Contents lists available at ScienceDirect

Technical Innovations & Patient Support in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/technical-innovations-and-patient-support-in-radiation-oncology

Validation testing of a language translation device for suitability in assisting Australian radiation therapists to communicate with Mandarin-speaking patients

Darren Hunter[a,*], Richard Oates[b], Nigel Anderson[c], David Kok[d], Daniel Sapkaroski[e], Caroline Wright[a]

[a] Department of Medical Imaging and Radiation Sciences, School of Primary and Allied Health Care, Monash University, Clayton, Victoria, Australia
[b] Radiation Therapy Services, Peter MacCallum Cancer Centre Bendigo, Victoria, Australia
[c] Radiation Oncology, Olivia Newton-John Cancer Wellness & Research Centre, Austin Health Heidelberg, Victoria, Australia
[d] Department of Radiation Oncology, Peter MacCallum Cancer Centre Moorabbin, Victoria, Australia
[e] Radiation Therapy Services Peter MacCallum Cancer Centre, Parkville, Victoria, Australia

ARTICLE INFO

Keywords:
Radiotherapy
Translation
CALD
Communication
Cancer
Validation

ABSTRACT

Introduction: Clear, timely communication between practitioners and patients is key in ensuring equitable access to health services and optimal care. Australia's linguistically diverse population adds complexity to healthcare provision. This paper describes a validation study to assess clinical suitability of a language translation device, intended for use with Mandarin speaking patients undergoing radiotherapy (RT).

Materials and methods: After a comprehensive device selection process, common phrases used in RT practice were curated within one clinical center and translated by interpreters. Phrases were categorized by conversation type and readability (according to Flesch-Kincaid and FORCAST scores). Validation of device performance was undertaken by purposely selected radiation therapists (RTTs) who tested and evaluated the device using a survey with 5-point Likert scale responses. Statistical analysis was undertaken on Excel using Pearson's chi-square, z-test, interrater reliability/agreement and linear regression analyses.

Results: Six RTTs and two interpreters volunteered to participate in this study. 188 common phrases were spoken verbatim into the device and scored on a 5-point Likert scale, yielding an overall output accuracy of 66%. A z-test confirmed significance against prior comparative research and Linear regression analysis observed improved output between consecutive participants. 62.7% of interpreter scores were identical; a further 29.1% constituted a single point scoring variation. Poorer outcomes were observed with colloquial English and lower readability.

Conclusions: This study found the device produced suitable translation accuracy and identified language styles that should be avoided with use. Further research could consider clinical application, expanded languages and/or health disciplines, and development of a national RTT phrase list.

Introduction

The Australian community comprises people who speak over 300 languages and dialects, with 21.0% of Australians speaking a non-English language at home [1,2]. Mandarin is the most common language spoken (2.5%), aside from English [2]. Health care systems must support the ability of health care professionals to deliver high quality and safe health care across the entirety of the culturally and linguistically diverse community, particularly when English is a second language or patients cannot converse in English at all [3].

Regional policies recognize that '*effective communication is central to quality health care and human services and … to provide person-centered care*' [4]. Within radiotherapy (RT) practice, effective communication is paramount to informed consent, information provision, support for treatment-related stress/anxiety, management of side effects, treatment decision making and participation in clinical trials [5]. Professional interpreters are considered the gold-standard to overcoming health care communication barriers [4]. However, there are documented concerns

[*] Corresponding author.
E-mail addresses: darren.hunter@monash.edu (D. Hunter), richard.oates@petermac.org (R. Oates), nigel.anderson2@austin.org.au (N. Anderson), david.kok@petermac.org (D. Kok), daniel.sapkaroski@petermac.org (D. Sapkaroski), caroline.wright@monash.edu (C. Wright).

https://doi.org/10.1016/j.tipsro.2023.100207
Received 9 December 2022; Received in revised form 27 March 2023; Accepted 14 April 2023
Available online 25 April 2023

relating to both the cost and availability of interpreters [5,6]. By contrast, informal interpreters (bilingual staff and carers) present further challenges of reliability and confidentiality [5,6]. Mobile language translation devices (mobile translators) may be a viable alternative to interpreters [7–10]. Mobile translators employ machine translation – algorithms that convert one language to another [11].

