Developing and Validating EFL Learners' Perception of Machine Translation Questionnaire

Vahid R. Mirzaeian[*]
Assistant Professor of CALL, English Language Department, Faculty of Literature, Alzahra University
Tehran, Iran
Email: mirzaeian@alzahra.ac.ir

ABSTRACT

With huge improvements in the quality of MT in recent years, more interest has also been directed to its application in language learning; however, there is no standard questionnaire to assess EFL learners' perception of this technology. Therefore, this paper reports an attempt to describe the processes undertaken to validate a questionnaire developed for the first time to measure EFL students' perception of Machine Translation. The validation processes employed were face and content validity, factor analysis informed construct validity, reliability via internal consistency using test-retest reliability as well as Cronbach's Alpha correlation efficient. The exploratory factor analysis yielded four factors, namely, familiarity, use, fear of detection and the importance of training among students. The items on the questionnaire yielded factor loading ≥0.5. Reliability checking indicated that MT questionnaire was reliable. Cronbach's Alpha obtained was 0.94 for the whole questionnaire and between 0.80-0.91 for the four subscales. Test-retest statistic examination displayed stability of the responses at two time points eight weeks apart. The final questionnaire consisted of 29 items and the psychometric analysis indicated that it was both reliable and valid.

---

[*] Vahid R. Mirzaeian received his Ph.D. in CALL from UMIST, UK. His main areas of interest include: CALL, corpus linguistics and MTALL (Machine Translation Assisted Language Learning.

1. Introduction

With the advancement of information and communication technology (ICT), its impact is evident in all areas of life, including education. More and more institutes and educators are using this technology for the purpose of teaching. With the sudden spread of Covid-19, all educators and educational institutes were forced to use this technology. One such technology is Machine Translation (MT). As the name suggests, it is used to translate texts from one language into another; however, since it deals with aspects of two languages, it can also be used for language learning.

MT started around 1960s and a lot of resources were spent on it. However, it was a not considered a success for many reasons. The cost was very high and the output was unacceptable. Despite these problems, the pioneers in the field continued their efforts and many years later, acceptable output could be achieved if the input was subject to some restrictions, e.g. domain specificity. With the introduction of Neural Based Machine Translation (NBMT) by Google in 2016, there was a revolution in the quality of MT (Tu, Lu, Liu, Liu, & Li, 2016).

Since the quality of MT was not acceptable at the beginning, it was not used in the field of translation as well as other areas such as language learning. Many people translated some sentences with the help of MT and made jokes out of them. As a result, many language teachers rejected the idea of using MT for language learning (Anderson, 1995). During the same period, some researchers started using MT output for language learning, i.e. they asked students to input data into MT, get the output and correct the mistakes in the output technically referred to as post-editing (Nino, 2004).

If research on using MT for language learning is reviewed, mixed results will be obtained. Some believe that it can be beneficial for language learning; some claim it is fruitless or even harmful; and some remain cautious. Since 2016, a great revolution occurred in the field of MT and massive improvements were achieved in the quality of MT and it seems advisable to propose more research to be carried out in the field to shed more lights on the use of MT for language learning. Since questionnaires are useful tools enabling researchers to conduct studies and understand more about different aspects, the authors of this article decided to conduct studies on the issue of the application of MT in language teaching. They could not find any standard questionnaire; as a result, they decided to develop one. This paper; therefore, reports an attempt to develop and validate a questionnaire specifically prepared to be used by researchers to know more about perceptions of EFL students regarding MT.

2. Review of Literature

To show the significance of research in the application of MT to language learning, first we review studies conducted without a questionnaire; next, we will discuss those conducted with questionnaires.

Studies without questionnaires

Research in the area of the application of MT to language learning can be traced back to 2008 when a researcher (Nino, 2008) evaluated the implementation of MT post-editing in foreign language teaching. Her results showed that post-editing activity was suitable for advanced

learners. Errors in the MT output were also found to be similar to student human translation. She continued her research next year and examined the perception of foreign language students about MT and came to the conclusion that MT could be used as a good tool for language learning. The use and attitudes of students and teachers regarding MT has also been investigated recently (Cetiner & Isiag, 2019). Based on the results, the researcher proposed a framework to develop best practices for using MT in a language learning context. He also suggested, that the students and teachers should receive training in relation to their potential educational applications.

