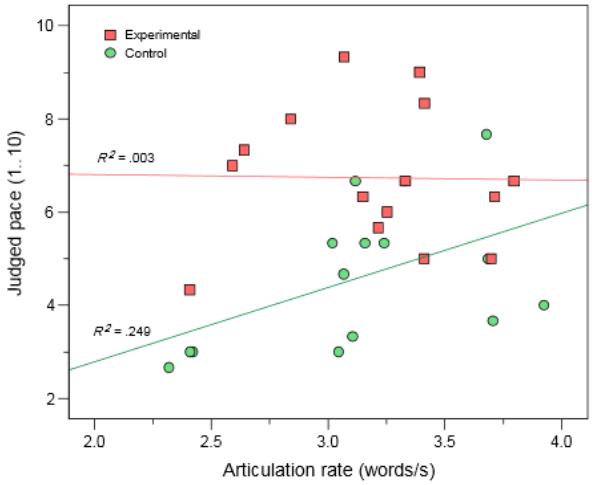

Figure 3*. Judged pace as a function of percentage of articulation rate (words/s) in spoken text, separately for members of experimental and control groups*

Again, it can be observed that the relationship between the objective measure and judged pace is fairly strong for the control group, R^2 = .249, and explains a quarter of the variance in the judgments. It can also be noticed that there was no correlation at all for the experimental group.

In order to understand the asymmetry in the results of the experimental and control groups, at first it is needed to examine the relationship between the predictor variables used here, viz. percentage pause and articulation rate (in words/s and in syllables/s). It turns out that it is rather immaterial whether articulation rate is expressed in words/s or in syllables/s. The intercorrelation between these two measures is almost perfect at r = .991 for the control group and r = .994 for the experimental group (N = 15, p < .001 in both cases). The intercorrelation between articulation rate and %-pause shows the same remarkable discrepancy between the two groups we met before, such

that the correlation is relatively strong and significant for the control group, $r = -.564$ ($p = .014$) and $-.618$ ($p = .007$) for words/s and sylls/s, respectively, but weaker and insignificant for the experimental group, $r = -.340$ ($p = .107$, one-tailed) and $-.367$ ($p = .089$, one-tailed) for words and syllables per second, respectively (see also Figure 3).

These results suggest that articulation rate and %-pause in the control group are both indices of cognitive difficulty in task performance: when these participants find it difficult to interpret the incoming message, they tend to speak more slowly, leading at the same time to fewer syllables (or words) per second and to more and/or longer pauses. These would be pauses for the sake of the speaker rather than for the sake of the listener. The speaker needs more time to find appropriate words and formulations to get the message across. The speaker does not insert pauses to help the listener by clearly marking off processing units (be they clauses or constituents). A reasonable prediction here would be that these speakers also produce relatively many filled pauses, which are the hallmark of problems with finding words or formulations. In contradistinction to this we would expect pauses in the experimental group to be used as structure markers for the sake of the listener. These would be relatively short pauses, which are planned by the speaker to guide the listener. Additionally, fewer filled pauses and other overt markers of planning difficulty should be observed for the experimental group.

These hypotheses can be tested by examining the number of disfluency markers, which is what it is done in the following section.

4.3 Disfluencies

In order to understand the discrepancy between the experimental and control group, let us now consider the number of disfluencies marked for each. We distinguish the following four categories:

1. Long silent pause, indicative of extra planning time needed. Assuming that pauses between IPUs in fluent speech normally do not exceed a duration of 1000 ms, any silent pause longer than 1000 ms was considered a (potential) disfluency.
2. Filled pause, indicative of hesitation, i.e., any instance of *eh*, *ehm* or *mm* that is not fluently coarticulated with whatever precedes it.
3. Lengthened vowel (i.e., a word-final vowel that is lengthened and is indicative of hesitation).
4. Repetition (i.e., the repetition of something that was said in the immediately preceding IPU) then broken off, and repeated in a second attempt. In a number of cases there was no break (no silent or filled pause after the false start); the repetition followed seamlessly after the false start. We decided to count the repetitions only (and only if the repetition was not an instance of stuttering – which happened on two occasions).

Table 5 lists the disfluencies found, for the experimental and control groups, together with the number of regular IPUs and short silent pauses. The latter two categories are indices of fluent speech, whereas the other four categories point to planning difficulties on the part of the speaker.