Google Translate, a commonly used mobile translator, encompasses 90 languages with varying accuracy across different languages [6,8,10,12–16]. In the context of Chinese (Mandarin), research reports 46.0–81.7% accuracy [8,12,13]. English to Mandarin translations are considered particularly problematic due to linguistic differences such as divergent syntactic structures, omitted pronouns in Mandarin, and a higher prevalence of English morphology [14]. Several studies either condone use in the clinical setting [6,8] or rather, advocate for clear communication practices to reduce the likelihood of error [12].

There are a number of objective measures used to evaluate clinical viability of mobile translators, however, inconsistency of quality metrics in machine translation evaluation have been observed [7,9,17]. An Australian study identified 15 mobile applications (apps) that met criteria for evaluation and two were deemed appropriate for everyday clinical use [9].

Validation plays an important role in managing the risk of harm relating to software failure during the implementation phase of new medical devices [18]. Whilst the design and methodology of software validation practices is typically tailored, the common focus remains on risk mitigation [18]. Thus, robust validation will identify faults, rather than merely prove success [18].

The aim of this paper is to describe the process of; (1) establishing mobile translator device selection/clinical suitability, (2) common radiotherapy phrase compilation, and (3) device validation (determining the reliability and accuracy of translation when the device is used by radiation therapists (RTTs) to communicate English phrases with a Mandarin output). This paper will describe the device accuracy results, stratified by conversational type and readability score. The objective was to determine how the device might facilitate the different steps of daily RTT engagement and treatment, with varying complexity of language. Cognizant of technological advancement, it was hypothesized that the output would fare better than prior research – specifically, the 57.7% output accuracy as reported in the 2014 study by Patil & Davies [8].

Materials and methods

Device selection & assessment of clinical suitability

In October 2017, an environmental scan, with the search term 'translation device' was conducted of relevant websites (Amazon (Australia), eBay (Australia), Kickstarter, Indiegogo, Google and Bing), to compile a list of translation devices that were available for purchase and use within Australia. Criteria used to evaluate initial suitability included; 1. Compliance with infection control (ability to sanitize and use between patients/RTTs), 2. Bi-directional capacity, 3. Breadth of languages supported, 4. Cost, and 5. Offline capability.

Curating a list of common RTT phrases

The researchers observed radiotherapy practice in a large metropolitan RT service over two days in November 2017, noting common phrases used by treating staff. Phrases were corroborated by 22 RTTs during a one hour discussion group in January 2018. Subsequently, additional phrases were added and consensus reached amongst the RTTs. As such, a collated list of common phrases used by RTTs was established in March 2018 (see Fig. 1 for further detail). The final list of phrases were approved for use in validation testing of the translation device.

Four phrase categories were pre-determined, relating to the sequence of conversation occurring between patients and RTTs during daily RT; (1) conversational language (daily greeting), (2) common enquiries (small talk with the patient), (3) simple identifiers (identification checks) and (4) treatment instructions (as part of the in-room procedure. Creation of the phrase categories assisted in managing data collection and analysis.

Device validation methodology

Device validation methodology was subject to ethics review (low and negligible risk) by the Peter MacCallum Cancer Centre (Ref 40756) and Monash University (Ref 16793) Human Research Ethics Committees. Ethics approval was granted on 10 July 2018 and 28 August 2018, respectively. In order to evaluate device accuracy and reliability, predetermined English phrases were spoken verbatim into the device by volunteer RTTs. Mandarin device output was scored by two interpreters.