Some scholars focused on the level of students and wanted to know if MT could be of more assistance to elementary students, for example. As a result, a researcher (Garcia & Pena, 2011) asked his elementary language learners to write in their native language and translate the writings into English using MT. He found that foreign language students communicated more while using MT. It was also found that MT was more beneficial to weaker students; it helped them to implement more vocabulary items and as a result communicate more effectively. He came to the conclusion that writing directly in the target language requires more effort on the part of the students in a foreign language. He advised instructors to implement MT at elementary levels of language instruction. Regarding the types of errors found in MT output, a researcher (Putri & Ardi 2015) identified errors in MT output when translating Indonesian folklore into English. He concluded that passages containing such errors could be used as a didactic medium to promote critical thinking among language learners regarding foreign language use. As for the frequency of errors found in MT

output, Napitupulu examined this issue in the translation of abstracts by MT using a model of error analysis (Keshavarz, 2011). Five types of errors were identified and the frequency of each type of error reported. It was found that MT's output was full of errors and was therefore not recommended for use in language learning environments (Napitupulu, 2017).

Another avenue of research in this field is related to investigating how faulty MT output could be used to create activities that helped raise awareness of grammatical differences between the two languages (Enkin & Mejias-Bikandi, 2016). Some also examined the use of MT as a complementary tool to help international students learn and develop their skills and vocabulary in learning a language (Bahri & Mahadi, 2016). The results suggested that most participants recognized MT as an effective tool for learning vocabulary, writing and reading.

Regarding the use of MT in English for certain purposes (ESP), a study showed that MT could affect the quality and quantity of student foreign language writing. The results showed that when using MT, students wrote more words, implemented more vocabulary and longer sentences (Kol, Schcolnik, & Spector-Cohen, 2018). It was concluded that MT was a useful tool for EAP students (English for Academic Purposes), if they were able to critically evaluate and correct MT output.

The use of MT as a CALL tool in EFL classes was also investigated (Lee, 2019). Students were instructed to write in their mother tongue and translate them once without MT and again with MT into a foreign language. Having compared the results, the researchers noticed improvements in vocabulary. It was also

644

concluded that MT helped students to focus on writing as a process. Interviews and reflection papers showed that the students rated the use of MT positively while writing.

There were studies showing MT was detrimental to language learning. Harris believed that the use of MT by language learners meant less commitment to the target language and less chance of learning it. He suggested that language teachers should discourage the use of MT among language learners (Harris, 2010).

Darancik studied the quality of MT when translating from German into Turkish and from Turkish into German. The analysis was based on word, syntax, semantics and grammar. The research concluded that MT could make students lazy. It has been suggested that language learners should be aware of the errors caused by MT systems (Darancik, 2016). Another survey showed various post editing issues in language learning related to learning orientation, ethics and translator status (Dongyun, 2017). It was concluded that research on the effect of post-editing in MT was immature and inadequate. It has been confirmed that potential MT effects on language learners deserve more attention. Instead of avoiding MT completely, it was suggested that language teachers learn more about its weaknesses to take proactive countermeasures.

And Finally, online translator, dictionary and search engine use among L2 students were investigated comparatively (O'Neill, 2019) and the researcher concluded that online translators were the most popular technological tool among students. In another similar study, he trained his students to use online translators and dictionaries and measured the impact on second language writing scores. The data revealed that again online translators had a greater effect on students' writing scores compared with online dictionaries.

Studies with the help of a questionnaire

Compared to the studies conducted without questionnaires, studies with the help of questionnaire are not as varied. To begin with, questionnaires were first used to examine students' attitudes towards MT while using MT to improve vocabulary, writing, and reading by other researchers (Alhaisoni & Alhaysony, 2017), (Jimenez-Crespo, 2018). It was discovered that MT was useful in helping students to understand the meaning of unknown words, improving writing skills and reading English textbooks.

Students' intentions to use MT, as well as their attitudes and behavior regarding the use of MT for language learning, were also examined by another researcher using a questionnaire (Sukkhwan, 2014). In addition to the questionnaire, a checklist and a translation task were also used. All students used MT; however, the frequency to find the meaning of unknown words was the lowest, followed by assignments of writing, and reading. Overall, the students showed a positive attitude towards MT.

A different methodology, that is, the design of a pre-test post-test questionnaire, was implemented to examine students' attitudes toward MT (Cetiner, 2018). The results showed that students developed a positive attitude toward MT having being exposed to it in a training course. In another similar study, the questionnaire results given to instructors showed that MT use by students was considered fraudulent (Case, 2015). Research concluded, however, that students had to use MT but teachers had to teach them how to use it.

As it can be seen, there is no consistent view

of using MT for language learning. In addition, all the questionnaires used were researcher-made and lacked the validity and reliability which are two essential components of any instrument. Therefore, since further studies are needed in this area and questionnaires play important roles in this regard, the authors of this article decided to develop the first standard questionnaire regarding the perceptions of MT by EFL language learners.