Table 5

Mean Duration (in seconds) and Number of IPUs, Regular Silent Pauses, Long Pauses and Filled Pauses Produced by Experimental and Control Groups

Disfluency	Control group		Experimental group		Δ (exp − cont)	
	Duration	N	Duration	N	Duration	N
Regular IPU	2.149	456	2.230	485	0.081	29
Silent pause short	.412	394	.438	392	0.026	−2
Silent pause long	3.220	43	3.219	43	−0.001	0
Filled pause	.363	73	.380	42	0.017	−31
Lengthened vowel	1.442	30		0		−30
Repetition	2.019	23	1.979	11	−0.041	−12

Table 5 shows no systematic differences in the behavior of the experimental and control participants in terms of the duration and number of disfluencies, with three notable exceptions. The number and mean duration of regular IPUs, as well as those of both short and long silent pauses (the latter would be indicative of planning problems on the part of the speaker) are virtually identical between the two groups. This also goes for the duration of the remaining categories of disfluencies but, remarkably, the number of disfluencies in the latter three categories differs between the groups such that the control group shows many more disfluencies in the categories filled pause, excessive prepausal vowel length and repetitions after a false start. These three categories, obviously, are indicative of planning problems. Before drawing any conclusions from these observations let us first see how the numbers are distributed over the 15 participants in each group.

Table 6 (next page) presents the numbers of disfluencies in the categories filled pause, excessive prepausal vowel lengthening and IPUs that repeat materials after a false start, broken down by the two groups of participants. In order to make the comparison maximally sensitive, the participants in the two groups have been matched on their TOEFL test scores.

Inspection of Table 6 reveals, first of all, that the TOEFL pre-test predicts the number of disfluencies observed in the interpretation tasks rather well. The correlations are negative since high TOEFL scores (indicating good proficiency in English) lead to better performance, with fewer hesitations in the interpreting task. The best fit was obtained when the TOEFL scores were used to predict the logarithm of the number of disfluencies. Quite a few participants fulfilled their interpreting task without any disfluency. Since the logarithm of 0 is undefined, we remedied this by incrementing the overall disfluency count for each of the 30 participants by 1. We then find the same asymmetry in the predictability that we met before. The interpreting performance of the control group can be predicted from objective measures

much better than the scores of the experimental group. The correlation coefficients are $r = -0.742$ ($N = 15$, $p = .001$, one-tailed) for the control group and $r = -0.440$ ($N = 15$, $p = 0.050$, one-tailed). Across all participants $r = -0.612$ ($N = 30$, $p < .001$, one-tailed).

Table 6

Number of over disfluencies in three categories (excessive pre-pausal vowel lengthening (L), filled pause (FP), repetition of words after a false start (R)) for participants in control and experimental groups. Participants are rank ordered within their group on the basis of their pre-test overall TOEFL score. Gender is indicated

Control Group							Experimental group						
Student	Gender	L	FP	R	Total	TOEFL	Student	Gender	L	FP	R	Total	TOEFL
C01	F	0	1	0	1	610.00	E01	M	0	0	1	1	613.33
C02	F	0	0	0	0	586.67	E02	F	0	1	1	2	603.33
C03	M	0	0	1	1	563.33	E03	M	0	0	0	0	566.67
C04	M	0	0	2	2	553.33	E04	M	0	0	0	0	563.33
C05	F	0	0	2	2	540.00	E05	F	0	0	1	1	553.33
C06	F	0	0	0	0	530.00	E06	F	0	4	1	5	553.33
C07	F	0	0	0	0	513.33	E07	M	0	3	0	3	550.00
C08	F	0	0	0	0	510.00	E08	F	0	0	0	0	550.00
C09	M	0	1	9	10	506.67	E09	F	0	2	0	2	546.67
C10	M	0	0	1	1	503.33	E10	M	0	1	0	1	523.33
C11	M	0	4	0	4	500.00	E11	F	0	0	0	0	516.67
C12	F	0	10	2	12	490.00	E12	F	0	0	0	0	493.33
C13	M	0	0	5	5	473.33	E13	M	0	0	4	4	480.00
C14	M	20	35	1	56	446.67	E14	M	0	0	2	2	476.67
C15	F	10	22	0	32	446.67	E15	F	0	31	1	32	446.67
Total		30	73	23	126		Total		0	42	11	53	

Although the total number of overt disfluencies in the performance of the control group (126) is more than twice as large as for the experimental group (53), the difference falls short of significance. A sign test on the counts (11 pairs matched on within-group TOEFL rank, excluding 4 tied scores) yields $p = .114$ (one-tailed), which is a (weak) trend at best.