Sampling & recruitment of RTTs and interpreters

A purposive sample of RTTs employed at the host clinical center, with broad experience levels were recruited by email. The first six RTTs to respond were accepted as participants. Interpreters who held National Accreditation Authority for Translators and Interpreters (NAATI) certification with experience in Mandarin translation, who were employed at the clinical center were invited to participate, via email. NAATI Certified Interpreters are skilled in both sight translation (i.e. read/write documentation) and monologue interpreting (i.e. interpretation of spoken word) – thus, best placed to complete required tasks of transferring written and oral communication from Mandarin to English and vice versa. Participants were exempt from a formal consent process; participation constituted informed consent.

Determining device performance

Prior to device testing, the phrase list and scoring parameters were sent to each interpreter. The interpreters independently translated each phrase into Mandarin. Device testing was conducted in October 2018, following the steps outlined in Fig. 2.

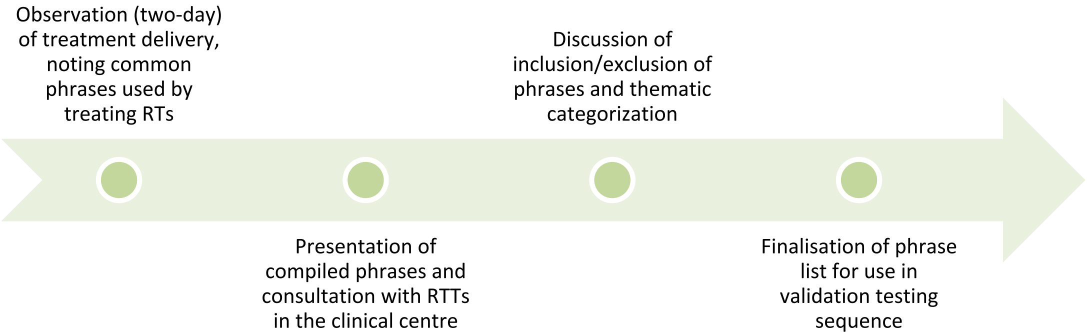
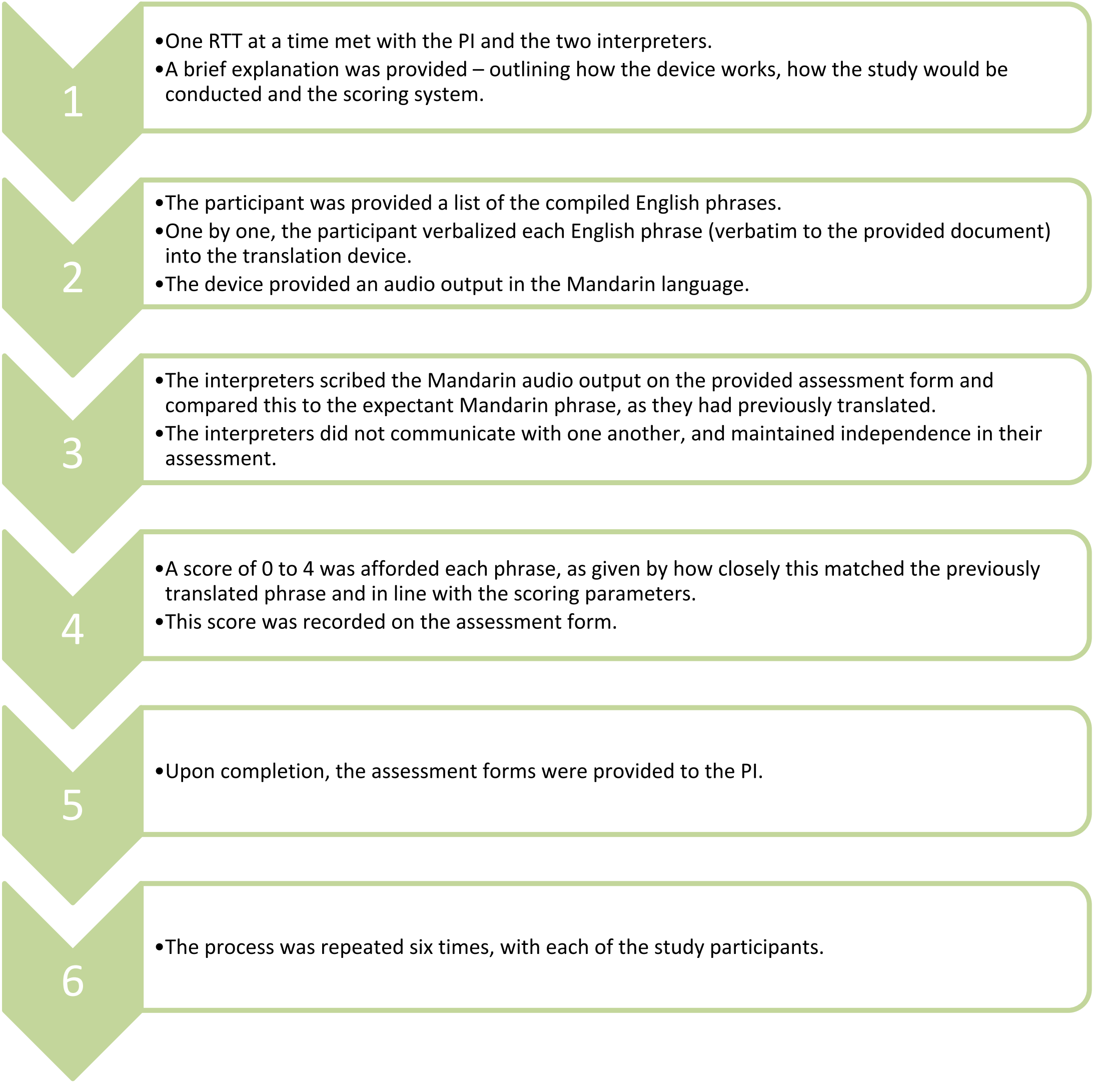

