The Effect of Self-repair on Judged Quality of Consecutive Interpreting: Attending to Content, Form and Delivery

Weiwei Zhang[1]

Dalian Maritime University; Macquarie University

Zhongwei Song

Macquarie University

Abstract

This paper investigates the correlations between self-repair and subjective assessments of student interpreters' performance in consecutive interpreting (CI). Twelve interpretations from an interpreting contest in China are transcribed, with the self-repairs identified and annotated based on Levelt's classification (1983), including both overt and covert repairs. In addition to the final scores awarded at the contest, different methods and raters are used to assess the comprising aspects of an overall quality, namely content, form and delivery. Statistical analysis shows that: (1) overt repairs have a strong positive correlation with content, and moderate negative correlations with form and delivery; (2) form and delivery are negatively correlated with covert repairs, in terms of the frequencies of repetitions and pauses, and the mean length of pauses; (3) the judges' overall assessments are more closely correlated with content than self-repairs. Finally, pedagogical implications for CI training are discussed, as are suggestions for future research.

Keywords: self-repair, interpreting quality, consecutive interpreting, subjective assessment, student interpreter

[1] *Correspondence to: weiwei.zhang@mq.edu.au*

The Effect of Self-repair on Judged Quality of Consecutive Interpreting: Attending to Content, Form and Delivery

Quality of interpreting is often defined as 'elusive' (Pöchhacker, 2012, p. 305) not only because the different parties involved in the interpreting interaction may have their own views and perceptions, but also because "the variability of perception extends… to the different criteria used in measuring it" (Pradas Marcias, 2006, pp. 25-26). Comprehensive surveys have been conducted to find out what professional interpreters perceive to be important criteria of quality assessment (Bühler, 1986; Chiaro & Nocella, 2004; Pöchhacker, 2012). For example, Pöchhacker's (2012) study showed that following the content-related criteria of *sense consistency* and *logical cohesion*, the prosody-related factor of *fluency* ranked as number three among the eleven linguistic-semantic and extra-linguistic factors carefully selected in his research. The criteria that are consistently used as variables of fluency include 'repairs', 'pauses', 'self-correction', 'fillers', 'hesitation', and 'repetition', as can be seen from Lee's (2015, pp. 231-236) summary of 24 previous studies on quality assessment methods. In Levelt's (1983) categorization, these factors are all regarded as forms of self-repair, and therefore self-repairs undoubtedly play a major role when assessing fluency in interpreting.

The phenomenon of self-repair has been of great interest in academia since Jefferson presented the 'Error Correction Format' (as cited in Rieger, 2002, p. 19). Research on classification, monitoring mechanisms, as well as internal and external triggers of self-repairs over the past decades has shed light on the cognitive process of language production. Speakers make pauses or corrections because of the presence of a monitoring mechanism in their mind; they balance their attention between different aspects of the speech and attend to certain kinds of errors or dysfluencies while ignoring others, depending on the context and on the task (Levelt, 1983). Certain subcategories of self-repair have also been investigated in studies of second language (L2) and interpreting quality assessment to explore the acoustic correlates of fluency, as well as the relationship between perceived fluency and perceived quality (Bosker, Pinget, Quené, Sanders, & De Jong, 2013; Cucchiarini, Strik, & Boves, 2000; Han, 2015; Pradas Marcias, 2006; Yang, 2002; Yu & van Heuven, 2017).

Most of the literature on self-repair in interpreting focuses on simultaneous interpreting (SI), while the role of self-repair in the quality assessment of consecutive interpreting (CI) is still under-researched. CI is not only an essential pedagogical component of conference interpreting training, it is also an important part of the curriculum for most undergraduate English majors in China. As CI requires coordinated cognitive efforts between listening comprehension, working memory, long-term memory and coherent target language production, CI training is believed to have not only effectively contributed to the skill set for conference interpretation, but also to students' overall language proficiency (Zhang & Wu, 2017). Therefore, a more comprehensive investigation of the effects of self-repairs on the perceived quality of interpreter performance may provide theoretical implications for interpreting studies. Moreover, the results may be of pedagogical significance as well because self-repair provides insight into interpreters' monitoring mechanisms in the interpreting process.

This paper sets out to investigate the effect of self-repair on the judged performance of undergraduate students in CI, using data from the semi-final of an English to Chinese interpreting contest for university students in China.

The authors acknowledge that a holistic assessment of interpreting quality involves an interplay of different criteria, however in this paper they will focus on an attempt to explore the correlations between self-repair and perceived quality in view of the three categories of *content*, *form* and *delivery*.

1. Self-repair

1.1. Self-repair in L1 and L2 language production

Levelt (1983) is widely acknowledged to have made a pioneering effort in describing a priori categories and the distribution of self-repairs. Based on an analysis of a large corpus of self-repairs spontaneously made by Dutch-speaking participants in a task, he argued that speaker's self-monitoring was probably based on parsing their own inner or overt speech. As interpreting is also a complex process of language production, his theory allows us a perspective on how interpreters monitor their inner speech, the *sense* derived from the source text, as well as their overt speech, the interpreting product.

Levelt (1983) describes a typical repair as comprising of three parts: the *original utterance* (OU), which contains the *reparandum*, the item to be repaired; the editing phase, which may or may not contain an *editing term* (ET), (uh, well, etc.); and the *repair* (R) proper, which contains an alteration (also referred to as *reparatum* in subsequent studies, e.g., van Hest, 1996). In light of this description, self-repairs are grouped into two categories: overt repairs and covert repairs, the difference being whether the *reparandum* is articulated or not. A summary of Levelt's classification is shown in Table 1.