We may also normalize the number of overt disfluencies by speech time. After all, when a speaker produces more speech materials (words, syllables) during a longer stretch of time, there is more opportunity to produce errors

and disfluencies. We therefore divided the total number of disfluencies per speaker by the duration of all his/her IPUs added together.

To conclude the analysis, we now try to establish a possible relationship between the incidence of overt disfluency markers and the pace of the interpreting performance as judged by the expert raters. The correlation between the raw number of disfluencies and judged pace is slightly poorer than when the logarithm of the number of disfluencies used, but even then r is rather weak at −.526 (N = 30, p < .001). Moreover, similar correlation coefficients are obtained between the disfluency counts and all other judged aspects of the interpreting performance (which tend to be strongly correlated, see Table 3). When we compute the correlations separately for experimental and control groups, we observe the same asymmetry as before: correlations are appreciably better for the control group than for the experimental group, not just for pace but for all judged aspects.

4.4 Predicting pace from multiple correlates of fluency

In the preceding sections we have seen that the prosodic parameter with the most tangible measureable correlates, (i.e., pace of delivery, correlates with a large number of variables). These variables can be located in the acoustical domain, (e.g., articulation rate (syllables per second) and percent pause). However, pace also correlated with the number of disfluencies per unit time as counted in the transcripts of the interpreting performances obtained from the participants. Interestingly, the intercorrelations between the disfluency counts and the acoustic correlates of pace were relatively modest, so that there is reason to try to predict judged pace from acoustic and count parameters together. Table 7 presents the correlation matrix for judged pace (dependent) and the acoustic and count parameters of (dis)fluency. Only the non-redundant lower triangle of the matrix is shown.

This table summarizes the information presented. We now see that the variability in the duration of the inter-pausal units (or fluent runs) is, in fact, fairly good predictor of judged pace, better, for instance, than articulation rate

or percent pause, though still weaker than the relatable number of disfluencies. This is somewhat unexpected, especially since the correlation is positive. One would expect competent speakers to divide their delivery into chunks of roughly equal size, which should yield a negative correlation with judged pace: the smaller the variability in the chunk size, the better the fluency. Variability in the pause duration, however, does not correlate with judged pace.

Table 7

Correlation matrix of judged pace (dependent) and five predictors: Percent pause, articulation rate (syllables/second), standard deviation of inter-pausal units (ms), standard deviation of (filled and silent) pauses (ms) and the relative number of disfluencies per unit time. N = 30 for each cell.

	Pace	%pause	Art rate	Sd sp	SD pause
Percent pause	−.469**				
Artic. rate (syll/s)	.314*	−.504**			
SD speech	.503**	−.181	.121		
SD pause	.075	.503**	.112	.561**	
Rel. disfluencies	-.543**	.646**	−.623**	−.267	−.035

* $r > .300$: $p < .05$, ** $r > .460$: $p < .01$ (one-tailed).

Table 8A-B contains the same correlation matrix as in Table 7 but now the data are presented separately for the experimental and control groups.

Table 8A-B.

Correlation matrix of judged pace (dependent) and five predictors. for more information, see Table 7, N = 15 per cell.

		Pace	%pause	Art rate	SD sp	SD pause
A.Control	Percent pause	−.848**				
	Artic. rate (syll/s)	.592**	−.618**			
	SD speech	.741**	−.773**	.337		
	SD pause	−.454*	.593**	.073	−.397	
	Rel. disfluencies	−.583*	.727**	−.727**	−.615**	−.007
B. Experimental	Percent pause	−.200				
	Artic. rate (syll/s)	−.009	−.367			
	SD speech	.487*	−.056	.064		
	SD pause	.338	.487*	.137	.732**	
	Rel. disfluencies	−.470*	.601**	−.434	-.210	−.034

* $r > .450$: $p < .05$, ** $r > .590$: $p < .01$ (one-tailed).