3. Method

The method employed to validate the questionnaire included translational validity via face validity and content validity; factor analysis informed construct validity; reliability checking through internal consistency (Cronbach's Alpha) as well as test-retest as shown in Figure 1.
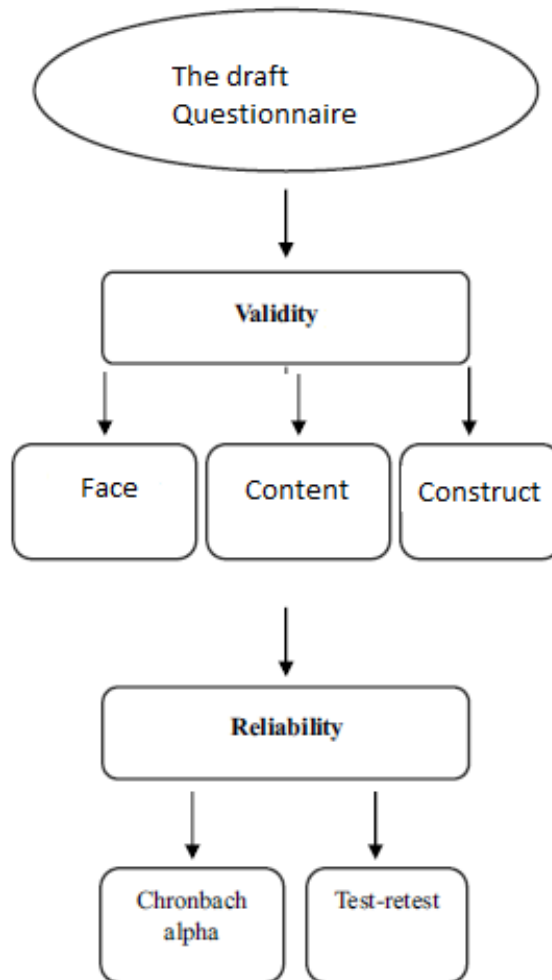
Figure 1. Machine translation questionnaire validation flow chart

The draft questionnaire was derived from the relevant literature and four existing questionnaires, namely, ICT familiarity questionnaire (Lorenceau, Marec, & Mostafa, 2020), digital citizenship questionnaire (Nordin et al., 2016), and digital literacy questionnaire (Holt, Overgaard, & Engel, 2020). The initial

646

draft of MT consisted of 35 items in five subsections, namely, familiarity, use, fear of detection, training, and open-ended questions.

Face validity

To determine the face validity for our questionnaire, an evaluation form was prepared to ease the process of assessing each item by

respondents in terms of the clarity of wording, the likelihood ability of the target audience to answer the questions, style and layout. Twenty-five EFL students in a state university in Tehran were randomly selected and completed face validity form on a Likert scale of 1-5, namely, 1= strongly disagree, 2 = slightly disagree, 3 = slightly agree 3= agree, 5= strongly agree.

Content validity

Content validity analysis was conducted to insure the appropriateness of the questionnaire content as well as relevancy to study purposes. To calculate the content validity of MT questionnaire, conceptual framework for the questionnaire had to be defined by thoroughly reviewing the literature as well as seeking expert advice. Once the conceptual framework was established, eight experts in the areas of ICT, digital literacy, digital citizenship, MT, language teaching, psychometrics, Computer Assisted Language Learning (CALL) and Data Driven Learning (DDL) were invited to review the draft 35 item questionnaire to ensure its consistency with the conceptual framework. Each expert independently rated item relevancy of questions based on a 5 point Likert scale (1 = no relevancy, 2 = minimum relevancy, 3 = relevancy, 4= maximum relevancy, 5= absolute relevancy). The validity of the items was estimated using Content Validity Index (Lynn, 1996).

Construct validity

The sampling population used for factor analysis was 160 EFL students from the general population of ELT students in Tehran using snowball sampling technique. To make sure a proportionate sample size has been implemented for this questionnaire to allow performing Exploratory Factor Analysis (EFA), two issues were taken into consideration, namely, Kaiser-Meyer-Olkin (KMO) sampling adequacy and Factor loadings using correlation between a factor and a variable (Hayes, 2002). The most commonly used orthogonal rotation (varimax) was implemented to rotate the factors so that the loading on each variable was maximized and the loading on other factors was minimized (Bryman & Cramer, 2005; Field, 2005).

Reliability

Having completed validity procedures, the researchers examined MT questionnaire to evaluate its reliability (Haladyna, 1999). Two estimators of reliability are commonly implemented: test-retest reliability and internal consistency reliability both of which were used in this study to evaluate the reliability of MT questionnaire.

Test-retest reliability

The test-retest reliability of our questionnaire was conducted by administering it to 25 EFL learners aged 18-28 randomly selected from a state university in the capital, Tehran. The questionnaire was completed on two different occasions with an eight-week interval. Since ordinal data were obtained from the questionnaire using a five Likert scale rated from very low to very high and the scale was not continuous, non-parametric statistical tests were preferred to Pearson Correlation coefficient. As a result, to determine if there were significant differences between responses at each time point, Wilcoxon Non-Parametric Statistical Test was implemented to analyze rsponsess between test and retest.