Observation (two-day) of treatment delivery, noting common phrases used by treating RTs
Presentation of compiled phrases and consultation with RTTs in the clinical centre
Discussion of inclusion/exclusion of phrases and thematic categorization
Finalisation of phrase list for use in validation testing sequence

Fig. 1. Curating a list of common RTT phrases.


1

- One RTT at a time met with the PI and the two interpreters.
- A brief explanation was provided – outlining how the device works, how the study would be conducted and the scoring system.

2

- The participant was provided a list of the compiled English phrases.
- One by one, the participant verbalized each English phrase (verbatim to the provided document) into the translation device.
- The device provided an audio output in the Mandarin language.

3

- The interpreters scribed the Mandarin audio output on the provided assessment form and compared this to the expectant Mandarin phrase, as they had previously translated.
- The interpreters did not communicate with one another, and maintained independence in their assessment.

4

- A score of 0 to 4 was afforded each phrase, as given by how closely this matched the previously translated phrase and in line with the scoring parameters.
- This score was recorded on the assessment form.

5

- Upon completion, the assessment forms were provided to the PI.

6

- The process was repeated six times, with each of the study participants.

Fig. 2. Device testing sequence.

In the absence of prior research, device accuracy assessment criteria were developed by the research team, and interpreters. A five-point Likert scale was devised to facilitate interpreter assessment of the phrases. Each interpreter awarded each phrase a score of 0–4 (Table 1). Individual scores of 3 or 4 indicated that the output phrase was correctly translated and understood by the interpreter. A ‘pass’ (i.e. translation deemed suitable for use) was only afforded when both interpreters scored the phrase a 3 or above. Thus, only four of the 25 combinations (16.0%) were awarded a pass by the research team (3–3, 3–4, 4–3, and 4–4). A ‘critical disagreement’ occurred when the interpreters scored on either side of the pass/fail threshold. In these circumstances a fail was defaulted as there was evidence of uncertainty. Interpreters were unaware of what combinations constituted a pass/fail to ensure unbiased assessment.

Theory/calculation

Validation study data included de-identified participant demographics and interpreter scores for the phrases. All data was entered into a Microsoft® Office Excel 2016 spreadsheet (Version v16.0, Microsoft ® Corporation, Redmond, USA). Descriptive statistics were used for primary data analysis. Where relevant, inferential statistics were analysed using R Software (Version 4.1.2, https://www.r-project.org, Vienna, Austria).