Breaking the correlations down separately for the experimental and control groups shows the by now familiar result that the correlations are clearly stronger for the control group than for the experimental group. There is, however, one parameter that behaves differently between the two groups. The variability in duration of the (filled and silent) pauses correlates negatively with judged pace in the control group ($r = -.454$, $p = .045$, one-tailed) but positively in the experimental group ($r = -.338$, $p = .109$, ins.). Variability in pause duration in the control group is typically caused by long silences and hesitations when the student interpreter is stuck for words. The better participants in this group have fewer of these long pauses, so that the variability in their pause durations is reduced. The experimental group, however, has fewer long pauses and disfluencies as a general characteristic; their pause variability is probably conditioned by the grammatical structure of their utterances such that light prosodic boundaries (at the phrase and clause level) have relatively short pauses and deeper boundaries (at the sentence level) are marked by longer pauses – as is typically found in other languages such as English (e.g. Grosjean, Grosjean & Lane 1979, Selkirk, 1984) and Dutch (e.g., Swerts, 1997). In that case, the more variable the pause duration, the more competently does the speaker use prosodic markers. Note also that for the experimental group longer pauses tend to go together with longer IPUs, whereas the correlation is reversed for the control group.

Multiple regression analyses were then conducted for the two groups combined (N =30) and for the experimental and control groups separately. All five predictors mentioned in the correlation matrix were entered simultaneously in one analysis and in step-wise mode in another. The results are shown in Table 9 A-B-C (next page).

For the total group of participants combined we find an R^2 of .473 when all five predictors are entered simultaneously. In the stepwise mode it turns out that only two predictors make a sufficient contribution to be included in an

optimal model, which then accounts for 43.3 percent of the variance (i.e., only 4 points less than the saturated model).

When the analysis is performed for the experimental group separately, no model is produced that is better than chance. As can be seen in Table 8B, two predictors correlate significantly (but only just) with the criterion when studied as single predictors, viz. SD speech and Relative number of disfluencies but they lose significance in the simultaneous entry of five predictors because of the increased degrees of freedom.

Judged pace is best predicted for the control group. Entering all five predictors simultaneously yields an R^2 of .785 (i.e., the model accounts for 79 percent of the variance). However, as was shown earlier, one single predictor, (i.e., percent pause duration) accounts for 72 percent of the variance; none of the remaining four predictors makes a further contribution that reached significance.

Table 9

Summary of multiple regression analysis with judged pace as the dependent from five predictors: Percent pause, Articulation rate (syllables/second), Standard deviation of interpausal units (ms), Standard deviation of (filled and silent) pauses (ms) and the Relative number of disfluencies per unit time. Analysis was run with simultaneous entry (left part of table) and in stepwise mode (right part of table) for all participants combined (panel A, N = 30) as well as for the control (panel B, N = 15) and experimental groups (panel C, N = 15) separately.

	Simultaneous entry					Stepwise mode				
Predictors	Beta	R^2	F	$df_{1,2}$	p	Beta	R^2	ΔR^2	F	$df_{1,2}$	p
				A. Combined groups							
Rel. disfl.	−.443					−.440	.295	.295	11.7	1, 28	
SD speech	.617					.385	.433	.138	6. 6	1, 27	.016
SD pause	−.380										
%Pause	.166										
Art. rate	.089	.473	4.3	5,24	.006						
				B. Experimental group							
SD pause	.642										
%Pause	−.444										
Artic. rate	−.419										
Rel. disfl.	−.376										
SD speech	−.059	.457	1.5	5, 9	.227						
				C. Control group							
% Pause	−.751					−.221	.719	.719	33.3	1, 13	<.001
Rel. disfl.	.376										
SD speech	.342										
Artic. rate	.277										
SD pause	.110	.785	6.6	5, 9	.008						

5. Discussion

We examined the relationships between the expert judgments of the quality of the participants' interpreting performance on the one hand and objective correlates of their performance on the other. In the quality judgments a rating instrument was used that was comprised of ten scales. Seven of these pertain to aspects of quality that can be (and actually were) established by examining written transcripts of the interpreting tasks. These aspects relate to abstract linguistic properties of the interpretations, such as the accuracy with which the ideas in the source text were expressed, number of words omitted, appropriateness of choice of words and terminology, number of grammatical errors, and overall coherence of the interpretation. The remaining three scales were meant to capture the phonetic aspects of the delivery of the interpretation, (i.e., the degree of accentedness, the pace (or fluency) of the delivery and the pleasantness of the voice). These three phonetic aspects all relate to relatively long-term aspects of speech, (i.e., are not properties of specific vowels or consonants, and are therefore prosodic features).

