Measuring translation revision competence and post-editing competence in translation trainees : methodological issues

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

Measuring translation revision competence and post-editing competence in translation trainees: Methodological issues

Isabelle S. Robert, University of Antwerp (https://orcid.org/0000-0002-8595-0691)

isabelle.robert@uantwerpen.be

Iris Schrijver, University of Antwerp (https://orcid.org/0000-0001-6091-024X)

iris.schrijver@uantwerpen.be

Jim J. Ureel, University of Antwerp (https://orcid.org/0000-0003-3584-5612)

jim.ureel@uantwerpen.be

Abstract

Translation proper is rarely the sole activity of professional translators, who regularly function also as revisers and/or post-editors. Various models of and studies into translation competence (TC), translation revision competence (TRC) and post-editing competence (PEC) exist. However, a fundamental question remains unanswered: how similar – or different – are TC, TRC and PEC? Before this question can be answered, a methodological issue must be addressed: how do we measure TRC and PEC? Using existing literature, we propose seven instruments to measure TRC and PEC. Our aim is to determine whether the instruments are exchangeable, that is, result in similar measures of the underlying variable. We conducted a small-scale study with translation trainees, who performed L1 Dutch–L2 French TR and PE tasks. The measuring instruments generated TRC scores that were significantly different and therefore not exchangeable. In contrast, PEC scores were not always significantly different. In conclusion, with measuring instruments for TRC and PEC being generally not exchangeable, it is imperative that researchers not only report on measuring instruments thoroughly in research in general, but also use measuring instruments designed according to the same principles when they investigate differences and similarities between TRC, PEC, and even TC.

Keywords: translation revision, post-editing, competence models, quality assessment, competence measuring instruments

1. Introduction

Whereas in the past translators started from a clean sheet to produce target texts, they nowadays work increasingly less from scratch (Jakobsen, 2019; Koponen et al., 2021). Translators also regularly revise, which entails reading a human translation to "find features of a draft translation that fall short of what is acceptable ... and make or recommend any needed corrections and needed improvements" (Mossop, 2020, p. 115). Moreover, with the development of computer-assisted translation (CAT) tools, translators rely on translation memories (TMs), which means that they are, in a sense, revising reused human translations. When there are no adequate stored translations, CAT tools often integrate machine translation (MT) as well. In that case, translators become post-editors, with post-editing (PE) being the term used to refer to revising machine-generated output.

In view of translators' evolving skill set, it is paramount that we examine the competences[i] required to cope with new working conditions, and, in particular, the difference(s) – if any – between translation competence (TC), translation revision competence (TRC) and post-editing competence (PEC). TC Models such as PACTE (2003), Göpferich (2009) and EMT Expert Group (2009, 2017) have been well-established in Translation Studies (TS) for quite some time. However, research interest in translation revision (TR) and PE has been gaining momentum only in recent years. Despite a growing number of TR and PE publications (for an overview, see Koponen et al., 2021), TRC and PEC research is still limited. As we will explain in Section 2, the few existing TRC and PEC models share some components or subcompetences while still being different. This is most likely the result of scholars generally building on existing models to design their own models.

The underlying hypothesis of existing TRC and PEC models is that TC, TRC and PEC are different but share common ground. This begs the fundamental question: *how* different are TC, TRC and PEC? One could hypothesize that TRC and PEC are more similar to each other than to TC, since TR and PE – contrary to translation – share the same starting point: an existing target text. Our rationale is based on Pym's (2003) minimalist definition of translation competence. If generating and selecting a target text is at the core of TC, this is where TC differs from TRC. In the initial TR process, text generation and text selection are not required, since (a version of) the target text has already been created. The same holds true for PE, except that the text created has been produced by a machine and not a human translator.

However, before we can empirically study the relationship between TC, TRC and PEC, a methodological issue must be solved: how do you measure TRC and PEC? Measuring TC is related to translation quality assessment and has been and is still being investigated (more than 1000 hits in the Translation Studies Bibliography in August 2021). However, with TRC and PEC having different starting points, the question remains unsolved. Consequently, the two research questions (RQs) that we will address in this paper are: (1) what kind of measuring instruments or indicators can be used and/or developed to measure TRC and PEC[ii] and (2) are they exchangeable, in other words, do they result in similar measurements of the underlying variable, that is, TRC and PEC?

To answer RQ1, we reviewed the different TRC and PEC models (Section 2) as well as TC measuring instruments. To answer RQ2, we conducted an experimental pilot study with translation trainees (Sections 3 and 4). We hypothesize that, if different measuring instruments yield the same results, it can be argued that they are exchangeable.

2. Translation revision competence (TRC) and post-editing competence (PEC)

2.1. Translation revision and post-editing competence models

To design their own TRC and PEC models, scholars generally build on existing TC models. TC and TC acquisition have been studied extensively in TS. When discussing TC, scholars generally provide overviews of existing TC models (see, for example, Chodkiewicz, 2020; Kornacki, 2018; Massey, 2017; PACTE, 2020; Tiselius & Hild, 2017). Such overviews almost always include the multicomponential construct models developed by the PACTE research group (2003, 2005; Hurtado Albir, 2017), Göpferich (2009) or the EMT Expert Group (2017), in addition to Pym's (2003) minimalist definition of TC. We will not address these models in

detail here. Suffice it to say that they all include a series of common subcompetences (e.g., bilingual competence, extra-linguistic competence, instrumental competence, strategic competence), as well as additional competences (e.g., knowledge-about-translation competence (PACTE), translation-routine-activation subcompetence (Göpferich), service-provision competence (EMT)).

In contrast, there is much less TRC research. Robert et al. (2017) designed a model based on established TC models (EMT, Göpferich, PACTE) and on related research on revision training and competence (models) (Bisaillon, 2007; Hansen, 2009; Kelly, 2005; Künzli, 2006; Mossop, 1992). Their TRC model consists of nine interconnected subcompetences, with some specific to revision, such as knowledge-about-revision subcompetence or strategic subcompetence for revision. Robin (2016) also proposed a TRC model, consisting of seven subcompetences: (1) ameliorative, (2) evaluative, (3) translation, (4) comparative-contrastive, (5) corrective, (6) linguistic and (7) decision-making subcompetence. This model appears more process-orientated. More recently, Scocchera (2017) suggested a multicomponential TRC model, consisting of six subcompetences: (1) analytical-critical, (2) operational, (3) metalinguistic-descriptive, (4) interpersonal, (5) instrumental and (6) psycho-physiological competence.