Table 1: Classification of self-repairs in Levelt (1983)

Overt Repairs	***Different information repairs (D-repairs)***	current message replaced by a different one
	Appropriateness repairs (A-repairs)	AA-repairs ─ ambiguity reduction AL-repairs ─ appropriate level terminology AC-repairs ─ coherence with previous text ALC-repairs ─ AL or AC
	Error repairs (E-repairs)	EL-repairs ─ lexical repairs ES-repairs ─ syntactic repairs EF-repairs ─ phonetic repairs
Covert Repairs	***Pauses***	silent and filled pauses
	Repetitions	repetitions of one or more lexical items
Rest Category	***R-repairs***	repairs that defy any systematic categorization

Variance in the distribution of self-repairs can be arguably explained by the speech production theory and perceptual loop theory formulated by Levelt (and others) (as cited in Kormos, 1999). According to this model of speech production, speech processing comprises of five principal components: the conceptualizer, the formulator, the articulator, the acoustic-phonetic processor, and the parser. When producing a speech, speakers first conceptualize the message, before formulating its language representation, and finally articulating it. The perceptual loop theory assumes that a speaker monitors the speech production processes through three loops: the preverbal message is inspected against the original intention before going into the formulator; the formulated message then goes through covert or pre-articulatory monitoring before articulation, and finally the generated utterance is checked after articulation, in very much the same way that we check others' utterances (Kormos, 1999).

Based on these theories, linguistic, cognitive, pragmatic and psychological triggers of self-repairs have been investigated, and the varied distributions can be explained with the assumption that speech production monitors are more sensitive to certain types of errors in line with the characteristics and the context of a particular task. For example, studies find that speakers adopt self-repair as a strategy to gain planning time when they are cognitively strained (Al-Harahsheh, 2015; Rieger, 2003). Further, significant correlations are found between working memory capacity (WMC) and the number of different types of self-repairs in L1 and L2: L1 speakers with a larger WMC can allocate more attention to the norms of appropriacy, thus more Appropriateness repairs (A-repairs), while L2 speakers with larger WMC pay more attention to form, hence more Error repairs (E-repairs) (Mojavezi & Ahmadian, 2014). Many other studies have confirmed a relationship between self-repair and L2 proficiency. L2 speakers with a higher proficiency attend more to A-repairs, in comparison to lower proficient L2 speakers who attend more to E-repairs and Different information repairs (D-repairs) (Hennecke, 2017; Kormos, 1999; van Hest, 1996; Yang, 2002).

1.2. Self-repair in interpreting

The phenomenon of self-repair has also been investigated in interpreting studies, drawing on findings from L1 and L2 language production research. SI studies focus more on covert repairs, including pauses (filled and unfilled/silent) and repetitions, in interpretations from the second or B-language into the native or A-language. Thus, Plevoets and Defrancq (2016) explore the relationship between informational load and the occurrence rate of the editing term *uh(m)* and find that interpreters produce significantly more *uh(m)'s* than non-interpreters, which lends support to the notion that interpreting is a cognitively more demanding task than L1 and L2 speech production. Their study also finds that the input side and the output side compete for cognitive resources and pose challenges of a different nature to interpreters. In an earlier study, Petite (2005) investigates the input and output triggers of repairs and finds strong evidence that simultaneous interpreters not only repair input-generated errors to achieve greater resemblance to the source text, but also more frequently resort to output-generated repairs to achieve greater relevance, to make it easier for the receivers to understand the message.

Research on self-repair in CI is more focused on describing the distribution of different types of self-repairs, especially with a view to the discrepancies between A and B target language production, and between higher and lower interpreting competence. Some argue that problems with retrieving information from notes contribute as much to pauses as language skills (Mead, 2000; Xu, 2010), which is another interesting topic that needs more empirical evidence from future research. Also, B language proficiency has been found to play a prominent role in the occurrence of self-repair in students' interpretations (Dai, 2011; Mead, 2000; Yu, 2012; Zeng & Hong, 2012). For this reason, the authors opted to focus on A language interpretations, in an attempt to explore the correlation between self-repair and perceived interpreting quality, rather than B language proficiency.

1.3. Self-repair and interpreting quality assessment

Interpreting quality assessment can be approached from multiple perspectives and dimensions using different sets of standards and criteria (Pöchhacker, 2001). Subjective judgements can be made from various perspectives of the interpreters themselves, the users (listeners, speakers) and also the commissioners of the interpreting services (Gile, 1991), while researchers as 'external observer[s]' may take an interest in 'objective' measures of the interpreting product (Viezzi, 1996, p.12, as cited in Pöchhacker, 2001, pp.411-412). A wide range of methodological tools have thus been developed to make both measurements possible, including, for example, surveys and impression/holistic scoring for subjective assessment, and error counts, propositional scores and even acoustic features as parameters for objective quality assessment (Lee, 2015; Pöchhacker, 2001).

In response to the notion that evaluating interpreter performance in the classroom should receive scholarly attention, Lee (2015) conducted an online questionnaire survey of interpreter trainers, and grouped a final list of 21 weighted criteria into three categories: content, form, and delivery. While perceptions certainly vary regarding what criteria should be incorporated to represent each category, the three broad categories described by Lee (2015) would be generally agreed upon, especially given the face-to-face communication genre of CI.