Internal consistency reliability

To evaluate the internal consistency of MT questionnaire, Cronbach's alpha was calculated.

If an instrument has more than one subscale, Cronbach's alpha has to be computed for each subscale and the whole scale separately (Nunnally & Bernstein, 1994). As a result, it was computed for each of our subscales.

4. Results

Face validity

All participants evaluated each question at three or four on a Likert scale of 1-4. 95% indicated the questions were comprehensible and easy to answer and 90% agreed with the layout and appearance of the questionnaire.

Content validity

Based on CVI index, a score of 3 or 4 indicates the validity of content as well as its consistency with the conceptual framework (Lynn, 1996). For instance, if an item is rated as relevant by 5 out of 8 experts, the CVI would be $5/8 = 0.62$ not meeting the required level of 0.87 (7/8) indicating the item should be eliminated (Devon et al., 2007). As a results, 3 items on draft questionnaire were considered invalid since they yielded CVIs of $5/8 = 0.62$ to $6/8 = 0.75$ and therefore deleted from the questionnaire. All the remaining items were valid since they enjoyed CVI ranging from 0.87 to 0.100.

Construct validity

The KMO sampling adequacy of the questionnaire was 0.9 ensuring that we had an appropriate sample size; therefore, we were able to undertake the EFA. A researcher (Kaiser, 1974) suggested accepting values ≥0.5 and described values between 0.5 and 0.7 as mediocre; 0.7 and 0.8 as good; 0.8 and 0.9 as great, and > 0.9 as superb. Therefore, based on Kaiser's scale, the sample adequacy of 0.9 for our questionnaire was considered superb. Likewise, another researcher (Stevens, 2002) suggested that if a factor has 10 or more variables with loadings of 0.4 and ≥150 participants, it is considered reliable. The sample size of 160 was considered adequate and allowed us to undertake EFA since the KMO of the first analysis of the draft questionnaire was 0.9.

On the first run PCA, the total variance of the draft questionnaire was 66.14% meaning it was considered to be reasonable since at least 50% of the variance could be explained by common factors (Field, 2005). The commonalities of the items on the questionnaire were >0.5. In the first run PCA, six factors had eigenvalues ≥1.00 when Kaiser's criterion was applied to the draft questionnaire. The complied scree plot on the first PCA indicated there were two to five factors. The scree plot and eigenvalues are considered accurate enough to determine the number of factors to be retained when variance of variables is ≥0.6 and the sample is ≥250 or when variance of variables is ≥0.7 and there are more than 30 variables. The scree plot by SPSS is given in Figure 2.
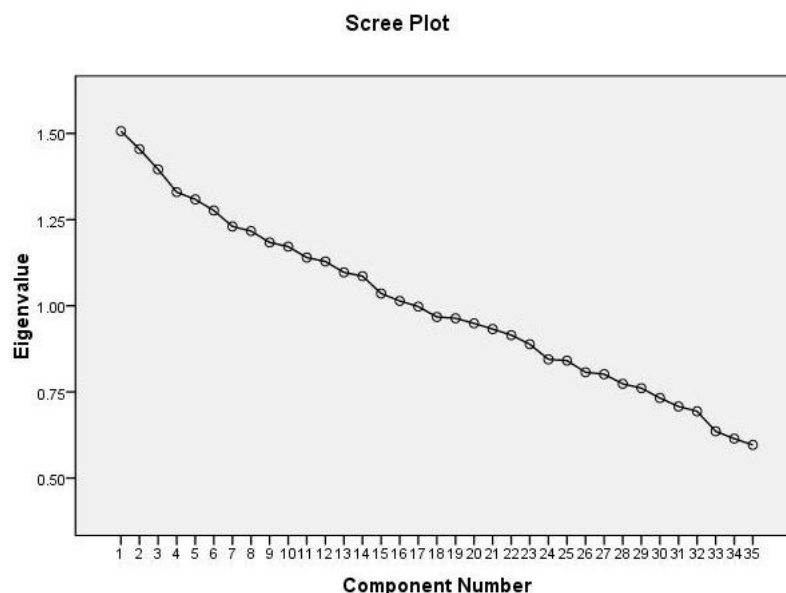
648

Figure 2. SPSS generated scree plot

Therefore, a four factor solution with Varimax rotation was considered suitable both conceptually and statistically after examining two to six factor solutions. In order to conduct the most appropriate interpretation, using guidelines for practical significance (Hair, Anderson, Tatham, & Black, 1998) the loading values were carefully examined. They indicated a factor loading of ±0.3 meaning the items were of minimal significance, ±0.4 meaning the items were important and ±0.5 indicating the significance of factor.