PEC research is also relatively limited, although the PE process has been investigated extensively (for an overview, see Koponen et al., 2021, pp. 1–17; Nunes Vieira et al., 2019). The first PEC model is Rico and Torrejón's (2012), who integrate Offersgaard et al. (2008) and O'Brien's (2002, 2010) insights into three sets of competences: (1) core competences, (2) linguistic skills and (3) instrumental competence. Core competences consist of, on the one hand, "the attitudinal or psycho-physiological competence that allows the post-editor to cope with subjectivity issues involved in defining and applying PE specifications" (p. 170) and, on the other hand, "the strategic competence that helps post-editors reach at informed decisions when choosing among different PE alternatives" (p. 170). The most recent PEC models are Nitzke et al.'s 2019 model and its refined version by Nitzke and Hansen-Schirra (2021), both based on PACTE's (2003) TC model and Robert et al.'s (2017) TRC model, which – according to Nitzke et al. (2019) – share some of the competences needed for post-editing machine translation output. The 2019 PEC model consists of four core competences and eight subsidiary subcompetences. The first core competence is the risk assessment competence, described by Nitzke et al. (2019) as "one of the most important competences a post-editor needs" and "the ability to assess the risk of the text to be translated" (p. 248). The second core competence is the strategic competence, based on risk assessment, which is the post-editor's ability to decide to apply either full or light PE for the translation task or to use only MT. The third core competence is the consulting competence, which is – depending on risk assessment and strategic decisions – the post-editor's ability to "inform the customer or project manager about potential risks as well as problem-solving strategies" (p. 248). Finally, the fourth core competence is the service competence. Additionally, Nitzke et al. (2019) list eight subsidiary subcompetences: (1) bilingual competence, (2) extralinguistic competence, (3) instrumental competence, (4) research competence, (5) revision competence, (6) translation competence, (7) machine translation competence and (8) post-editing competence. As in PACTE's (2003) and Robert et al.'s (2017) models, Nitzke et al. (2019) also include factors such as psycho-physiological components, post-editors' self-perception, the PE brief including guidelines for the PE task and affinity for ICT. It has to be noted that in Nitzke et al.'s (2019) model, translation competence, revision competence and post-editing competence are themselves considered subsidiary subcompetences, that is, they "support the core competences" (p. 249).

Regarding the revision subcompetence, the authors state that "the post-editor must handle the trade-off between necessary changes and over-editing, that is, to spot significant mistakes" (p. 249), which echoes the definition of the strategic subcompetence in Robert et al.'s model. In reference to the post-editing subcompetence, Nitzke et al. explain that errors in neural MT are harder to identify because the MT output is more fluent and correct. Consequently, the risk of overlooking mistakes is real and therefore post-editors must be trained in "spotting exactly these more fine-grained problems" (p. 250). In other words, although the structure of Nitzke et al.'s model is different from Rico and Torrejón's (2012) model, problem detection and solving are considered central too. This is also the case in the 2021 PEC model by Nitzke and Hansen-Schirra. PEC is represented as a "house of PE competences" (p. 69) whose architecture is grounded on translation competence (including bilingual, extralinguistic and research competence). In other words, it is expected that post-editors are skilled translators since "they need the same basic skill set" (p. 70). The house model further consists of three pillars defining three additional competences: error handling, MT engineering and consulting. Depending on the job profile and the specialisation of the post-editor, these three additional competences can play either a major or minor role. For example, when practical PE is at the core of the job profile, the main focus is on error handling, that is, error spotting (or 'problem detection'), error classification and error correction (that is, 'problem solving'). Next, the house model also includes a roof representing the soft skills for post-editors, such as risk assessment and service provision. Finally, psycho-physiological components, such as stress resistance or quick-wittedness, are also part of the model.

2.2. Measuring translation revision and post-editing competence

In search of possible TRC and PEC measuring instruments, we examined PACTE's TC research (Hurtado Albir, 2015; PACTE, 2011a, 2011b), in which indicators of strategic subcompetence are used to measure variables of translation competence. Consequently, we started from what Robert et al. (2017) also consider the central subcompetence of TRC, that is, strategic subcompetence, defined as follows:

> Procedural and conditional knowledge to guarantee the efficiency of the revision process and solve the problems encountered. [...]. Its functions are to (1) plan and carry out the revision task: selecting the most adequate procedure in view of the task definition, reading for evaluation, applying a detection strategy (anticipation and/or comparison), applying an immediate solution or problem-solving strategy, *making only the necessary changes* [emphasis added], taking the main revision principle into account; [...]. (p. 14)

As emphasized in the definition above, 'necessary changes' is key to TRC. However, necessary changes are not the only type of revision interventions. TR scholars, such as Brunette et al. (2005), Künzli (2005) and Robert and Van Waes (2014), have generally also distinguished between 'underrevisions' (failed necessary changes), 'hyperrevisions' (changes that do not make translations better or worse) and 'overrevisions' (changes that introduce errors into translations). This typology can be used to measure TR quality, which, in turn, can be used as a measuring instrument for or indicator of TRC. The first attempt in that respect dates back to the early 1980s. In 1983, Arthern, then head of the English translation division of the Council of the European Communities in Brussels, was required to write evaluation reports on his collaborators. Arthern (1983) defined the following categories of revision intervention: "Substantive error left or introduced (=X); formal error left or introduced (=F); unnecessary intervention (=U); necessary correction of sense or improvement in readability (=C)" (p. 55). After some trial and error, Arthern developed a mathematical formula to assign a score to each

reviewer: S (score) = X + F/2 + U/3. In 1991, aware of the risk of introducing subjectivity in distinguishing between substantial and formal errors, or between necessary and unnecessary interventions, Arthern decided to revisit his calculations, eliminating the U-category and combining the X- and F-categories: S (score) = X + F. In so doing, he re-ranked the same reviewers and found that the four best reviewers with the first calculation were the same as those with the second calculation. Similarly, the three worst reviewers were also virtually identical, regardless of calculation method.

In other words, when taking TR quality assessment as a measuring instrument for TRC, a first perspective is the type of TR intervention, for which counting the number of necessary changes is a first step. This is also what Robert and Van Waes (2014) did to measure TR quality, but they also looked at the number of underrevisions. The sum of necessary changes and of underrevisions was the indicator for what they called 'revision detection potential'. They included underrevisions, because these revision interventions showed that revisors had indeed detected errors, even though the errors remained in the translation.