Examiners' holistic judgement is still heavily relied upon in interpreter training programs, and one of the problems with subjective assessment as revealed in previous experiments is the so-called 'halo effect' (Fulcher, 2010, p.209, as cited in Lee, 2015, p.230), where the judgement of one particular criterion affects that of other criteria used in one assessment. For example, pauses as a subparameter of fluency have a negative effect on fluency evaluation (Pradas Marcias, 2006), while a lower fluency in turn shows a tendency to impact negatively on perceived quality (Rennert, 2010; Yu & van Heuven, 2017). Therefore, it is necessary to explore further how these variables of interpreting quality are related to and affect one another.

Research on how prosodic factors influence interpreting performance mainly draws on quantitative assessment of L2 fluency, such as the ones shown in Table 2, in which filled pauses, repetitions, restarts, and repairs are transcribed exactly as they were uttered and then counted, and silent pauses are detected by running software and then automatically computed. Different quantitative measures are adopted in these three studies, and the variables which fall into categories of self-repairs are bolded by the authors.

It should be noted, however, that a cut-off point of 0.25 seconds is taken in Yu and van Heuven's (2017) research, but no differentiation is made between syntactic and non-syntactic pauses. The position of pauses plays a strong role in subjective experience because we are more tolerant of pauses between sentences or phrases and sometimes perceive a pause where there is no actual interruption (Rennert, 2010). Therefore, acoustic measures should be combined with subjective assessment to make better judgement of perceived pauses.

Table 2: Self-repair in quantitative measures of fluency

Cucchiarini et al. (2000), L2	**Bosker et al. (2013), L2**	**Yu & van Heuven (2017), CI**
Seven primary variables	*Speed*	Articulation rate
Articulation rate	Mean length of syllables (MLS)	Speech rate
Rate of speech		Effective speech rate
Phonation/time ratio	*Breakdown*	**Number of silent pauses**
Mean length of runs (Mean number	**Number of silent pauses (NSP)**	**above 0.25 seconds in**
of phonemes between silent pauses)	**Number of filled pauses (NFP)**	**duration**
Mean length of silent pauses	**Mean length of silent pauses (MLP)**	**Mean length of silent pauses**
Duration of silent pauses per minute		**longer than 0. 25 seconds**
Number of silent pauses per minute	*Repair*	**Number of filled pauses (uh,**
	Number of repetitions (NR)	**er, mm, etc.)**
Two secondary variables	**Number of corrections (NC)**	**Mean length of all filled**
Number of filled pauses per minute		**pauses;**
Number of disfluencies per minute		**Number of pauses**
		Mean length of pauses
		Number of other disfluencies
		Mean length of fluent runs
		Phonation/time ratio

Another problem with quantitative assessment is that using a larger number of acoustic measures would increase the chance of confounding the different measures. For example, the relative contribution of *speech rate* and *mean length of silent pauses* to perceived fluency would remain unclear because both measures depend on the duration of silent pauses. Therefore, correlations among acoustic measures should also be taken into account (Bosker et al., 2013), and for practical purposes, a small number (three in Mead's research) of parameters may be predictive enough to assess interpreter's fluency (Mead, 2005).

In short, exploring the correlations between self-repair and the perceived quality of interpreting performance may involve more theoretical and pedagogical implications than investigating individual subcategories of self-repair as temporal parameters of fluency, because self-repair also offers insight into how interpreters monitor the interpreting process. Secondly, in investigating such correlations, acoustic measures should be selected rationally and combined with subjective assessment for better judgement. Finally, data from real-life settings should complement finding (Yu & van Heuven, 2017).

2. Research questions

The design of the current study was inspired by the three categories of content, form and delivery in the analytic rating scale developed by Lee (2015). These three categories correspond roughly to the rubrics adopted in a semi-final interpreting contest for undergraduate English majors in China, namely *information*, *delivery of message*, and *professionalism* (See Appendix). Data from this contest were also used because the contest involved a simulated real-life CI setting, where student interpreters faced a large audience of judges and students and teachers from participating universities, and were hence under multiple pressures including overcoming nervousness and striving to meet professional standards. The audio and video recordings of the contest were acquired with the consent of the organizers of the contest. The judges' assessments were also attained, and the authors designed the ratings of individual aspects of the performances (see section 4.2.2). The data were used to try and answer the following two research questions (RQ):

> RQ1: What are the characteristics and the distributions of different types of self-repairs in student interpreters' CI performance?
> RQ2: What are the relationships between self-repairs and interpreting quality assessments, from the perspective of both perceived quality overall as well as the categories of content, form and delivery?

3. Methods

3.1. Participants

A total of 35 students from over twenty Chinese universities attended the contest, but interpretations by only 12 participants were used as research data for reasons that will be discussed later. All participants were senior English majors aged 21 or 22, who had started to learn English at the age of 12 or earlier, and who had undertaken intensive training in English for 3.5 years since they started university. They had just finished one semester of interpreting classes and undergone some additional CI training, including note-taking skills, by the time of the contest. The contestants also had to complete a preliminary round of competition at their respective universities in order to qualify for the semi-final.