Table 1
Scoring Criteria for device output accuracy.

Score	Explanation	Example
0	Complete mismatch, wrong meaning	Cairns in leash
1	One or few correct words, wrong meaning	Cans are English
2	Impartial match, ambiguous	Will discuss English?
3	Impartial match, same meaning	Can you talk in English?
4	Complete match	Can you speak English?

Nb. Green shaded cells indicate individual scores deemed a pass. Red cells indicate a fail.

Overall device accuracy

Overall device accuracy was evaluated using combined interpreter scores – indicative of a pass, 'critical disagreement' or fail. Interpreter phrase scoring variation was analyzed using frequency and magnitude of change. Descriptive statistics outlined the proportion of matched scores for phrases, instances where one interpreter scored the device output higher than the other, and the degree of variation between the two interpreters' scores.

Hypothesis testing

A single proportion hypothesis test (z-test) was employed to refute a null hypothesis of 57.7% translation accuracy ($P_0 = 0.577$), as previously reported in a related study by Patil and Davies [8]. The z-test considered probability of chance in achieving the overall device accuracy results as outlined above.

Performance by conversation categories

Pearson's chi-square test of independence investigated ordinal data relationships with pass/fail rates and the impact of categorisation. The *p*-values were computed by Monte Carlo simulation with 10,000 replicates. All tests were two-sided and alpha was set at 0.05.

Interrater reliability and interrater agreement were subject to further analysis by means of Cohen's Kappa analysis and Kendall's rank correlation coefficient, respectively [19]. In both equations, values close to 1 indicate strong agreement/association.

Cohen's Kappa (k) is given by:

$$k = (p_0 - p_e) / (1 - p_e)$$

By contrast, Kendall's rank correlation coefficient (τ) is given by:

$$\tau = ((\text{number of concordant pairs}) - (\text{number of discordant pairs})) / \text{number of pairs}$$

Performance by readability scores

Readability was specified by Flesch-Kincaid grade levels [20,21]. This is a validated measure of literacy for US students and commonly reported in health academia [22,23]. Flesch-Kincaid scores do not hold an upper limit, but can fall as low as −3.4 where there are few, monosyllabic words within a passage of text. [20] It is calculated by the following equation:

$$\text{Flesch-Kincaid Grade Level} = 0.39\ (\text{total words/total sentences}) + 11.8\ (\text{total syllables/total words}) - 15.59$$

One of the key limitations to the Flesch-Kincaid grade levels (and related readability metrics) is that it is best applied to narrative text [24–26]. By contrast, the FORCAST index does not account for sentence length and was developed to analyze non-narrative text, such as phrase lists [24,27]. For this reason, the authors elected to report the FORCAST grade levels in addition to Flesch-Kincaid grade levels. FORCAST is calculated by the following equation:

$$\text{FORCAST Grade Level} = 20 - (\text{number of single-syllable words in a 150-word sample} / 10)$$

Flesch-Kincaid and FORCAST grade levels are given by; (1) preschool/kindergarten, (2) elementary school, (3) middle school, (4) high school and (5) college. Table 2 outlines the scores and corresponding US grade levels relating to benchmarked literacy levels. These scores allow for meaningful comparison with related research in this field. As above, chi-square tests were used to assess device accuracy rates and readability score variance.

Table 2
Flesch-Kincaid & FORCAST grade levels (US) for phrases.