Drawing on this first perspective, one can assess TR quality and consequently measure TRC, using three TRC measuring instruments or indicators, depending on the type of interventions considered:

1. number of necessary revisions only ('lenient revision quality score' below[iii])
2. number of necessary revisions minus (–) number of overrevisions (difference score, 'strict revision quality score' below)
3. number of necessary revisions plus (+) number of underrevisions (sum score, 'revision detection score' below)

However, as Robert and Van Waes (2014) argue, the question as what weight to assign to the different types of interventions remains. In other words, a second perspective would entail considering not only the *number* of interventions, but also the *impact* of each error on the quality of the revised translation. For example, in PE research, Daems and Macken (2021, p. 54) assign a severity weight to each problem or error. They use a severity weight of 3 for critical problems that have a major impact on the accuracy and/or intelligibility of translations; 2 for problems that cause a shift in meaning between source texts and target texts or affect the intelligibility of target texts; 1 for minor problems, where target texts can still be understood without effort and the information contained in them is equal to that of source texts, but there is a small error; and 0 for differences that are not actual problems (e.g., explicitations or omission of non-essential information). In the same vein, and this would entail the second perspective on TR quality assessment and thus on measuring TRC, one can therefore suggest a lenient and a strict *weighted* TR quality score as two additional TRC measuring instruments or indicators:

4. lenient *weighted* revision quality score:

$$\frac{\text{(number of necessary changes * weight)}}{\text{(number of errors * weight)}}$$

5. strict *weighted* revision quality score:

$$\frac{\text{(number of necessary changes * weight)} - \text{(number of overrevisions * weight)}}{\text{(number of errors * weight)}}$$

Besides, Daems and Macken's (2021) severity weights are reader-orientated. They represent the impact of the problem (e.g., omission of infrequent collocations) on readers. However, in didactic contexts, another approach can be adopted, where a different weight is attributed to errors, depending on the course contents and course-level intended learning outcomes. In our case, since the tasks were into the foreign language (L2 French), we knew from experience that some issues would not – or only exceptionally – be detected by students, because their L2 linguistic competence is not fully developed or because they do not have the necessary tools at their disposal. On the contrary, some issues that have only a minor impact on readers, such as many grammar mistakes, are, in fact, included in the intended learning outcomes of our curriculum and therefore should receive a higher weight. Consequently, TRC Measuring Instruments or Indicators 4 and 5 can be further divided into 4R, 4D, 5R and 5D, with 'R' standing for 'reader' and 'D' for 'didactic'. This is a third perspective.

Finally, and this is the fourth and last perspective, a measure of TRC can also be based on what is called the item-based assessment method (Bachman, 1990) in language testing or the 'rich points method' in TC research (PACTE, 2008). 'Items' in revision (see, for example, Robert & Van Waes, 2014) are generally specific text segments (words, expressions, sentences) that require revision to adhere to the revision brief provided. Consequently, working with items would mean that the number of items is considered for each of the seven measuring instruments described above.

As an answer to our first research question, we can therefore state that TRC can be measured in at least seven ways, which correspond to seven measuring instruments or indicators. It seems reasonable to apply the same reasoning to PEC as to TRC and to distinguish seven measuring instruments constructed from the same perspectives. This rationale is based on the common aspects shared by TRC and PEC models, such as problem detection and problem solving, which is in line with aspects for good PE performance described by de Almeida and O'Brien (2010):

> The ability to identify issues in the raw MT output that need to be addressed and to fix them appropriately. We call these "Essential Changes"; [...] The ability to adhere to the guidelines, so as to minimise the number of preferential changes, which are normally outside the scope of PE. We call these "Preferential Changes". (p. 2)

3. Methodology

To answer our second research question, we conducted a small-scale experimental pilot study with students (L1 Dutch, L2 French) in the Master's in Translation programme at the University of Antwerp. We worked with a convenience sample of 11 students,[iv] enrolled in the course *Dutch–French Translation and Revision*, a weekly 2-hour on-campus course (13 weeks, Semester 2, Academic Year 2018–2019). The course focuses equally on L1 Dutch–L2 French translation, TR and PE.

The study took place in June 2019, with informed consent from the students to collect necessary data. All students carried out three tasks (Dutch–French, approx. 300 words): (1) a TR task, (2) a PE task and (3) a translation task. We will not address the translation task in this paper, unless it is necessary to understand the context of the experiment. Students were provided clear instructions for all three tasks. They were free to complete the tasks in any order, but all students chose to work in the following order: TR, PE and translation. They were allowed to work for approximately four hours. Students had access to the same tools: *Le Grand Robert*

(monolingual French dictionary), *Van Dale* (bilingual Dutch–French dictionary) and *Antidote* (writing assistance software package, including language corrector, dictionaries with search tools and language guides, all directly integrated into MS Word, see https://www.antidote.info/en). No other tools were allowed to ensure the 'tools' variable remained constant for the three tasks. Product (MS Word files) and process data were collected using Inputlog 8.0.02 (https://www.inputlog.net, Leijten & Van Waes, 2013) for all three tasks.

Following our fourth perspective in measuring TRC, the TR task included 13 items that had been inserted by us into an existing published translation. The validation of the items was carried out by a renumerated independent external professional translator. The selected items represented different types of translation errors, which we labelled using a slightly adapted version of Mossop's typology of 14 grouped revision parameters (2020, pp. 136–137), the same typology as the one used by the students in class during the semester. Mossop's Parameters 10 and 12 (p. 137, Group D, *Problems with the Visual and Organizational Aspects of the Text*) and Parameters 13 and 14 (p. 137, Group E, *Problems Related to Specifications and Policies*) were not included.

For the PE task, the selection, that is, the identification of items was different. We translated the source text using DeepL and identified 18 items that we believed our students should be able to revise. This identification process was far from straightforward, since errors in neural MT are not easy to detect by L2 language users. We had already observed this phenomenon in our TR and PE classes at the University of Antwerp. We submitted the source text and target text for item validation to the professional translator, who also carried out item validation for the TR task. After a debriefing session, we retained 16 items, which were classified according to the same typology used for the TR task.

4. Results

Although our methodology is mainly product-based, with TR and PE quality assessment at its core, we also considered some process measures, in particular task time, which we measured with Inputlog.

4.1. Process data

Students worked approximately as long for the translation task as for the TR task, but less long for the PE task (Table 1).