Based on the above tests, when the factor loading was ≤0.5, items were eliminated from the factor pattern matrix of the questionnaire. Based on guidelines by a researcher (Stevens, 2002), the decision to eliminate these items was confirmed. His guideline is based on sample size suggesting that the statistically acceptable loading for 50 participants is 0.72, for 100 participants 0.51, and for more than 200 participants 0.29-0.38. since the sample size used in our validation process was 160, three items with a loading <0.5 were eliminated and the remaining items with a loading of ≥0.5 were retained. One remaining item had a loading of 0.74; however, it was retained due to its importance to the relevant factors. The final PCA of the four-factor solution with twenty-nine times accounted for 62.17% of variance. The results for the final four factor solution of the questionnaire based on PCA is given in Table 1.

Table 1. The results for the final four factor solution of the MT questionnaire based on PCA

	Item	Factor 1	Factor 2	Factor 3	Factor 4
	MT familiarity (α=0.91)	Loadings			
1	I am familiar with digital technology	0.84			
2	I am familiar with the Internet	0.83			
3	I am familiar with an operating system	0.81			
4	I know how to use digital devices	0.80			
5	I know how to use my phone or computer	0.69			
6	I know what machine translation means	0.64			
7	I know different types of machine translation systems	0.62			

8	I know Google Translate	0.61			
9	I am happy with machine translation output	0.58			
10	I have a machine translation app on my phone	0.55			
	MT use (α=0.91)		Loadings		
1	I use machine translation a lot.		0.82		
2	I use machine translation to translate from English to my mother tongue.		0.82		
3	I use translation to translate from my mother tongue to English		0.81		
4	I use MT on my device such as phone, tablet, etc.		0.80		
	Fear of MT detection (α=0.80)			Loadings	
1	I use MT a lot to do my homework			0.77	
2	My teacher told me not to use MT			0.67	
3	My teacher dislikes my using MT			0.65	
4	I do not want the teacher to know I have used MT			0.57	
5	I do things so that my teacher does not realize I have used MT			0.55	
6	MT use is against our institute's regulations			0.47	
	Importance of MT training (α=0.89)				Loadings
1	I should be trained to use machine translation effectively				0.74
2	My teacher should teach me how to use machine translation				0.69
3	My university/institute should conduct workshops on how to use MT.				0.66
4	Teachers should also be taught how to use machine translation				0.64
5	Machine translation improves learning in general if I receive training				0.57
6	Machine translation improves language learning if I receive training				0.55
7	Since I was not trained, I could not use MT correctly				0.55
8	I underestimated the power of MT because I had no training				0.55
9	I overestimated the power of MT because I had no training				0.50

Reliability

Test-retest

Twenty EFL learners completed the questionnaire in test and retest with an eight-week interval and Wilcoxon non-parametric test showed no significant differences between the two tests. The final MT questionnaire included four subscales, namely, familiarity, use, fear of detection and the importance of training. The first subscale (familiarity) accounted for 37.11% of the total variance. This factor included ten items and reflected information about participants' familiarity with MT. The second subscale (use) accounted for 13.03% of total variance including four items with very high factor loadings ranging from 0.79 to 0.82. These items referred to participants' use of MT technology. The third subscale (fear of MT detection) accounted for 6.31% of the total variance and included six items. It focused on students' perception of MT use detection by their instructors. And the last subscale (the necessity of MT training) accounted for 5.71% of the total variance and included nine items. The items were rated on a Likert scale of 1-5 where 1 = very low; 2 = low; 3 = medium; 4 = high and 5 = very high.

internal consistency reliability

Construct validation for the questionnaire was found to be 0.94 indicating consistent

reliability between the items and the questionnaire. Based on that, Cronbach's alpha was calculated for the revised questionnaire. The alpha computed for each subscale exceed the minimum value of 0.70. Table 2 shows the results of MT questionnaire calculated via Wilcoxon's non-parametric test.

Table 2. Test-Retest results of MT questionnaire using Wilcoxon's non-parametric test.

No	Items	P value
1	*I am familiar with digital technology*	0.157
2	*I am familiar with the Internet*	0.480
3	*I am familiar with an operating system*	0.083
4	*I know how to use digital devices*	0.705
5	*I know how to use my phone or computer*	0.655
6	*I know what machine translation means*	0.180
7	*I know different types of machine translation systems*	0.705
8	*I know Google Translate*	0.180
9	*I am happy with machine translation output*	0.157
10	*I have a machine translation app on my phone*	1.000
11	*I use machine translation a lot.*	0.414
12	*I use machine translation to translate from English to my mother tongue.*	1.000
13	*I use translation to translate from my mother tongue to English*	0.317
14	*I use MT on my device such as phone, tablet, etc.*	0.180
15	*I use MT a lot to do my homework*	0.234
16	*My teacher told me not to use MT*	0.655
17	*My teacher dislikes my using MT*	0.655
18	*I do not want the teacher to know I have used MT*	0.317
19	*I do things so that my teacher does not realize I have used MT*	0.096
20	*MT use is against our institute's regulations*	0.157
21	*I should be trained to use machine translation effectively*	0.763
22	*My teacher should teach me how to use machine translation*	0.157
23	*My university/institute should conduct workshops on how to use MT.*	0.564
24	*Teachers should also be taught how to use machine translation*	0.564
25	*Machine translation improves learning in general if I receive training*	0.564
26	*Machine translation improves language learning if I receive training*	0.157
27	*Since I was not trained, I could not use MT correctly*	1.000
28	*I underestimated the power of MT because I had no training*	0.783
29	*I overestimated the power of MT because I had no training*	0.180