It was reported before (Yenkimaleki & Van Heuven, 2018, 2021, 2022) that the seven textual/linguistic scales intercorrelate very strongly, as do the three prosodic scales, but the correlations between scales in different categories are relatively low. The possibility to divide the ten scales into one group of seven non-prosodic and three prosodic scales was borne out by a factor analysis, which showed opposite factor loadings by the two groups of scales on the two principal components extracted in the analysis.

The results presented here bear out, quite clearly, that the expert judgments of the non-phonetic aspects can be related in a rather straightforward manner to a number of structural properties that could be quantified or counted in written transcripts of the interpreters' deliveries.

Since the seven rating scales are very strongly intercorrelated there is little point in trying to predict each of these scales separately from objective counts. It would be sufficient, therefore, to summarize the most striking correlations found.

It turned out, then, the total number of errors in the interpreted passages (i.e., wrong words and number of omitted words added together) afford excellent prediction of the accuracy (and omissions) rating, with correlation coefficients in excess of .900. The actual numbers of wrong words and omissions were quite disparate, however. The conclusion follows, therefore, that the expert judges were not able to differentiate between these two aspects of accuracy even though they were clearly different in the interpreters' productions. This conclusion does not reflect negatively on the quality of the raters – it just shows that these two closely related aspects are extremely difficult to distinguish when asked to give an on-the-spot evaluation of an interpreter's performance. Proper differentiation between the two types of inaccuracy in interpreting can only be achieved when a written transcript is available for a detailed and more time-consuming analysis.

These lexical accuracy parameters (words incorrectly translated or omitted altogether) are the two most important aspects of the overall rating of the students' interpreting performance. Incorrect words were weighted by 20, omissions by 15, so that together they represent 35 percent of the overall score. The other eight aspects together, with weights of either 7 or 10, represent the remaining 65 percent.

It should be noted in this context that the objective measures that predict the judged accuracy of the interpretation performance so well, are also the quantitative measures that optimally differentiate between the participants in the experimental and the control groups. The experimental group produced a very significantly smaller number of (lexical) inaccuracies than the control

group (with a mean of 55 versus 37 lexical inaccuracies per speaker). It remains unclear at this stage why the experimental group would produce fewer inaccuracies than the control group. Why would a 36-hour training module emphasizing prosody and prosodic differences between English and Persian, which is what differentiates the experimental groups from the control group, lead to a reduction in number of lexical errors?

The total number of disfluencies counted in a participant's delivery proved to be a reasonable predictor for the rated adequacy of the speaker's expression and coherence, explaining between a quarter and a third of the variance in the ratings. Interestingly, the ratings could be better predicted by a relative than an absolute count of the number of disfluencies. In the relative measure the number of disfluencies were related to the duration of the interpretation. So the expert judges did not just keep track of the number of disfluencies they heard in the interpreter's delivery but normalized for the length of the delivery.

There is no point in trying to predict the ratings of grammatical correctness of the interpretations. Since the interpretation was from English into Persian, all participants spoke the target language as their native language. Although numerous disfluencies were found in the Persian utterances produced, no ungrammatical structures were observed.

Turning now to the prosodic rating scales, it appeared that the pace of the delivery is clearly related to a number of objective parameters. The three phonetic-prosodic evaluation scales are very highly intercorrelated, even if the correlation coefficients are computed for the experimental and control groups combined ($.888 < r < .976$). We decided to concentrate on the prediction of pace (fluency) as this parameter has rather straightforward acoustical correlates. The results show that the pace rating for the control group can be predicted most successfully by a single parameter (i.e., percent

pause duration) which by itself explains 72 percent of the variance in the pace rating. Curiously enough, no predictive model is possible for the experimental group and only two single predictors yield marginally significant correlations with pace, (i.e., the variability in the duration of the interpausal units and the relative number of disfluencies).

In the overall prediction of pace for the group of 30 participants combined a regression model was found that explains 43 percent of the variance. The best predictor here was the number of disfluencies (normalized for the total duration of the interpretation), followed by the variability in the duration of (filled and silent pauses).