Table 1

Summary Task Times: Absolute Comparisons (Hours:Minutes:Seconds, per Task) (N = 11, for Each Task)

Task	*M*	*SD*	Min	Max
Translation	1:28:23	0:16:45	0:44:07	1:45:55
Translation revision	1:27:35	0:12:30	0:57:33	1:45:30
Post-editing	0:53:06	0:10:42	0:39:41	1:17:42

Students were free to work on each task as long as they wanted. However, they were informed that they had a limit of four hours (10-minute tolerance margin). Although it is strange that students appear to revise as fast as they translate, one must not forget that the TR task was of a didactic nature. This meant that students had to justify all changes that they introduced. In other words, students had to add comments (balloons) for every change, mentioning Mossop's

parameters as seen in class. However, this was also the case for the PE task. The fact that students worked less long on the PE task might be due to poor time management. Since not all students worked for the full four hours, comparing total task times by means of (relative) percentages reflects actual task times more accurately (see Table 2).

Table 2

Summary Task Times: Relative Comparisons (Percentage of Task Times Combined) (N = 11, for Each Task)

Task	*M*	*SD*	Min	Max
Translation	38.4	5.6	26.6	45.8
Translation revision	38.3	4.6	29.8	48.0
Post-editing	23.3	4.2	17.2	32.2

Because of our small sample size (N=11), we conducted a non-parametric test (Friedman's ANOVA) to determine if the differences between task time percentages were statistically significant. Because the test was significant ($\chi^2(2) = 16.91$, $p < .01$), we followed up the result with three paired sample tests (Wilcoxon signed-rank tests) to determine where the difference was statistically significant. A Bonferroni correction[v] was applied and all effects are reported at a .017 level of significance. The difference in relative duration between the translation task and the revision task was not significant ($z = –.445$, $p > .05$), but the differences in duration between the translation task and the PE task and between the TR task and the PE task were significant ($z = –2.934$, $p = .001$; $z = –2.934$, $p = .001$; both 2-tailed).

Consequently, we decided that we would consider both the absolute and relative duration of each task to calculate the scores for each task. As we will show in Section 4.2, all scores, calculated as described in Section 2, will be reported as both time-*independent* scores (task duration not considered) and time-*dependent* scores (task duration considered). The time-dependent scores were calculated by dividing each score by the number of minutes devoted to the task and multiplying the resulting quotient by 60 to report a score per hour.

4.2. Product data

4.2.1. Translation revision task

Time-independent and time-dependent TR scores are summarized in Table 3.

Table 3

Descriptives Time-Independent and Time-Dependent Translation Revision (TR) Scores (N = 11)

	TRC indicator	TR score (%)			
		Time-independent		Time-dependent	
		M	*SD*	*M*	*SD*
1.	Lenient TR quality score	62.9	9.0	44.5	11.3
2.	Strict TR quality score	46.9	14.4	33.3	13.5
3.	TR detection score	66.4	8.6	46.8	10.5
4.	Lenient weighted R TR quality score	63.2	11.1	44.7	12.0
5.	Lenient weighted D TR quality score	69.7	11.5	49.2	12.3
6.	Strict weighted R TR quality score	52.4	13.5	37.2	13.1
7.	Strict weighted D TR quality score	56.8	15.8	40.3	14.6

The time-independent TR scores were normally distributed (statistically non-significant tests of normality). Therefore, we conducted a repeated measures (RM) parametric test (one-way ANOVA) to determine if the TR scores calculated according to all seven measuring instruments were significantly different. Mauchly's test of sphericity was significant ($\chi^2(20) = 55.94$, $p < .01$), so we corrected the degrees of freedom, using Greenhouse-Geisser estimates of sphericity; $\varepsilon = .37$. The within-subjects test was significant, $F(2.21, 22.21) = 18.55$, $p < .001$. Post-hoc pairwise comparisons revealed statistically significant differences for ten pairs (Appendix, Table 1). In other words, the measuring instrument or indicator is indeed important and can lead to statistically significantly different TR scores, which means that measuring instruments do not seem to be exchangeable. However, correlation tests indicate that each score shows significant positive correlations with all other scores (Appendix, Table 2), which could suggest that all time-independent TR scores seem to measure the same construct.

For the time-dependent TR scores, we adopted the same analysis as we did for the time-independent scores. As shown in Table 3, all time-dependent scores were lower than their corresponding time-*independent* scores. The series of time-dependent scores were normally distributed (statistically non-significant tests of normality). Therefore, we conducted a RM parametric test (one-way within-subjects ANOVA) to determine if TR scores calculated according to all seven measuring instruments were significantly different. Mauchly's test of sphericity was significant ($\chi^2(20) = 55.35$, $p < .001$), so the degrees of freedom were once again corrected using Greenhouse-Geisser estimates of sphericity; $\varepsilon = .37$). The within-subjects test was again significant, $F(2.24, 22.44) = 18.70$, $p < .001$. Post-hoc pairwise comparisons revealed statistically significant differences for nine pairs (Appendix, Table 3). In other words, the measuring instrument or indicator appears once again important and can lead to statistically significantly different TR scores, meaning that measuring instruments are again not exchangeable. However, correlation tests indicate that each score shows significant positive correlations with all other scores (Appendix, Table 2), which could mean that all time-dependent TR scores too seem to measure the same construct.

Finally, we conducted paired samples *t*-tests for each pair of scores (time-independent and time-dependent) to determine if differences between time-independent scores and their time-dependent counterparts were statistically significant. All tests were significant (Appendix, Table 4). In other words, task duration does indeed have an impact on TR scores.

4.2.2. Post-editing task

For the PE task, we adopted the same approach as we did for the TR task. Table 4 shows the time-independent and time-dependent scores for the PE task.