In the first round of the English to Chinese interpreting task, the recording of a speech titled "currency manipulation" was played once, non-stop. The speech had 224 words, lasting one minute 15 seconds at a rate of 179 words per minute. The contestants began to interpret after a beep signal, and had to finish within one minute and 30 seconds of hearing the speech. Five judges rated the interpreted renditions in accordance with the rubrics used for assessment at the contest, the main criteria including information, delivery of message, and professionalism (see Appendix). The contest recessed after the first three contestants so that the judges could compare their grades, given on a 100-point scale, and have discussions to reach an agreed understanding of the rubrics. The final score for each contestant was the sum of the five grades given by the judges. The 12 contestants who scored the highest were selected as the research participants because their general interpreting qualities were comparable and well above average, so it is safe to assume that the subjective assessments would not be predominantly undermined by any particularly poor aspect of the performance. For example, when a contestant made major mistakes or missed too much of the message, the assessments of form and delivery may have only played a negligible role as compared to content, and it would therefore make little sense to analyze any correlations between self-repair and the perceived assessments of interpreting quality agreed upon by the judges.

The twelve interpretations were transcribed by a postgraduate assistant, and checked by the first author to make sure that all repairs and filled pauses were transcribed exactly as they were uttered.

3.2. Material

3.2.1 Self-repair

All overt repairs were annotated in accordance with Levelt's classification (1983) (summarized in Table 1) by the same specially trained postgraduate assistant, and checked by the first author. Inconsistencies were discussed, and the second author was consulted until an agreement was reached. A closer look at an undefinable group of repairs in the data shows that they were the results of failed attempts by the student interpreters to correct a mistake or modify an expression, and were thus annotated as FRs (failed repairs) instead of Rest Repairs.

The annotation of covert repairs was more complicated. Firstly, in addition to repetitions of the same lexical items (annotated as S1), another type of repetition was also found in the data. The interpreter in question had apparently used a different expression to render the same information unit that had been rendered immediately before, and therefore such renderings were also annotated as repetitions but with 'S2', that is a similar rendering of the same lexical items in the ST.

> Example 1:
> Source Text (ST): And is there really anything wrong with this…
> Target Text (TT): 又是怎样的一个错误，或者说一个其它的问题所在呢 [S2]?
> Literal translation of TT: What kind of mistake is it? Or is it another problem?

Secondly, a range of 0.25 to 2 seconds of the cut-off criterion for pauses has been used in the interpreting literature (Han, 2015). Considering the direct communicative nature of CI, another three raters who had not been informed of the ST watched the video of the first three interpreters, and noted down the perceived pauses. Based on this, a cut-off criterion of 0.5 seconds for non-syntactic (silent, filled) pauses was agreed upon after discussion, and filled pauses shorter than 0.5 seconds were annotated together with editing terms in overt repairs as ET. The detection and measuring of pauses in the recordings was done manually with the WavePad Sound Editor at a sampling frequency of 44100 Hz. Silences at the very beginning and the end of every interpretation were discarded by the trim function in the software.

Four objective measures were calculated for each interpretation to reflect the self-repair features as shown in Table 3. The aspect of overt repairs was represented by one measure: the number of overt repairs. The number of editing terms (ETs) was added to the total number of overt repairs because disguising editing terms is regarded as a tactic by experienced interpreters (Petite, 2005) and the pronounced fillers would nevertheless affect the flow of the speech. The covert repair aspect was represented by three measures: the number of repetitions (NR), the number of pauses (NP) and the mean length of pauses (MLP). All frequency measures were calculated using spoken time to avoid confounding the different measures (Bosker et al., 2013)

Table 3: List of four objective measures of self-repair

Category	Measures	Calculation
Overt repairs	Number of overt repairs (NOR)	(Number of D/A/E repairs + ETs+FRs)/spoken time
Covert repairs	Number of repetitions (NR)	Number of S1S2 repairs/spoken time
	Number of pauses (NP)	Number of silent and filled pauses/spoken time
	Mean length of pauses (MLP)	Sum of length of silent and filled pauses/number of silent and filled pauses

3.2.2 Subjective assessments

The overall perceived quality was represented by the final scores given by the five judges in the semi-final. Four of the five judges were experienced interpreter trainers from four Chinese universities, while the other judge was a senior interpreter from the local municipal government. For the subjective assessments of individual aspects of the performance, different raters and different methods were used to avoid a halo effect.

3.2.2.1 Content

The method for rating content was a proposition-based one, which has been proved effective in evaluating the accuracy of interpreting performance (Chen, 2017). The ST was divided into 25 scoring propositions by the first author and a second rater, who had both had training and had experience undertaking proposition-based scoring following a previous research project. Inconsistencies were discussed until an agreement was reached. In addition, fourteen cohesive links in the ST were identified in the same way. The assessment of cohesive links was also included because logical cohesion ranked among the top criteria of quality based on the existing literature, and "the cohesion…can be described in terms of the formal (syntactic and semantic) links" (Widdowson, 1978, as cited in Li, 2016, p.38). A score of 1 was given when the meaning of a proposition was correctly interpreted. For a link to gain an extra score of 0.5, it had to be immediately adjacent to the proposition that it was used to link, as indicated in the ST. In addition, the rendering of the link played a comparable linking role in the TT, but did not necessarily represent the same syntactic or semantic form as the corresponding one in the ST. Added information and erroneous renderings were not penalized. Kendall's *W* was run to determine if there was agreement. The agreement between the two raters was statistically significant: $W = .967$, $p < .0005$.

3.2.2.2 Form

The definition of form was readjusted in the present study by taking into consideration both the criteria used in previous studies and the rubrics of the competition semi-final so as to reflect a fairly independent parameter of the quality assessment. The three raters, who had participated in the judgement of perceived pauses (see 4.2.1), were asked to rate the form of performances on a 10-point scale (1 being very poor and 10 very good) while watching the video recordings. They were instructed to base their judgments on (1) impression of confidence, (2) poise, (3) appropriate register, (4) pleasant appearance, (5) pleasant voice (lively intonation and stress, unambiguous and clear diction), and (6) perceived reliability. A short interview was also conducted after the assessment as to which criteria might have affected their ratings the most.