It remains unclear at this time why the pace (or fluency) judgments can be predicted in a rather straightforward fashion from a number of objective properties of the speech produced by the student interpreters in the control group, whereas no convincing relationships could be found between the acoustic measurements and counts of errors and disfluencies for the experimental group. Part of the solution of this problem may be that the assumption underlying the analysis we applied is that the relationships between the predictors and the criterion should be linear. Already we briefly speculated that it might be more reasonable to assume a U-shaped (i.e., quadratic or parabolic) relationship between such parameters as speech rate and percent pause on the one hand and judged pace on the other. Obviously, when there is excessive pausing or an exceedingly slow speaking rate, which would cause poor judgments of pace (or fluency). However, a speaker may also speak so fast and with so few pauses that the listener suffers from information overload – which would yield unfavorable ratings of pace. We argued that speech rates and speech pause ratio (i.e., percent pause duration) should ideally be somewhere in the middle of the range, neither too slow nor too fast.

No signs of non-linearity can be observed in the results obtained for the control group. For this group the overall tendency is: the faster the better. However, when we examine the results of the experimental group more closely, we may observe a tendency in the scatterplots (Figures 2-3) to reveal non-linear, in fact, parabolic relationships between the acoustic predictors and judged pace. Table 10 lists side-by-side the correlation coefficients between the acoustic predictors and judged pace obtained for linear and quadratic (U-shaped) regression functions for the experimental and the control group separately.

Table 10
Correlation coefficients (Pearson's r) between acoustic predictors and judged pace for experimental and control groups, assuming linear versus quadratic relationships.

Acoustic predictor	Experimental group			Control group		
	Linear	Quadratic	Δ	Linear	Quadratic	Δ
% Pause	.200	.531	.331	.848	.887	.039
Articulation rate	.001	.430	.429	.592	.642	.050

Table 10 shows that the U-shaped function fits the data much better (by 33 to 43 points) than a linear function. For the control group, however, the difference between linear and quadratic functions is almost negligible (5 points or less). We are inclined to interpret this difference as an indication that some speakers in the experimental group speak so fast and pause so little that the raters judge this speed of delivery (or pace) uncomfortable.

6. Conclusion

Overall, we addressed the usability of objective assessment procedure for the quality of interpretation performance. The results of the study showed that interpreter trainees' performance can be systematically assessed by relating the intersubjective expert judgments to objective measures. It is suggested that the expert judgment in the evaluation of interpretation performance can be predicted by combination of objective correlates as well.

We should point out that the present study was undertaken on a relatively small scale, with two groups of fifteen interpreting trainees in the beginning stages of their professional education. Moreover, the source and target languages were Persian and English, so that we can only speculate as to the generalisability of the results to other combinations of source and target languages, where the typological differences between the languages are either greater or smaller than in the case of English and Persian– which are distantly related members of the Indo-European language family.

The pedagogical implications of this study would pertain to interpreting programs. The policy makers, curriculum developers, practitioners and administrators need to make a number of changes in their overall approach in evaluation of interpreter trainer performance at different training programs.

Acknowledgement

We would like who thank all the students and colleagues who collaborated with us in this study.

References

Bialystok, E. (1978). A theoretical model of second language learning. *Language Learning*, *28*(1), 69–83. https://doi.org/10.1111/j.1467-1770.1978.tb00305.x

Boersma, P. & Weenink, D. J. M. (1996). Praat, a system for doing phonetics by computer, version 3.4 *Report* 132, Institute of Phonetic Sciences University of Amsterdam. www.praat.org

Buhmann, J., Caspers, J., Heuven, V. J. van, Hoekstra, H., Martens, J.-P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus. *Proceedings of LREC 2002*. ELRA, 779−785. http://lrec-conf.org/proceedings/lrec2002/pdf/96.pdf

Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: Trends and perspectives* (pp. 205–224). Multilingual Matters.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Dawson, M., & Schell, A. (1987). Human autonomic and skeletal classical conditioning: the role of conscious cognitive factors. In G. Davey (Ed.), *Cognitive processes and Pavlovian conditioning in humans* (pp. 27–55). John Wiley & Sons.

Derwing, T. M., Munro, M. J., & Wiebe, G. E. (1998). Evidence in favour of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Etemadi, S. H., & Abbasian, G. R. (2023). Dynamic assessment and EFL learners' writing journey: Focus on DA modalities and writing revision types. *Teaching English Language*, 17(1), 53–79. https://doi.org/10.22132/TEL.2022.162923

Garzone, G. (2002). Quality and norms in interpretation. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century* (pp. 107–119). Amsterdam/ John Benjamins. https://doi.org/10.1075/btl.43.11gar

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: performance structures in sentence production. *Cognitive Psychology*, 11, 58−81. https://doi.org/10.1016/0010-0285(79)90004-5

Gile, D. (1995). Fidelity assessment in consecutive interpretation: An experiment. *Target*, *7*(1), 151–164. https://doi.org/10.1075/target.7.1.12gil

Gile, D. (2005). Empirical research into the role of knowledge in interpreting: Methodological aspects. In H. V. Dam, J. Engberg & H. Gerzymisch-Arbogast (Eds.), *Knowledge systems in translation* (pp. 149–171). Mouton de Gruyter.