Table 4

Descriptives Time-Independent and Time-Dependent Post-Editing (PE) Scores (N = 11)

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

PEC indicator		PE score (%)			
		Time-independent		Time-dependent	
		M	*SD*	*M*	*SD*
1.	Lenient PE quality score	25.6	14.4	28.2	14.8
2.	Strict PE quality score	19.9	15.5	21.8	17.1
3.	PE detection score	48.3	12.5	55.3	12.2
4.	Lenient weighted R PE quality score	20.8	10.7	23.1	11.2
5.	Lenient weighted D PE quality score	22.9	11.8	25.4	12.3
6.	Strict weighted R PE quality score	16.0	11.9	17.7	13.3
7.	Strict weighted D PE quality score	16.5	13.0	18.2	14.7

We conducted a RM parametric test (one-way ANOVA) to determine if the differences were significant and we started with the time-independent scores. Mauchly's test of sphericity was significant ($\chi^2(20) = 108.61$, $p < .001$). Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity; $\varepsilon = .30$. The within-subjects test was significant, $F(1.82, 18.23) = 32.23$, $p < .001$. Post-hoc pairwise comparisons revealed statistically significant differences for six pairs (Appendix, Table 5). For the PE task, the difference was consistently significant between one particular score or indicator, that is the PE detection score, and all the other indicators. In other words, in this case, the measuring instrument did not seem to make a difference, except for the PE detection score, which was significantly different from all others. Correlation tests revealed significant positive correlations between each score and all other scores, except for the PE detection score (Appendix, Table 6), where it was not always the case (3 significant correlations, out of 6). In other words, in the case of PE, the measuring instrument seems to make less difference, except for the PE detection score.

Contrary to the TR scores, the time-dependent PE scores are higher than their time-independent counterparts (Table 4). This was to be expected, since participants generally worked longer for the TR task than the PE task. Again, we conducted a RM parametric test (one-way ANOVA) to determine if the differences were significant. Mauchly's test of sphericity was significant ($\chi^2(20) = 114.82$, $p < .001$). Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity; $\varepsilon = .28$. The within-subjects test was significant, $F(1.67, 16.70) = 28.16$, $p < .001$. Post-hoc pairwise comparisons revealed significant differences for six pairs (Appendix, Table 7). Again, the difference was consistently significant between the PE detection score and all the other scores. Correlation tests revealed significant positive correlations between each score and all other scores, except for the PE detection score (Appendix, Table 6), where there was no significant correlation at all. In other words, as observed above, the measuring instrument seems to make less difference, except for the PE detection score.

Finally, we conducted paired samples *t*-tests for each pair of scores (time-independent and time-dependent) to determine if the differences between time-independent scores and their time-dependent counterparts were statistically significant. No test was significant, except for the pair comparing the time-independent PE detection score with the time-dependent PE detection score (Appendix, Table 8). In other words, in the case of PE, taking the task duration into account does not change the results in statistically significant ways, except for the PE detection score.

5. Conclusions

The few existing TRC and PEC models seem to be built on the underlying hypothesis that these two competences are actually different but do share some common ground. Before conducting any empirical research into the degree of similarity or difference between TRC and PEC, the methodological issue of how to measure TRC and PEC must be addressed. In this paper, we have discussed four perspectives on how to measure TRC and PEC: the type of revision or post-editing intervention, the impact of errors to be revised or post-edited on the quality of the revised or post-edited translation, the context in which the competence is measured (i.e., didactic or professional) and the item-based or rich point assessment method. From these perspectives, we have proposed seven measuring instruments or indicators for TRC and PEC, while also considering the task duration, that is, the time taken to perform the tasks used as material to measure TRC and PEC.

Our preliminary analyses show that how TRC is measured generally leads to significantly different results: scores generated by different measuring instruments (time-independent *and* time-dependent) are significantly different in most cases. This is true not only within each group (i.e., among the time-independent scores and the time-dependent scores), but also between scores (i.e., between time-independent scores and their time-dependent counterparts). In other words, the TRC measuring instruments do not seem to be exchangeable, although their positive correlations seem to indicate that they measure the same construct. For PE, the results are slightly different: scores generated by different measuring instruments are not significantly different, except for one score, the PE detection score, which is systematically different from the others and measures what can be called 'post-editing detection potential', that is, the capacity to detect an error or problem in a machine translation without necessarily being able to solve it. Task duration does not play a significant role either, except for the PE detection score.

The instruments for measuring TRC and PEC that we propose in this paper are mainly product-based and quality-focused. Further research is needed to examine if, and if so, which process variables may be useful measuring instruments as well, in both didactic and professional settings. Although the present study is limited in scope, in terms of participants and tasks used as well as its specific focus on L1–L2 directionality, the findings provide valuable pointers for further research into TRC and PEC. First, it is paramount that scholars report in detail on the measuring instruments used, since different instruments may yield different results. Second, comparative research into TRC and PEC must be conducted, using instruments that measure the same variable in the same way. How different TC, TRC and PEC are remains to be seen, but the need for more empirical research into TRC and PEC is crystal-clear given the translator's evolving skill set.

6. References

Arthern, P. J. (1983). Judging the quality of revision. *Lebende Sprache, 2*, 53–57.

Arthern, P. J. (1991). Quality by numbers: Assessing revision and translation. In C. Picken (Ed.), *Fifth conference of the Institute of Translation and Interpreting* (pp. 85–94). Aslib.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bisaillon, J. (2007). Professional editing strategies used by six editors. *Written Communication*, *24*(4), 295–322. https://doi.org/10.1177/0741088307305977

Brunette, L., Gagnon, C., & Hine, J. (2005). The GREVIS Project: Revise or court calamity. *Across Languages and Cultures*, *6*(1), 29–45. https://doi.org/10.1556/Acr.6.2005.1.3

Chodkiewicz, M. (2020). *Understanding the development of translation competence*. Peter Lang. https://doi.org/10.3726/b17378

Daems, J., & Macken, L. (2021). Post-editing human translations and revising machine translations: Impact on efficiency and quality. In M. Koponen, B. Mossop, I. S. Robert, & G. Scocchera (Eds.), *Translation revision and post-editing: Industry practices and cognitive processes* (pp. 50–70). Routledge. https://doi.org/10.4324/9781003096962-5

de Almeida, G., & O'Brien, S. (2010, May 27–28). *Analysing post-editing performance: Correlations with years of translation experience* [Paper presentation]. 14th Annual Conference of the EAMT, S. Rafael, France.

EMT Expert Group. (2009). *Competences for professional translators, experts in multilingual and multimedia communication*. Retrieved from https://ec.europa.eu/info/sites/info/files/emt_competences_translators_en.pdf

EMT Expert Group. (2017). *EMT competence framework 2017*. Retrieved from https://ec.europa.eu/info/sites/info/files/emt_competence_fwk_2017_en_web.pdf

Göpferich, S. (2009). Towards a model of translation competence and its acquisition: The longitudinal study *TransComp*. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation process research* (pp. 11–37). Samfundslitteratur.