The three raters were all PhD candidates in Linguistics, and as they had not been informed of the ST, their assessments could be justified as representative of end users of the interpreting service, whose presumed expectations were also an important part of quality assessment (Kurz, 2001). Also, raters' subjective assessments of particular aspects of a speech or interpretation were proven to be valid in previous studies (Bosker et al., 2013; Yu & van Heuven, 2017). Kendall's *W* showed that the agreement among the three raters was statistically significant: $W=.858$, $p< .0005$.

3.2.2.3 Delivery

Fluency is arguably a more dominant criterion of delivery. The objective acoustic measure of *effective speech rate* was adopted to represent the judged fluency of performances for two reasons. Firstly, effective speech rate was found to be the best indicator of perceived fluency in CI (Yu & van Heuven, 2017); secondly, another subjective assessment of fluency would have been contaminated by the halo effect, thus undermining the interpretability of the relative contributions of self-repair to the perceived assessments of form and delivery. The effective speech rate was calculated as number of syllables, excluding disfluencies from the syllable count, divided by total duration of speech including all (silent and filled) pauses.

4. Results and discussion

The results of the self-repair distributions are presented first, followed by the subjective assessments of different aspects of the interpreting performances. Finally, the correlations between these are discussed.

4.1. Distribution of self-repairs

The authors found Levelt's classification (1983) to be effective in describing self-repairs in students' interpretations from English into Chinese. The data revealed significant differences in the distribution of self-repairs between students' A language interpretations and L1 production based on previous findings (Brédart, 1991). The percentage of covert repairs was 47.1% (since only pauses >0.5 seconds were counted in the present study, the actual number could be even larger), much higher than the 24.6% in Levelt's and the 41.5% in Brédart's studies on L1 self-repairs (Brédart, 1991). According to speech production theory and perceptual loop theory, covert repairs occur before an utterance is articulated, when the speaker is inspecting a preverbal message against the original intention, the ST in this case, or is wording the message with the language formulator. Given that the students interpreted into their A language Chinese, it can be argued that conceptualizing the message in the ST, that is deverbalization in the interpreting process, is a more prominent cause of covert repairs in students' interpretations. Secondly, a certain level of professionalism can be inferred from the distribution of repairs. A-repairs were found in the data, indicating that students were also monitoring the production process in the interest of the audience, while only a very small number of errors were left unrepaired (4.8%) to maintain the integrity of the interpretation. Detailed information about students' interpretations and the self-repair distribution is shown in Table 4.

4.2. Subjective assessments of interpreting performances

In rating the content aspect of interpretations, it was found that the students had little difficulty re-presenting the common linking devices such as "for example" and conjunctions like "but". However, they were not as good at identifying the relationships implied by some adverbs, which compromised the coherence of the interpretation. Examples include "though" in "China has a mighty wand on its hand though", and "however" in "it is encouraging to note, however, that…'. Also, the assessment of cohesive links could get a little tricky because of the syntactic differences between the Chinese and English language.

> Example 2:
> SS: This thus keeps the Yuan undervalued.
> TT: 这使得人民币得以贬值。
> Literal translation of TT: This leads to the devaluation of RMB.

As is shown in Example 2, the adverb "thus" is used as a cohesive link to indicate a cause-effect relationship. The same relationship was built semantically with the predicate "leads to". General agreement was reached after discussion, and an extra 0.5 was awarded for the achievement of the coherent relation, but further research into the differences in syntactic and semantic linking devices between English and Chinese is suggested for future studies.

After rating the form aspect of interpretations, the raters briefly discussed the criteria used in the assessment. Raters seemed to be more sensitive to different problems in the performances in their own particular ways, and this again reflects how highly subjective interpreting quality assessment can be. "No eye contact whatsoever", "mistake against common sense", "unease voice", "not very confident", etc. were reported to have affected their assessment. However, a satisfactory concordance was found eventually between the raters, which means that the ratings could nevertheless represent how the audience might have perceived interpreting quality in general. Moreover, the mention of "mistake against common sense" indicated that perhaps users of interpreting services with no knowledge of the ST may still reach sound judgments of the accuracy of interpretations. As *no opposite meanings* is also a heavily weighted criterion in Lee's (2015) scale, it may be reasonable to introduce penalties for opposite meanings in the assessment of content to better reflect the correlation between content and perceived quality in future studies.