Heuven, V. J. van (1994). Introducing prosodic phonetics. In C. Odé & V. J. van Heuven (Eds.), *Experimental studies of Indonesian prosody*. Semaian 9. Leiden: Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië, Leiden University. https://hdl.handle.net/1887/2574

Heuven, V. J. van (2008). Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing*, 2, 39–62. https://doi.org/10.3366/E1753854809000305

Heuven, V. J. van (2017). Prosody and sentence type in Dutch. *Nederlandse Taalkunde*, 22, 3–29, 43–46. http://doi: 10.5117/NEDTAA2017.1.HEUV

Heuven, V. J. van (2018). Notes on the phonetics of word and sentence stress: A cross-linguistic (re-)view. In H. van der Hulst, J. Heinz & R. Goedemans (Eds.), *The study of word stress and accent: Theories, methods and data* (pp. 13–59). Cambridge University Press. https://doi.org/ 10.1017/9781316683101.002

Heuven, V. J. van (2022). Resolving the prosody paradox. In P. Arantes & A. Post da Silveira (Eds.), *Prosody and Bilingualism* (pp. 168–204). Araraquara: Letraria. https://www.letraria.net/prosodia-e-bilinguismo/

Heuven, V. J. van, & Sluijter, A. M. C. (1996). Notes on the phonetics of word prosody. In R. Goedemans, H. van der Hulst & E. Visch (Eds.), *Stress patterns of the world, Part 1: Background* (pp. 233–269), HIL Publications (volume 2), The Hague: Holland Academic Graphics.

Jones, R. (2014). *Conference interpreting explained*. Routledge.

Jong, N. de, & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533−568. https://doi.org/https://doi.org/10.1111/j.1467-9922.2010.00620.x

Khodashenas, M. R., Khodabakhshzadeh, H., Baghaei, P. & Motallebzadeh, K. (2023). Language assessment literacy components; now and then: A case of Iranian EFL head teachers. *Teaching English Language*, 17(1), 107−138. https://doi.org/10.22132/TEL.2022.163654

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41(3−4), 295−321. https://doi.org/10.1177/002383099804100404

Kopczynski, A. (1994). Quality in conference interpreting: Some pragmatic problems. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research into simultaneous interpretation* (pp. 87–99). John Benjamins. https://doi.org/10.1075/btl.2.24kop

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13–104). Macmillan.

Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 640–673). Basil Blackwell.

O'Neal, G. (2010). The effects of the presence and absence of suprasegmental on the intelligibility and assessment of an expanding-circle English according to other expanding-circle English listeners. *JAIRO* (Japanese Institutional Repositories Online).

Pöchhacker, F. (2004). *Introducing Interpreting Studies*. Routledge.

Rutherford, W., & M. Sharwood Smith (1985). Consciousness-raising and universal grammar. *Applied Linguistics*, *6*(3), 274–282. https://doi.org/10.1093/applin/6.3.274

Sawyer, D. B. (2004). *Fundamental Aspects of Interpreter Education: Curriculum and Assessment*. John Benjamins.

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (Eds.), *Proceedings of CLaSIC 2010, Singapore, December 2-4* (pp. 721–737). Singapore: National University of Singapore, Centre for Language Studies. https://doi.org/10.1515/9781614510932.27

Selkirk, E. O. (1984). *Phonology and syntax. The relation between sound and structure*. MIT Press.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1), 514−521. https://doi.org/10.1121/1.418114

Whalley, K., & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading*, *29*(3), 288303. https://doi.org/10.1111/j.1467-9817.2006.00309.x

Wu, S. C. (2010). Assessing simultaneous interpreting. PhD Thesis, School of Modern Languages, Newcastle University. https://theses.ncl.ac.uk/jspui/bitstream/10443/1122/1/Wu%2011.pdf

Yenkimaleki, M. (2018). Implicit vs. explicit prosody teaching in developing listening comprehension skills by interpreter trainees: an experimental study. *International Journal of English Language and Linguistics Research*, 6, 11–21.