Hansen, G. (2009). The speck in your brother's eye – the beam in your own: Quality management in translation and revision. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (pp. 255–280). John Benjamins. https://doi.org/10.1075/btl.80.19han

Hurtado Albir, A. (2015). The acquisition of translation competence: Competences, tasks, and assessment in translator training. *Meta, 60*(2), 256–280. https://doi.org/10.7202/1032857ar

Hurtado Albir, A. (Ed.). (2017). *Researching translation competence by PACTE group*. John Benjamins. https://doi.org/10.1075/btl.127

Jakobsen, A. L. (2019). Moving translation, revision, and post-editing boundaries. In H. V. Dam, M. Nisbeth Brøgger, & K. Korning Zethsen (Eds.), *Moving boundaries in Translation Studies* (pp. 64–80). Routledge. https://doi.org/10.4324/9781315121871-5

Kelly, D. (2005). *A handbook for translator trainers: A guide to reflective practice*. St. Jerome.

Koponen, M., Mossop, B., Robert, I. S., & Scocchera, G. (Eds.). (2021). *Translation revision and post-editing: Industry practices and cognitive processes*. Routledge. https://doi.org/10.4324/9781003096962

Kornacki, M. (2018). *Computer-assisted translation (CAT) tools in the translator training process*. Peter Lang. https://doi.org/10.3726/b14783

Künzli, A. (2005). What principles guide translation revision?: A combined product and process study. In I. Kemble (Ed.), *Translation norms: What is 'normal' in the translation profession? Proceedings of the conference held on 13th November 2004 in Portsmouth* (pp. 31–43). University of Portsmouth, School of Languages and Area Studies.

Künzli, A. (2006). Teaching and learning translation revision: Some suggestions based on evidence from a think-aloud protocol study. In M. Garant (Ed.), *Current trends in translation teaching and learning* (pp. 9–24). Helsinki University.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze writing processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Loewen, S., & Plonsky, L. (2016). *An A–Z of applied linguistics research methods*. Palgrave. https://doi.org/10.1007/978-1-137-40322-3

Massey, G. (2017). Translation competence development and process-oriented pedagogy. In J. W. Schwieter & A. Ferreira (Eds), *The handbook of translation and cognition* (pp. 496–518). John Wiley & Sons. https://doi.org/10.1002/9781119241485.ch27

Mossop, B. (1992). Goals of a revision course. In C. Dollerup & A. Loddegaard (Eds.), *Teaching translation and interpreting: Training, talent and experience. Papers from the first Language International Conference Elsinore, Denmark, 31 May–2 June 1991* (pp. 81–90). John Benjamins. https://doi.org/10.1075/z.56.14mos

Mossop, B. (with Hong, J., & Teixeira, C.). (2020). *Revising and editing for translators* (4th ed.). Routledge. https://doi.org/10.4324/9781315158990

Nitzke, J., Hansen-Schirra, S., & Canfora, C. (2019). Risk management and post-editing competence. *JoSTrans: The Journal of Specialised Translation*, *31*, 239–259.

Nitzke, J. & Hansen-Schirra, S. (2021). *A short guide to post-editing* (Translation and Multilingual Natural Language Processing 16). Language Science Press. DOI: 10.5281/zenodo.5646896

Nunes Vieira, L., Alonso, E., & Bywood, L. (Eds.). (2019). Post-editing in practice: Process, product and networks. *JoSTrans: The Journal of Specialised Translation*, *31*.

O'Brien, S. (2002). Teaching post-editing: A proposal for course content. In *Proceedings of the Sixth EAMT Workshop: Teaching Machine Translation* (pp. 99–106). Retrieved from https://www.aclweb.org/anthology/2002.eamt-1.11.pdf

O'Brien, S. (2010, October 31–November 4). *Introduction to post-editing: Who, what, how and where to next?* [Paper presentation]. AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas, Denver, CO, United States. http://www.mt-archive.info/10/AMTA-2010-OBrien.pdf

Offersgaard, L., Povlsen, C., Almsten, L., & Maegaard, B. (2008, September 22–23). *Domain specific MT in use* [Paper presentation]. 12th EAMT conference, Hamburg, Germany.

PACTE. (2003). Building a translation competence model. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 43–66). John Benjamins. https://doi.org/10.1075/btl.45.06pac

PACTE. (2005). Investigating translation competence: Conceptual and methodological issues. *Meta*, *50*(2), 609–619. https://doi.org/10.7202/011004ar

PACTE. (2008). First results of a translation competence experiment: 'Knowledge of translation' and 'efficacy of the translation process'. In J. Kearns (Ed.), *Translator and interpreter training: Issues, methods and debates* (pp. 104–126). Continuum.

PACTE. (2011a). Results of the validation of the PACTE translation competence model: Translation project and Dynamic Translation Index. In S. O'Brien (Ed.), *Cognitive explorations of translation* (pp. 30–53). Continuum.

PACTE. (2011b). Results of the validation of the PACTE translation competence model: Translation problems and translation competence. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in Translation Studies* (pp. 317–341). John Benjamins. https://doi.org/10.1075/btl.94.22pac

PACTE. (2020). Translation competence acquisition: Design and results of the PACTE group's experimental research, *The Interpreter and Translator Trainer, 14*(2), 95–233. https://doi.org/10.1080/1750399X.2020.1732601

Pym, A. (2003). Redefining translation competence in an electronic age: In defence of a minimalist approach. *Meta*, *48*(4), 481–497. https://doi.org/10.7202/008533ar

Rico, C., & Torrejón, E. (2012). Skills and profile of the new role of the translator as MT post-editor. *Revista Tradumàtica: Tecnologies de la Traducció*, *10*, 166–178. https://doi.org/10.5565/rev/tradumatica.18

Robert, I. S., & Van Waes, L. (2014). Selecting a translation revision procedure: Do common sense and statistics agree? *Perspectives*, *22*(3), 304–320. https://doi.org/10.1080/0907676X.2013.871047

Robert, I. S., Remael, A., & Ureel, J. J. J. (2017). Towards a model of translation revision competence. *The Interpreter and Translator Trainer, 11*(1), 1–19. https://doi.org/10.1080/1750399X.2016.1198183

Robin, E. (2016). The translator as reviser. In I. Horváth (Ed.), *The modern translator and interpreter* (pp. 45–56). Eötvös University Press.

Scocchera, G. (2017). *La revisione nella traduzione editoriale dall'inglese all'italiano tra ricerca accademica, professione e formazione: Stato dell'arte e prospettive future*. Aracne editrice.