Table 4: Interpretation measures and self-repair distribution

No.	Measures of Interpretations						Overt Repairs						Covert Repairs					
	Duration second	***SP length***	***FP length***	***Word count***	***Word excl. repairs***	***ESR***	***D***	***A***	***E***	***FR***	***ET incl. FP<0.5s***	***Total***	***S1***	***S2***	***Subtotal S1+S2***	***SP***	***FP***	***Subtotal SP+FP***
1	80.07	_	1.37	346	330	4.12	1	1	2	1	4	9	5	1	6	0	2	2
2	86.36	0.60	5.15	465	346	4.01	3	0	4	3	12	22	6	2	8	1	7	8
3	72.73	0.50	1.09	350	338	4.65	2	0	1	1	1	5	2	1	3	1	1	2
4	59.39	0.57	5.75	245	233	3.92	0	2	1	0	4	7	0	0	0	1	7	8
5	71.39	0.83	1.08	349	341	4.78	0	2	0	2	3	7	0	1	1	1	2	3
6	51.93	0.65	0.54	248	240	4.62	0	2	0	0	1	3	2	0	2	1	1	2
7	43.03	0.50	_	213	205	4.76	4	1	0	0	1	6	4	1	5	1	0	1
8	69.39	1.13	3.43	306	290	4.18	2	0	1	0	4	7	2	0	2	1	6	7
9	71.78	_	4.11	355	321	4.47	0	0	1	1	11	13	3	0	3	0	6	6
10	84.94	_	4.81	366	340	4.00	1	2	1	0	6	10	4	1	5	0	7	7
11	69.95	5.08	1.21	286	251	3.59	2	1	1	2	3	9	6	0	6	7	1	8
12	74.43	–	0.76	378	348	4.68	1	2	2	0	7	12	1	1	2	0	1	1
Repair Total	208						16	13	14	10	57	110 (52.9%)	35	8	43 (20.7%)	14	41	55 (26.4%)

Table notes. SP=silent pause, FP=filled pause, ESR=effective speech rate, D=different information repairs, A=appropriateness repairs, E=error repairs, FR=failed repairs, ET=editing term, S1=repetition of the same lexical items, S2=similar rendering of the same lexical items in ST

Finally, the delivery aspect of the performances was represented by the objective measure of effective speech rate. It is worth noting, however, that the different criteria adopted to define and measure pauses (see 4.2.1.) may have led to a variance in the explanatory power of delivery in the present study. A summary of all subjective assessments is shown in Table 5.

Table 5: Overall perceived quality and subjective assessments of content, form and delivery

No.	Judges	Content			Form				Delivery
		Rater 1 L+P	***Rater 2 L+P***	***Mean***	***Rater 1***	***Rater 2***	***Rater 3***	***Mean***	***ESR***
1	470.15	3.5+9.5	2.5+10	12.75	7	8	8	7.67	4.12
2	459.05	1.5+7.5	1+6.5	8.25	6	7	6	6.33	4.01
3	455.31	1.5+5.5	1+5.5	6.75	8	9	7	8.00	4.65
4	461.12	1.5+9.5	2+8	10.50	5	6	6	5.33	3.92
5	468.94	1.5+10.5	1.5+10	11.75	9	8	9	8.67	4.78
6	456.46	1+7	1+6	7.50	7	7	8	7.33	4.62
7	460.08	2.5+11.5	2.5+10.5	13.50	4	5	5	4.67	4.76
8	478.99	0.5+9.5	1+9	10.00	7	6	7	6.67	4.18
9	468.86	1+12	1+10	12.00	6	6	7	6.33	4.47
10	462.92	3.5+12.5	3.5+10.5	15	6	5	6	5.67	4.00
11	456.51	2+10	2.5+10.5	12.5	5	4	6	5.00	3.59
12	465.68	3.5+12	3+12.5	15.50	8	9	7	8.00	4.68

Table notes. L=cohesive links, P=propositions

4.3. Correlations between self-repair measures and subjective assessments of CI quality

The self-repair measures were compared with different categories of subjective ratings to determine how, and to what extent, these were related. Pearson's *r* correlations were run to assess such relationships. Any outlier was removed when detected in the preliminary analysis of whether the variables were normally distributed, because it would otherwise have compromised the data, and this would have run contrary to one of the research goals, which was to generalize any findings. To this end, possible reasons for the occurrence of outliers were analysed. Pearson's *r* results are shown in Table 6.

4.3.1 Correlations between overt repairs (represented by NOR) and subjective assessments

In the preliminary analysis of whether the variables were normally distributed, student 2 was determined to be an outlier in the linear relationship between NORs and content. The distribution of self-repairs in Table 4 shows that the total number of overt repairs in his/her interpretation was 22, nearly twice as many as the second highest number, which was 13 in student 9's interpretation. This could lead to the abnormal distribution of values, so student 2's data were removed.

A strong positive correlation of r=0.712* was found between NORs and content. Overt repairs are made when an error is detected in an articulated utterance: D-repairs (see Table 1) are made when the interpreter wants to provide a better organized message or replace it with a new one, A-repairs are made when the interpreter wants to rephrase an expression mainly for the benefit of the audience, and E-repairs are made when a linguistic error is detected in the utterance. As the distribution of self-repairs reflects the monitor bias of the speaker in the language production process, it can be assumed that the students who scored higher on content may be linguistically more competent or have a larger WMC, since they could cognitively manage to monitor their interpreting output. Little correlation was found between NORs and the overall perceived quality, indicating that the judges did not seem to penalize overt repairs in students' interpretations. However, moderate negative correlations of *r*=-0.330 and *r*=-0.299 were found between NORs and form and delivery respectively, which shows that an audience's subjective assessments of whether the interpreter looked professional and whether the message was delivered fluently may have been affected by overt repairs to a certain extent. The case of student 2, though regarded as an outlier, also shows that when the number of overt repairs exceeds a certain level, the above analysis would no longer be applicable to any correlations between the variables.