5. Discussion

When exploring complex phenomena in the humanities such as perception, the integrity of the research is highly dependent upon the accuracy of the measures used. Two essential concepts ensuring this accuracy are validity and reliability both of which have been taken into consideration while developing this questionnaire.

Face validity ensures questionnaire's

appropriateness to the content area as well as study purpose. It is the weakest form of validity; however, it is the easiest to conduct. It evaluates the appearance of the questionnaire in terms of the clarity of the language used, consistency of style and formatting, readability and feasibility. While face validity is considered the lowest form of validity, checking it was useful since it yielded important information regarding the operationalization of the questionnaire by EFL learners.

Content validity ensures that the content reflects a comprehensive range of features and is usually performed by seven or more experts (Pilot & Hunger, 1999). To calculate the content validity of MT questionnaire, conceptual framework for the questionnaire had to be defined by thoroughly reviewing the literature as well as seeking expert advice. The results of the validity testing of the MT questionnaire revealed that it was an accurate measure of assessing students' perception of MT. The processes used to validate the questionaries' were rigorous but appropriate.

To ensure construct validity, factor analysis is commonly used during questionnaire development to cluster items into common factors enabling interpretation of each factor based on items having the maximum loading on it and finally summarizing the items into minimum number of factors (Bryman & Cramer, 1999). Loading refers to the degree of association between a factor and an item (Bryman & Cramer, 2005). A factor is a list of items belonging together in which related ones define the part of the construct grouped together. Unrelated items do not define any construct and therefore should be eliminated (Munro, 2005).

Exploratory Factor Analysis (EFA) is a common method employed to investigate the relationship among variables without employing a specific hypothetical model (Bryman & Cramer, 2005). EFA helps researchers define each construct base on the theoretical framework indicating the measure direction (Devon et al., 2007) as well as identifying the greatest variance in scores with smallest number of factors (Delaney, 2005).

It is necessary to have a large sample so that EFA can be conducted reliably (Bryman & Cramer, 2005). Although there is no census among researchers as to the number of participants for a reliable EFA, a minimum of five participants for each variable is usually recommended (Munro, 2005).

Factor analysis assessed the theoretical construct of our questions. Many different types of extraction methods are employed to conduct EFA. The two most common methods are Principal Axis Factoring (PAF) and Principal Component Analysis (PCA) (Bryman & Cramer, 2005). In PAF only common variance is analyzed; however, in PCA, total variance of a variable will be analyzed (Bryman & Cramer, 2005). Total variance consists of both common and specific variances. Specific variance refers to the specific variation within a variable; however, common variance describes the variance shared among the scores of subjects with other variables (Bryman & Cramer, 2005). As a result, PCA seems to be perfectly reliable and error-free (Bryman & Cramer, 2005) and therefore has been selected for the 32 items MT questionnaire.

Based on two researchers (Bryman & Cramer, 2005), the two essential criteria to specify how many factors should be retained are

652

a) A scree Plot depicting the descending variances accounting for factors obtained from graph form. These are the factors lying before the point at which eigenvalues begin to drop; therefore, they can be retained and b) The Kaiser criterion for selection of factors having an eigenvalue ≥1. However, the common criterion of an eigenvalue ≥ 1.00 may misrepresent the maximum number of relevant factors (Heppner, Heppner, Lee, Wang, & Park, 2006).

Reliability of a questionnaire refers to the ability of a questionnaire to measure a feature consistently and to ensure that items fit together conceptually (Haladyna, 1999). Although ensuring that a questionnaire is reliable is considered vital; however, it is not sufficient since a questionnaire may be reliable but lack validity (Beanland, Scheneideer, Lobiondo-Wood, & Haber, 1999). Two researchers (Cronbach & Shavelson, 2004) suggested that issues such as how the instrument is to be used, content heterogeneity, sampling independence and standard error of instrument should be considered while determining the reliability of the questionnaire.

Two estimators of reliability are commonly implemented: test-retest reliability and internal consistency reliability both of which were used in this study to evaluate the reliability of MT questionnaire.