Shreve, G. M., Angelone, E., & Lacruz, I. (2018). Are expertise and translation competence the same? In R. Jääskeläinen & I. Lacruz (Eds.), *Innovation and expansion in translation process research* (pp. 37–54). John Benjamins. https://doi.org/10.1075/ata.18.03shr

Tiselius, E., & Hild, A. (2017). Expertise and competence in translation and interpreting. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 423–444). John Wiley & Sons. https://doi.org/10.1002/9781119241485.ch23

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

Appendix

Table 1

Translation revision (TR) scores (time-independent scores, pairwise comparisons)

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	16.084*	3.016	**0.007**	3.915	28.253
	3	–3.497	1.211	0.34	–8.384	1.391
	4	–0.266	1.884	1	–7.868	7.336
	5	–6.760*	1.65	**0.045**	–13.419	–0.101
	6	10.556	2.679	0.058	–0.253	21.365
	7	6.119	3.415	1	–7.658	19.896
2	3	–19.580*	3	**0.001**	–31.684	–7.477
	4	–16.350*	3.849	**0.036**	–31.879	–0.822
	5	–22.844*	3.707	**0.002**	–37.802	–7.886
	6	–5.528	2.443	0.989	–15.383	4.327
	7	–9.965*	2.16	**0.02**	–18.681	–1.249
3	4	3.23	2.423	1	–6.545	13.005
	5	–3.263	2.2	1	–12.139	5.612
	6	14.053*	2.815	**0.011**	2.696	25.409
	7	9.615	3.255	0.303	–3.519	22.749
4	5	–6.494*	0.944	**0.001**	–10.304	–2.683
	6	10.823*	2.233	**0.014**	1.814	19.831
	7	6.385	3.296	1	–6.912	19.683
5	6	17.316*	2.48	**0.001**	7.311	27.321
	7	12.879	3.298	0.062	–0.427	26.185
6	7	–4.437	1.465	0.267	–10.35	1.476

1 = Lenient TR quality score; 2 = Strict TR quality score; 3 = TR detection score; 4 = Lenient weighted R TR quality score; 5 = Lenient weighted D TR quality score; 6 = Strict weighted R TR quality score; 7 = Strict weighted D TR quality score

Based on estimated marginal means

* Mean difference significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Table 2

Pearson Correlations for Time-Independent and Time-Dependent Translation Revision (TR) Scores

TRC indicators	1	2	3	4	5	6	7
1. Lenient TR quality score	—	.86**	.97**	.93**	.95**	.88**	.84**
2. Strict TR quality score	.73**	—	.87**	.76**	.78**	.91**	.94**
3. TR detection score	.90**	.73**	—	.89**	.92**	.88**	.87**
4. Lenient weighted R TR quality score	.83**	.52*	.69**	—	.99**	.93**	.86**
5. Lenient weighted D TR quality score	.89**	.57*	.77**	.96**	—	.91**	.86**
6. Strict weighted R TR quality score	.76**	.83**	.73**	.84**	.79**	—	.98**
7. Strict weighted D TR quality score	.72**	.89**	.76**	.72**	.72**	.96**	—

Note. The results for the time-independent TR scores are shown below the diagonal. The results for the time-dependent TR scores are shown above the diagonal.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

Table 3

Translation Revision (TR) Scores (Time-Dependent Scores, Pairwise Comparisons)

(I) Factor2	(J) Factor2	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	11.180*	2.058	**0.006**	2.875	19.486
	3	–2.339	0.818	0.355	–5.638	0.96
	4	–0.209	1.359	1	–5.694	5.275
	5	–4.711	1.207	0.062	–9.581	0.16
	6	7.283	1.849	0.058	–0.179	14.744
	7	4.21	2.38	1	–5.392	13.812
2	3	–13.519*	2.05	**0.001**	–21.79	–5.249
	4	–11.389*	2.674	**0.035**	–22.178	–0.601
	5	–15.891*	2.579	**0.002**	–26.298	–5.484
	6	–3.897	1.736	1	–10.903	3.108
	7	–6.970*	1.558	**0.025**	–13.256	–0.685
3	4	2.13	1.639	1	–4.484	8.744
	5	–2.371	1.509	1	–8.461	3.718
	6	9.622*	1.886	**0.01**	2.013	17.231
	7	6.549	2.245	0.323	–2.509	15.607
4	5	–4.502*	0.642	**0.001**	–7.092	–1.911
	6	7.492*	1.495	**0.011**	1.46	13.524
	7	4.419	2.261	1	–4.706	13.543
5	6	11.994*	1.655	**0.001**	5.316	18.671
	7	8.921	2.257	0.057	–0.184	18.025
6	7	–3.073	1.022	0.277	–7.196	1.05

1 = Lenient TR quality score; 2 = Strict TR quality score; 3 = TR detection score; 4 = Lenient weighted R TR quality score; 5 = Lenient weighted D TR quality score; 6 = Strict weighted R TR quality score; 7 = Strict weighted D TR quality score

Based on estimated marginal means

* Mean difference significant at the .05 level.

Table 4

Pairwise Samples T-Tests (Time-Independent vs Time-Dependent TR scores)

Paired Samples Test								
	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1	18.43625	8.72746	2.63143	12.57306	24.29944	7.006	10	0.000
Pair 2	13.53264	8.03197	2.42173	8.13669	18.92858	5.588	10	0.000
Pair 3	19.59358	9.23063	2.78314	13.39236	25.79480	7.040	10	0.000
Pair 4	18.49353	8.68766	2.61943	12.65708	24.32997	7.060	10	0.000
Pair 5	20.48550	9.47064	2.85551	14.12304	26.84797	7.174	10	0.000
Pair 6	15.16301	8.04947	2.42701	9.75530	20.57072	6.248	10	0.000
Pair 7	16.52722	9.12213	2.75042	10.39890	22.65555	6.009	10	0.000

1 = Lenient TR quality score; 2 = Strict TR quality score; 3 = TR detection score; 4 = Lenient weighted R TR quality score; 5 = Lenient weighted D TR quality score; 6 = Strict weighted R TR quality score; 7 = Strict weighted D TR quality score

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

Table 5

Post-Editing Scores (Time-Independent Scores, Pairwise Comparisons)