Table 6: Correlations between self-repair measures and subjective assessments

Self-repair Measures	Content	Form	Delivery	Overall Perceived Quality
NOR	0.712* (No. 2 removed)	-0.330	-.299	0.005
NR	0.142	-0.774** (No. 4 removed)	-0.541 (No. 7 removed)	-0.430 (No. 4 removed)
NP	-0.118	-0.866** (No. 7 removed)	-0.789**	-0.208 (No. 8 removed)
MLP	-0.112	-0.303 (No. 7 removed)	-.492	-0.208

Table notes.
*p<.05 (two-tailed)
**p<.01(two-tailed)

4.3.2 Correlations between covert repairs (represented by NR, NP and MLP) and subjective assessments

In the preliminary analysis of the linear relationships between covert repair measures and the subjective assessments, a few values were identified as outliers and therefore removed where appropriate. For the variable of NR, as student 4 made no repetitions, while student 7 had the highest NR value, they were identified as outliers. Student 7 was also an outlier in the relationships between form and NP and MLP, probably because he/she received the lowest score for form. For similar reasons, student 8, who received the highest score from the judges, was identified as an outlier in the linear relationship between NP and overall perceived quality.

The results show that the three covert repair measures have low correlations with the assessment of content, but stronger negative correlations with the other three subjective assessments. That is to say, the raters and the judges seemed to be inclined to penalize covert repairs more than overt repairs.

Both NRs and NPs were found to have strong negative correlations with form and delivery, which denotes that repetitions and pauses in interpretations may have seriously diminished the end users' evaluation of the student interpreters' fluency and credibility. MLPs were found to only have a moderate correlation with form and delivery, probably due to the fact that students may have strived to shorten pauses and finish interpreting within a given set time in the contest. As for the judges' assessment, a higher negative correlation was found between NRs than NPs and the overall perceived quality, indicating that the judges may have been more sensitive to repetitions than to pauses. Based on our analysis above, covert repairs may have resulted from cognitive strains in the conceptualizing part of the language production process. In other words, it is very likely that student interpreters were struggling to reformulate the message by repeating the same or similar lexical items or employing pauses to gain some thinking time.

4.3.3 Content, form and delivery as predictors of overall perceived quality

From the analysis above, it seems that self-repair played a less notable role in the judges' evaluation than the other subjective assessments. In an attempt to explore further the relationship, a multiple regression was run to predict the dependent variable of overall perceived quality from the three independent variables of 'content', 'form' and 'delivery'. Unfortunately, no other statistically significant results were found aside from a value of R=0.471, indicating a moderate linear association between the dependent variable and the independent variables. This result has the following implications. Firstly, different criteria may have influenced the judges to a varying extent despite the use of set rubrics in the contest. A Pearson's correlation of r=0.559 (outlier student 8's data removed) shows that the judges' scores were strongly related to the content aspect of the interpreting performances. While it would be impossible to find out how much the judges weighted each criterion, it could be inferred that content may have played a predominant role in their assessment. In other words, the judges, who understood both the ST and the TT, still relied heavily on accuracy of information in their assessment of interpreting quality. Secondly, only twelve interpretations from the total number of 35 were analyzed for the present study. Considering that the

twelve participants were randomly distributed among the 35 contestants, whom the judges evaluated consecutively at one sitting, it cannot be ruled out that the factors of physical or mental exertion and an order effect could also have played a role. In addition, the results of subjective assessments of interpreting performances in a contest may have been influenced by many other unidentified factors. Finally, the correlations between self-repair and form and delivery may suggest a stronger impact of self-repair on the subject assessment of interpreting quality in real-life settings.

5. Conclusion

This paper investigated the effects of self-repairs on the subjective assessments of the three aspects of CI performance, namely content, form and delivery, and an overall perceived quality. The data collected from an interpreting contest were arguably more representative of how student interpreters would monitor their interpreting process in a real-life CI setting than in experiments. In an experimental setting, participants often perform the interpreting tasks in a more solitary environment with no added concerns about the quality of their interpretations. By comparison, the contestants had to manage the stresses of interpreting in the presence of a large audience, having a time limit to complete the task, while striving for a high quality performance in accordance with the assessment rubrics used at the contest. Our analysis showed that students who scored higher on content used more overt repairs, while those who scored lower used more repetitions and pauses. The negative ratings of form and delivery were closely correlated with covert repairs, in terms of the frequencies of repetitions and pauses and the mean length of pauses. Lastly, even though self-repair did not seem to have affected the judges' assessment in the contest, it may play a more prominent role in real-life interpreting settings.

One of the limitations of the study is the comparatively small data set for determining correlations between quantitative variables. The twelve participants were selected for reasons stated in the methodological part of the paper, however future research may benefit from a larger data set involving interpretations of more mixed quality, and subjective assessments by raters from more varied backgrounds.

The findings of the study provide some reference values for interpreter trainers. As a contribution of the study to CI training, the types and distribution of self-repairs may be used as additional information for trainers to identify problems in students' performance, and thus help students address them accordingly. As suggested in data analysis, a higher presence of overt repairs may indicate that students have more developed interpreting competence and larger WMC, while more covert repairs could indicate that students still have difficulty conceptualizing the sense of the ST. Secondly, while self-repairs are generally discouraged in interpreter training, trainers could differentiate between overt and covert repairs as subjective assessments of interpreting quality seem to be more sensitive to covert repairs including repetitions and pauses. Finally, interpreter trainers should be aware that, despite the existence of rubrics, their assessments of student's performance may still be predominantly influenced by accuracy, because trainers understand both the source and target spoken texts. Awareness of this bias may result in the development of new evaluation practices in interpreting training, for example, introducing additional examiners, who do not listen to the source text but only the interpreted renderings, to the examiners panel.