Yenkimaleki, M., & Heuven, V. J. van (2016). Effect of prosody awareness training on the performance of consecutive interpretation from Farsi into English: An experimental study. *Asia Pacific Translation and Intercultural Studies*, *3*(3), 235–251. https://doi.org/ 10.1080/23306343.2016.1233930.

Yenkimaleki, M., & Heuven, V. J. van (2017). The effect of memory training on consecutive interpreting performance by interpreter trainees: An experimental study. *FORUM: International Journal of Interpretation and Translation*, 157–172. DOI: 10.1075/forum.15.1.09yen.

Yenkimaleki, M., & Heuven, V. J. van (2018). The effect of teaching prosody teaching on interpreting performance: An experimental study of consecutive interpreting from English into Farsi. *Perspectives: Studies in Translatology*, 26, 84–99. https://doi.org/0.1080/0907676X.2017.1315824.

Yenkimaleki, M., & Heuven, V. J. van (2019a). Effects of prosody awareness training on the intelligibility of Iranian interpreter trainees in English. *Dutch Journal of Applied Linguistics*, *8*(2), 291–309. https://doi.org/10.1075/dujal.17023.yen

Yenkimaleki, M., & Heuven, V. J. van (2019b). The relative contribution of computer assisted prosody training vs. instructor-based prosody teaching in developing speaking skills by interpreter trainees: an experimental study. *Speech Communication*, 107, 48–57. https://doi.org/10.1016/j.specom.2019.01.006

Yenkimaleki, M., & Heuven, V. J. van (2019c). Prosody instruction for interpreter trainees: does methodology make a difference? An experimental study. *Across languages and cultures: A Multidisciplinary Journal for Translation and Interpreting Studies*, 20(1), 117–133. https://doi.org/10.1556/084.2019.20.1.6

Yenkimaleki, M. & Heuven, V. J. van (2020). Relative contribution of explicit teaching of segmentals vs. prosody to the quality of consecutive interpreting by Farsi-to-English interpreting trainees. *Interactive Learning Enviornments*, xx, 1–12. https://doi.org/10.1080/10494820.2020.1789673

Yenkimaleki, M., Heuven, V. J. van, & Moradimokhes, H. (2021). The effect of prosody instruction in developing listening comprehension skills by interpreter trainees: does methodology matter? *Computer Assisted Language Learning*, xx, 1-37. https://doi.org/10.1080/09588221.2021.1957942

Yenkimaleki, M., & Heuven, V. J. van (2021). Effects of attention to segmental *vs.* suprasegmental features on the speech intelligibility and comprehensibility of the EFL learners targeting the perception or production-focused practice. *System*, 100, 1–12. https://doi.org/10.1016/j.system.2021.102557

Yenkimaleki, M., & Heuven, V. J. van (2022). Comparing the nativeness vs. intelligibility approach in prosody instruction for developing speaking skills by interpreter trainees: An experimental study. *Speech Communication*, 137, 92–102. https://doi.org/ 10.1016/j.specom.2022.01.007

Yenkimaleki M., Heuven V. J. van, & Afshar H. S. (2022), The efficacy of segmental/suprasegmental vs. holistic pronunciation instruction on the development of listening comprehension skills by EFL learners. *The Language Learning Journal*, xx, 1–17. https://doi.org/ 10.1080/09571736.2022.2073382

Yu, W., & Heuven, V. J. van (2017). Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogy. *Interpreting: International Journal of Research and Practice in Interpreting*, *19*(1), 47–68. https://doi.org/10.1075.intp.19.1.03yu

Yu, W., & Heuven, V. J. van (2021). Quantitative correlates as predictors of judged fluency in consecutive interpreting: Implications for automatic assessment and pedagogy. In J. Chen & C. Han (Eds.), *Testing and assessment of interpreting* (pp. 117–142). Springer Nature. https://doi.org/10.1007/978-981-15-8554-8_6

Zhang, X. (2016). Semi-automatic simultaneous interpreting quality evaluation. *International Journal on Natural Language Computing*, 5, 1–12. https://doi.org/10.48550/arXiv.1611.04052

2023 by the authors. Licensee Journal of Teaching English Language (TEL). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0 license). (http://creativecommons.org/licenses/by-nc/4.0).