Grade Level Score	Corresponding US Educational System
<0.99	Preschool/Kindergarten
1.00–5.99	Elementary School
6.00–9.99	Middle School
10.00–12.99	High School
>13.00	College

Results

Selection of translation device for study use

The environmental scan yielded four devices available in the Australian market; two handheld (Logbar Ili and Travis Translator) and two earpiece (Waverly Labs Pilot and Google Pixel Buds) devices. Table 3 outlines the key features of the devices and why Travis Translator was selected for use within this study. Key attributes of the Travis device included compliance with infection control guidelines, capability to provide bi-directional communication, scalability across a large bank of languages, moderate cost and limited offline usability (twenty languages were available for use in offline and online environments; a further sixty were only available online). The two earpiece devices were considered incompliant with infection control guidelines, as they were not recommended for use between individuals, and presented greater challenge in maintaining sanitary conditions. Furthermore, in 2017 the first generation of Google Pixel buds required a second set (and two corresponding mobile phones) to facilitate bi-directional communication.

Common RTT phrases

A list of 188 common Australian RTT phrases was compiled for use in the study (Appendix 1). Observation of conversations during daily radiotherapy identified 180 original phrases. A further five phrases that reflected pre-treatment patient instructions, and three additional iterations on existing phrases were included. Phrase types were defined as; (1) conversational language, (2) common enquiries, (3) simple identifiers and (4) treatment instructions. Categorization by readability was given by Flesh-Kincaid/FORCAST grade levels. In light of limited phrases of middle school level and above (per Flesch-Kincaid calculation), these were grouped for meaningful analysis.

Study participants

Participant demographic data is provided in Table 4. All six RTT participants were born in Australia, with three males and three females, aged from 24 to 43 years (mean = 32, SD = 7.24) and from a variety of ethnic backgrounds. By contrast, both interpreters were born in China with Level 3 Advanced Diploma translation qualifications, one male and one female, and ages 30 and 34, respectively.

Scoring variation

188 phrases were spoken verbatim into the device by each of the six RTT participants (1128 phrases in total). In 62.7% of phrase trials (n = 707), the interpreters agreed on the same score (Fig. 3). Interpreter 1 recorded a score that was higher than Interpreter 2 in 27.8% of trials (n = 314), whereas the opposite was true in only 9.5% of trials (n = 107). By magnitude of scoring variation, a 1-point disparity of scores occurred in 29.1% of trials (n = 328), followed by 2-point (6.0%, n = 68), 3-point (1.4%, n = 16) and 4-point (0.8%, n = 9) disparities.

Applying Cohen's Kappa calculation for interrater reliability yielded a 92.3% agreement, and k = 0.82, thus indicating an almost perfect agreement. Similarly, Kendall's τ = 0.95, demonstrates a strong

Table 3
Criteria for selection of translation devices (correct as at October 2017).

Device Name	Cost (US$)	Languages supported (n)	Offline capability	Bi-directional communication	Infection control compliance
Google Pixel Buds	$159*	40	✓	Requires second device	×
Logbar Ili	$189	3	✓	×	✓
Travis Translator	$199	80	Limited	✓	✓
Waverly Labs Pilot	$299	20	×	✓	×

* Price is given per device, excluding corresponding mobile handset.

Table 4
Participant & Interpreter Demographics.

Age	Gender	Ethnicity	Birth Nation	Profession
26	Female	Chinese	Australia	Radiation Therapist
28	Female	European	Australia	Radiation Therapist
36	Male	Pakistani	Australia	Radiation Therapist
24	Male	Vietnamese	Australia	Radiation Therapist
43	Female	European	Australia	Radiation Therapist
35	Male	Malaysian	Australia	Radiation Therapist
30	Male	Chinese	China	Certified Interpreter (Level 3 – NAATI Adv Dip)
34	Female	Chinese	China	Certified Interpreter (Level 3 – NAATI Adv Dip)

association between the two interpreters – thus satisfying reliable and consistent scoring.

Determining device performance

Overall device performance

Evaluation of device performance considered the accuracy and reliability of translating spoken English to the audio Mandarin output. This study demonstrated a phrase pass rate of 66.0% (n = 744) across 1128 trials (188 phrases × 6 participants). By contrast, 26.3% (n = 297) constituted a fail, and 7.7% (n = 87) introduced uncertainty between the interpreters, thus also failed.