Test-retest reliability is usually conducted by administering the same instrument to the same population on two different occasions. This is based on the assumption that there should be no substantial change in the construct between the two sampling time points (Trochim, 2001). The duration of the time between the two sets is highly crucial since if the interval is short, the correlation will be higher and if the interval is too long, the correlation will be lower (Trochim, 2001). Since changes take place in participants or the environment, the very long test intervals will definitely affect the results (Linn & Gronlund, 2000). There is no definitive standard for the best time interval between the test and the retest; however, factors such as learning and exposure to the new technology in our case are essential and should be taken into consideration (Concidine, Botti, & Thomas, 2005).

Internal consistency evaluates inter-item correlation of an instrument and shows how well items are conceptually linked together (Nunnally & Bernstein, 1994). In addition, to estimate the consistency of the whole questionnaire, total score of all items is also calculated. Internal consistency can be measured in two ways: Cronbach's alpha correlation coefficient and Split-Half reliability (Trochim, 2001). Cronbach's alpha is the most frequently used statistic to determine internal consistency reliability; however, in Split-Half, all items measuring the same construct are divided into two sets and the correlation between the two is calculated (Trochim, 2001).

To evaluate the internal consistency of MT questionnaire, Cronbach's alpha was calculated.

If an instrument has more than one subscale, Cronbach's alpha has to be computed for each subscale and the whole scale separately {Nunnaly, 1994. As a result, it was computed for each of our subscales.

6. Conclusion

While MT has recently gained popularity due to its rapid output quality improvement in 2016, educators may find it difficult to measure

what they assess when they ask their students about MT largely because it is highly subjective and the educators may not have a clear understanding of MT. This paper reported psychometric validation of MT questionnaire to measure EFL students' perception according to a specific definition and context. However, to strengthen the rigor of the questionnaire for further research, it is advisable to undertake discriminant and convergent validity to examine the similarity and differences of the MT questionnaire with other similar tools such as computer familiarity questionnaires, ICT familiarity questionnaires, digital citizenship questionnaires and digital literacy questionnaires. It is also recommended that confirmatory factor analysis and structured equation modelling be undertaken in a larger sample with diverse EFL learners to support generalizability of the questionnaire. The MT questionnaire is both reliable and valid and can be generalized to a wider population of EFL learners in a variety of educational settings.

References

Alhaisoni, E., & Alhaysony, M. (2017). An Investigation of Saudi EFL University Students' Attitudes towards the Use of Google Translate. *International Journal of English Language Education, 5*(1), 72-82.

Anderson, D. (1995). Machine Translation as a Tool in Second Language Learning. *CALICO, 13*(1), 68-97. doi:http://www.jstor.org/stable/24147744

Bahri, H., & Mahadi, T. (2016). Google Translate as a Supplementary Tool for Learning Malay: A Case Study at Universiti Sains Malaysia *Advances in Language and Literary Studies, 7*(3), 161-167. doi:10.7575/aiac.alls.v.7n.3p.161

Beanland, C., Scheneideer, Z., Lobiondo-Wood, G., & Haber, J. (1999). *Nursing Research: methods, critical appraisal and utilisation.* Sydney: Mosby.

Bryman, A., & Cramer, D. (1999). *Quantitative Data Analysis*

with SPSS Release 8 for Windows. A Guide for Social Scientists. London: Routledge.

Bryman, A., & Cramer, D. (2005). *Quantitative Data Analysis*

with SPSS12 and 13. A Guide for Social Scientists: East Sussex Routledge.

Case, M. (2015). Machine Translation and the Disruption of Foreign Language Learning Activities. *eLearning Papers 45*, 4-16.

Cetiner, C. (2018). Analyzing the attitudes of translation students towards cat (computer-aided translation) tools. *Jounral of Language and Linguistic Studies, 14*(1), 153-161.

Cetiner, C., & Isiag, K. (2019). Undergraduate Level Translation Students' Attitudes towards Machine Translation Post-Editing Training. *International Journal of Languages' Education and Teaching, 7*(1), 110-120. doi:22.01.2019 20.02.2019 25.03.2019 10.18298/ijlet.3242

Concidine, J., Botti, M., & Thomas, S. (2005). Design, format,

validity and reliability of multiple choice questions for use in nursing

research and education. *Collegian, 12*(1), 19-24.

Cronbach, L., & Shavelson, R. (2004). My Current Thoughts

on Coefficient Alpha and Successor Procedures *Educational and Psychological Measurement, 64*(3), 391-418.

Darancik, Y. (2016). The Effect of Data-Based Translation Program Used in Foreign Language Education on the Correct Use of Language. *TOJET: The Turkish Online Journal of Educational Technology 15*(4), 88-106.

Delaney, C. (2005). The Spirituality Scale, Development and

Psychometric Testing of a Holistic Instrument to Assess the Human Spiritual

Dimension. *Journal of Holistic Nursing, 33*(2), 145-167.