(I) factor2	(J) factor2	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	5.682	1.968	0.34	–2.26	13.623
	3	–22.727*	3.293	**0.001**	–36.016	–9.439
	4	4.789	1.993	0.78	–3.253	12.831
	5	2.673	1.287	1	–2.522	7.867
	6	9.551	2.643	0.1	–1.114	20.216
	7	9.07	2.467	0.09	–0.884	19.024
2	3	–28.409*	4.72	**0.003**	–47.452	–9.366
	4	–0.893	2.89	1	–12.552	10.767
	5	–3.009	2.469	1	–12.971	6.953
	6	3.869	2.081	1	–4.527	12.266
	7	3.388	1.342	0.633	–2.026	8.802
3	4	27.516*	3.414	**0**	13.743	41.289
	5	25.400*	3.209	**0**	12.451	38.349
	6	32.278*	4.6	**0.001**	13.718	50.839
	7	31.797*	4.719	**0.001**	12.758	50.836
4	5	–2.116	0.748	0.376	–5.136	0.903
	6	4.762	1.816	0.536	–2.566	12.09
	7	4.281	2.337	1	–5.15	13.712
5	6	6.878	1.999	0.133	–1.187	14.944
	7	6.397	2.238	0.357	–2.635	15.429
6	7	–0.481	0.94	1	–4.272	3.31

1 = Lenient PE quality score; 2 = Strict PE quality score; 3 = PE detection score; 4 = Lenient weighted R PE quality score; 5 = Lenient weighted D PE quality score; 6 = Strict weighted R PE quality score; 7 = Strict weighted D PE quality score

Based on estimated marginal means

* Mean difference significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Table 6

Pearson Correlations for Time-Independent and Time-Dependent Post-Editing (PE) Scores

PE scoring method	1	2	3	4	5	6	7
1. Lenient PE quality score	—	.90**	.36	.88**	.96**	.77**	.82**
2. Strict PE quality score	.91**	—	.02	.77**	.84**	.90**	.97**
3. PE detection score	.68*	.39	—	.17	.28	–.16	–.09
4. Lenient weighted R PE quality score	.90**	.79**	.54*	—	.98**	.86**	.80**
5. Lenient weighted D PE quality score	.97**	.86**	.62*	.98**	—	.83**	.82**
6. Strict weighted R PE quality score	.79**	.91**	.22	.86**	.84**	—	.97**
7. Strict weighted D PE quality score	.83**	.97**	.25	.80**	.83**	.97**	—

Note. The results for the time-independent PE scores are shown below the diagonal. The results for the time-dependent PE scores are shown above the diagonal.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

Table 7

Post-Editing Scores (Time-Dependent Scores, Pairwise Comparisons)

(I) Factor2	(J) Factor2	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	6.472	2.304	0.389	–2.824	15.768
	3	–27.028*	4.635	0.003	–45.729	–8.326
	4	5.153	2.174	0.825	–3.62	13.925
	5	2.839	1.375	1	–2.707	8.386
	6	10.556	2.891	0.094	–1.111	22.222
	7	10.029	2.694	0.083	–0.843	20.9
2	3	–33.500*	6.26	0.007	–58.758	–8.241
	4	–1.319	3.33	1	–14.757	12.118
	5	–3.632	2.851	1	–15.135	7.87
	6	4.084	2.305	1	–5.218	13.386
	7	3.557	1.418	0.651	–2.164	9.277
3	4	32.181*	4.549	0.001	13.826	50.535
	5	29.867*	4.447	0.001	11.926	47.808
	6	37.584*	5.868	0.002	13.905	61.262
	7	37.056*	6.035	0.002	12.706	61.407
4	5	–2.313	0.855	0.463	–5.762	1.135
	6	5.403	2.076	0.554	–2.974	13.78
	7	4.876	2.695	1	–6	15.751
5	6	7.717	2.248	0.135	–1.354	16.787
	7	7.189	2.528	0.366	–3.01	17.389
7	7	–0.527	1.107	1	–4.994	3.939

1 = Lenient PE quality score; 2 = Strict PE quality score; 3 = PE detection score; 4 = Lenient weighted R PE quality score; 5 = Lenient weighted D PE quality score; 6 = Strict weighted R PE quality score; 7 = Strict weighted D PE quality score

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Table 8

Pairwise samples t-tests for time-independent versus time-dependent PE scores

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1	–2.65117	5.42225	1.63487	–6.29389	0.99155	–1.622	10	0.136
Pair 2	–1.86120	4.93203	1.48706	–5.17458	1.45219	–1.252	10	0.239
Pair 3	–6.95172	9.81013	2.95787	–13.54226	–0.36119	–2.350	10	0.041
Pair 4	–2.28743	4.36800	1.31700	–5.22189	0.64703	–1.737	10	0.113
Pair 5	–2.48444	4.80944	1.45010	–5.71546	0.74659	–1.713	10	0.117
Pair 6	–1.64619	3.98643	1.20196	–4.32431	1.03193	–1.370	10	0.201
Pair 7	–1.69247	4.25841	1.28396	–4.55331	1.16837	–1.318	10	0.217

1 = Lenient PE quality score; 2 = Strict PE quality score; 3 = PE detection score; 4 = Lenient weighted R PE quality score; 5 = Lenient weighted D PE quality score; 6 = Strict weighted R PE quality score; 7 = Strict weighted D PE quality score

This is an Accepted Manuscript (postprint). The Version of Record of this manuscript has been published and is available in Perspectives: https://doi.org/10.1080/0907676X.2022.2030377

[i] We are aware of the debate on the notion of competence in TS, especially in relation to expertise. We operationalize competence as "a pedagogical construct used to describe ideal skill/ability/knowledge sets for education and training purposes" (Shreve et al., 2018, p. 47).

[ii] With our first RQ, we follow PACTE's (2005) steps when they started investigating TC. One of their first objectives was indeed "to validate the TC measuring instruments" (p. 610).

[iii] In naming each measuring instrument or indicator, we opted for the term 'score' because, as we will explain in Section 3, the experiments in our pilot study included tasks.

[iv] 15 students were enrolled in the course *Dutch–French Translation and Revision* in 2018–2019. 11 students provided us with informed consent to use the data for the three tasks reported in this article. These tasks constituted their final evaluation for the course.

[v] The Bonferroni correction is a method commonly used in statistics to reduce the negative effects of conducting multiple statistical analyses on the same dataset (Loewen & Plonsky, 2016). When researchers conduct multiple analyses on the same dataset, they run the risk of falsely determining that statistically significant relationships exist, when, in fact, they do not. Such errors are known as Type I errors. To reduce Type I error rates, researchers will use a more conservative significance level, by dividing the commonly used alpha value .05 by the number of comparisons. The resulting conservative alpha value is then used to determine the significance of *p*-values against.