References

Al-Harahsheh, A. M. A. (2015). A Conversation Analysis of self-initiated repair structures in Jordanian Spoken Arabic. Discourse Studies, 17(4), 397-414.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. Language Testing, 30(2), 159-175.

Brédart, S. (1991). Word interruption in self-repairing. Journal of Psycholinguistic Research, 20(2), 123-138.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. Multilingua, 5(4), 231-235.

Chen, S. (2017). Note-taking in consecutive interpreting: New data from pen recording. Translation & Interpreting, 9(1), 4-24.

Chiaro, D., & Nocella, G. (2004). Interpreters' perception of linguistic and non-linguistic factors affecting quality: A survey through the World Wide Web. Meta: Journal des traducteurs/Meta: Translators' Journal, 49(2), 278-293.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America, 107(2), 989-999.

Dai, Z. (2011). A study on disfluency in Chinese to English interpretations of Chinese EFL learners. Shanghai Journal of Translators, 1, 38-43.

Gile, D. (1991). A communication-oriented analysis of quality in nonliterary translation and interpretation. Translation: Theory and practice. Tension and interdependence, 5, 188-200.

Han, C. (2015). (Para) linguistic Correlates of Perceived Fluency in English-to-Chinese Simultaneous Interpretation. International Journal of Comparative Literature and Translation Studies, 3(4), 32-37.

Hennecke, I. (2013). Self-repair and language selection in bilingual speech processing. Discours: Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, 12, 1-20.

Kormos, J. (1999). Monitoring and self-repair in L2. Language learning, 49(2), 303-342.

Kurz, I. (2001). Conference interpreting: Quality in the ears of the user. Meta: journal des traducteurs/Meta: Translators' Journal, 46(2), 394-409.

Lee, S. B. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. Interpreting, 17(2), 226-254.

Levelt, W. J. (1983). Monitoring and self-repair in speech. Cognition, 14(1), 41-104.

Li, X. (2016). *Patterns of Coherence in Translation: A Case Study of Hong Lou Meng*. Unpublished doctoral dissertation, Macquarie University, Sydney.

Mead, P. (2000). Control of pauses by trainee interpreters in their A and B languages. The Interpreters' Newsletter, 10, 89-102.

Mead, P. (2005). Methodological issues in the study of interpreters' fluency. The Interpreters' Newsletter, 13, 39-63.

Mojavezi, A., & Ahmadian, M. J. (2014). Working memory capacity and self-repair behavior in first and second language oral production. Journal of psycholinguistic research, 43(3), 289-297.

Petite, C. (2005). Evidence of repair mechanisms in simultaneous interpreting: A corpus-based analysis. Interpreting, 7(1), 27-49.

Plevoets, K., & Defrancq, B. (2016). The effect of informational load on disfluencies in interpreting. Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association, 11(2), 202-224.

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. Meta: Journal des traducteurs/Meta: Translators' Journal, 46(2), 410-425.

Pöchhacker, F. (2012). Interpreting quality: Global professional standards. In W. Ren (Ed.), *Interpreting in the Age of Globalization: Proceedings of the 8th National Conference and International Forum on Interpreting* (pp.305-318). Beijing: Foreign Language Teaching and Research Press.

Pradas Marcias, E. M. (2006). Probing quality criteria in simultaneous interpreting. Interpreting, 8(1), 25-43.

Rennert, S. (2010). The impact of fluency on the subjective assessment of interpreting quality. The Interpreters' Newsletter, 15, 101-115.

Rieger, C. L. (2002). *Self-repair strategies of English-German bilinguals in informal conversations: The role of language, gender, and linguistic proficiency*. Unpublished doctoral dissertation, University of Alberta, Edmonton.

Rieger, C. L. (2003). Repetitions as self-repair strategies in English and German conversations. Journal of Pragmatics, 35(1), 47-69.

Van Hest, G. W. C. M. (1996). *Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press.

Xu, H. (2010). Pauses in conference consecutive interpreting from English into Chinese: An empirical study. Foreign Languages Research, 1, 64-71.

Yang, L. (2002). The effect of English proficiency on self-repair behavior of EFL Learners. Shandong Foreign Language Teaching Journal, 4, 74-76.

Yu, W. (2012). *Self-monitoring and Self-repair Patterns in Consecutive Interpreting*. Unpublished doctoral dissertation, Shanghai International Studies University, Shanghai.

Yu, W., & van Heuven, V. J. (2017). Predicting judged fluency of consecutive interpreting from acoustic measures. Interpreting, 19(1), 47-68.

Zhang, T., & Wu, Z. (2017). The Impact of Consecutive Interpreting Training on the L2 Listening Competence Enhancement. English Language Teaching, 10(1), 72-83.

Zeng, J., & Hong, M. (2012). A study on self-repair in Chinese-English consecutive interpreting by student interpreters. Foreign Languages and Their Teaching, 2, 68-72.

Appendix. Rubrics for the English to Chinese Consecutive Interpreting contest

Categories	Criteria
Information (50%)	• Accurate rendition of main ideas • Maintaining the tone and the style of the source speech
Delivery of message (25%)	• Grammatical correctness • Logical cohesion • Appropriate register • Appropriate use of interpreting strategies, e.g. addition, abstraction, paraphrasing and omission • Natural/idiomatic target-language expressions • Variety of words and expressions
Professionalism (25%)	• Pleasant voice • Good/natural pronunciation • Not speaking too fast/slowly • Fluency of delivery • Eye contact with the audience • Giving an impression of confidence • Showing professionalism • Extra-linguistic communication skills