Devon, H., Block, M., Moyle-Wright, P., Ernest, D., Hayden, S., & Lazzara, D. (2007). A psychometric Toolbox for testing

Validity and Reliability. *Journal of Nursing scholarship, 39*(2), 155-164.

Dongyun, S. (2017). Application of Post-Editing in Foreign Language Teaching: Problems and Challenges *Canadian Social Science 13* (7), 1-5 doi:DOI:10.3968/9698

Enkin, E., & Mejias-Bikandi, E. (2016). Using Online Translators in the Second Language Classroom: Ideas for Advanced level Spanish *LACLIL, 9*(1), 138-158 doi:doi:10.5294/laclil.2016.9.1.6

Field, A. (2005). *Discovering Statistics using SPSS*. London: Sage.

Garcia, I., & Pena, M. (2011). Machine translation-assisted language learning: writing for beginners. *Computer Assisted Language Learning, 24*(5), 471-487. doi:10.1080/09588221.2011.582687

Hair, J., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate Data Analysis*. Upper Saddle River: Prentice-Hall.

Haladyna, T. (1999). *Developing and Validating*

multiple-choice test items. New Jersey: Lawrence Erlbaum.

Harris, h. (2010). Machine Translations revisited: Issues and Treatment Protocol *The Language Teacher, 34*(3), 1-29.

Hayes, S. (2002). Acceptance, Mindfulness, and Science. *American Psychological Association, 12*, 101-105.

Heppner, P., Heppner, M., Lee, D., Wang, Y., & Park, H. (2006). Development and validation of a Collectivist Coping Style

Inventory. .*Journal of Counselling Psychology, 53*(1), 107-125.

Holt, A., Overgaard, D., & Engel, L. (2020). Health literacy, digital literacy and eHealth

655

literacy in Danish nursing students at entry and graduate level: a cross

sectional study. *BMC Nurse, 19*(22). doi:https://doi.org/10.1186/s12912-020-00418-w

Jimenez-Crespo, M. (2018). The role of translation technologies in Spanish language learning, Journal of Spanish Language Teaching. *Jounral of Spanish Language Teaching*, 1-13. doi:10.1080/23247797.2017.1408949

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31-36.

Keshavarz, M. (2011). *Contrastive analysis and error analysis*. Tehran: Rahnama Publication.

Kol, S., Schcolnik, M., & Spector-Cohen, E. (2018). Google Translate in Academic Writing Courses? *The EUROCALL Review, 26*(2), 50-57.

Lee, S. (2019). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*. doi:10.1080/09588221.2018.1553186

Linn, R., & Gronlund, N. (2000). Measurement and assessment in

teaching. In. New Jersey: Prentice Hall.

Lorenceau, A., Marec, C., & Mostafa, T. (2020). Upgrading the ICT questionnaire items in Pisa 2021. In: Organisation for Economic Co-operation and Development.

Lynn, M. (1996). Determination and quantification of

content validity. *Nursing Research, 35*, 382-385.

Munro, B. (2005). *Statistical methods for health care*

research. Philadelphia: Lippincott: Williams & Wilkins.

Napitupulu, S. (2017). Analyzing Indonesian_English Abstracts Translation in View of Translation Errors by Google Translate. *International Journal of English Language and Linguistics Research, 5*(2), 15-23.

Nino, A. (2004). *Recycling MT: A course on foreign language writing via MT post-editing.*

Nino, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning, 21*(1), 29-49. doi:http://dx.doi.org/10.1080/09588220701865482

Nordin, M., Ahamad, T., Zubairi, A., Ismail, N., Rahman, A., Trayek, F., & Ibrahim, M. (2016). Digital citizenship questionnaire. *International Education Studies, 9*(3), 71-80.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Neill, E. (2019). Training students to use online translators and dictionaries: The impact on second language writing scores. *International Journal of Research Studies in Language Learning (IJRSLL), 8*(2), 47-65.

Pilot, D., & Hunger, B. (1999). *Nursing research:*

principals and methods. Philadelphia: Lippincott Williams & Wilkins.

656

Putri, G., & Ardi , H. (2015). *Types of errors found in google translation: a modle of MT evaluation*. Paper presented at the Proceedings of ISELT FBS Universitas Negeri Padang.

Stevens, J. (2002). *Applied Multivariate Statistics for the*

Social Sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Sukkhwan, A. (2014). *Students' Attitudes and Behaviours towards the Use of Google Translate.* (Master Thesis). Prince of Songkla University, Thailand,

Trochim, W. M. K. (2001). *Research methods knowledge base* (2nd ed. ed.). Cincinnati, Ohio: Atomic Dog Pub.

Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). *Modeling coverage for neural machine translation.* Paper presented at the Proceedings of ACL 2016.