Performance by sequential participant

Successful device output accuracy in order of RTT participation was found to be 60.6%, 64.4%, 65.4%, 65.4%, 70.2% and 69.7%, respectively (see Fig. 4). Linear regression modelling was employed to consider significance in the progressive improvement of scores across sequential participants. A *p*-value of 0.0049 confirmed statistical significance in the rate of successful device accuracy with repeated use.

Performance by conversation categories

The device accuracy results, given by conversation type, are outlined in Fig. 5. Device accuracy was reported at 72.6% (conversational language), 82.7% (common enquiries), 82.7% (simple identifiers), and 56.0% (treatment instructions), respectively. A chi-square test yielded a chi-square statistic of 73.1 and corresponding *p*-value of <0.05. Thus, results were statistically significant.

The fourth (and largest) category – treatment instructions – bore results that were lower than all other categories by a considerable margin. This category comprised a large number of phrases with colloquialisms and slang terms. For example, the base phrase ‘please move up’ scored a pass in all trials, whereas similar phrases ‘please *shuffle* up’ and ‘please *wriggle* up’ failed.

Performance by readability scores

When adjusted by Flesch-Kincaid grade level (Fig. 6), the results demonstrated marginal improvement in pass rate for phrases of middle school, high school or college level. A pass rate of 78.7% (n = 85) was observed for this cohort, as compared with preschool/kindergarten (64.2%, n = 497) and elementary school (65.9%, n = 162), respectively.

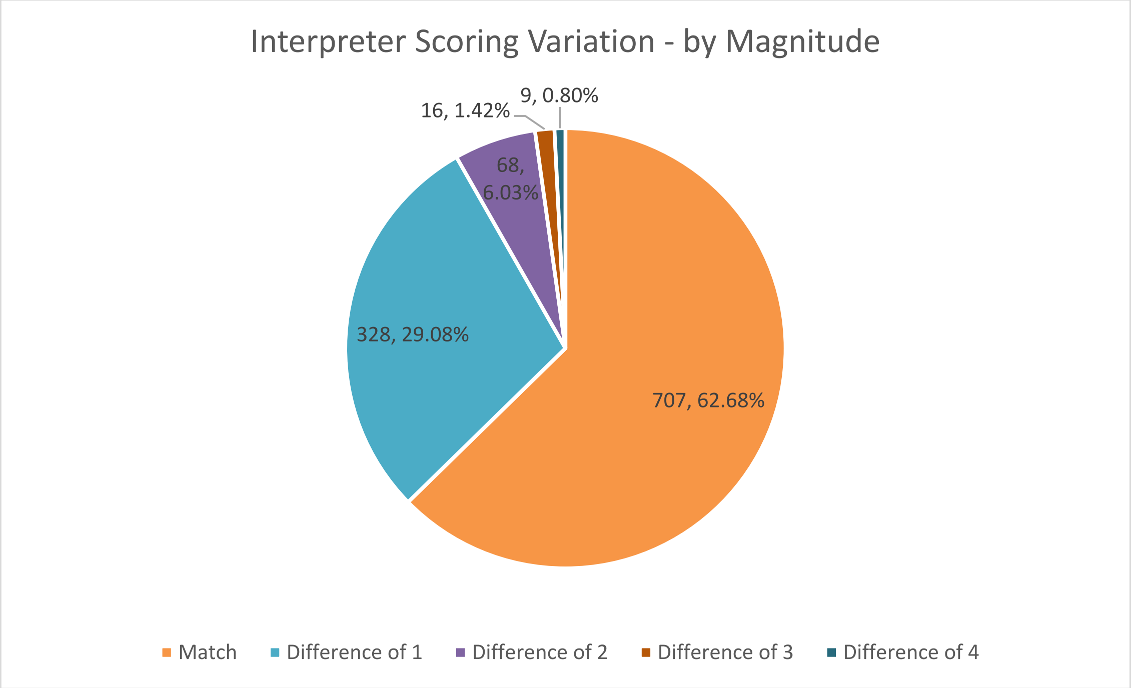


Fig. 3. Interpreter scoring variation – magnitude of difference between interpreters’ scores.

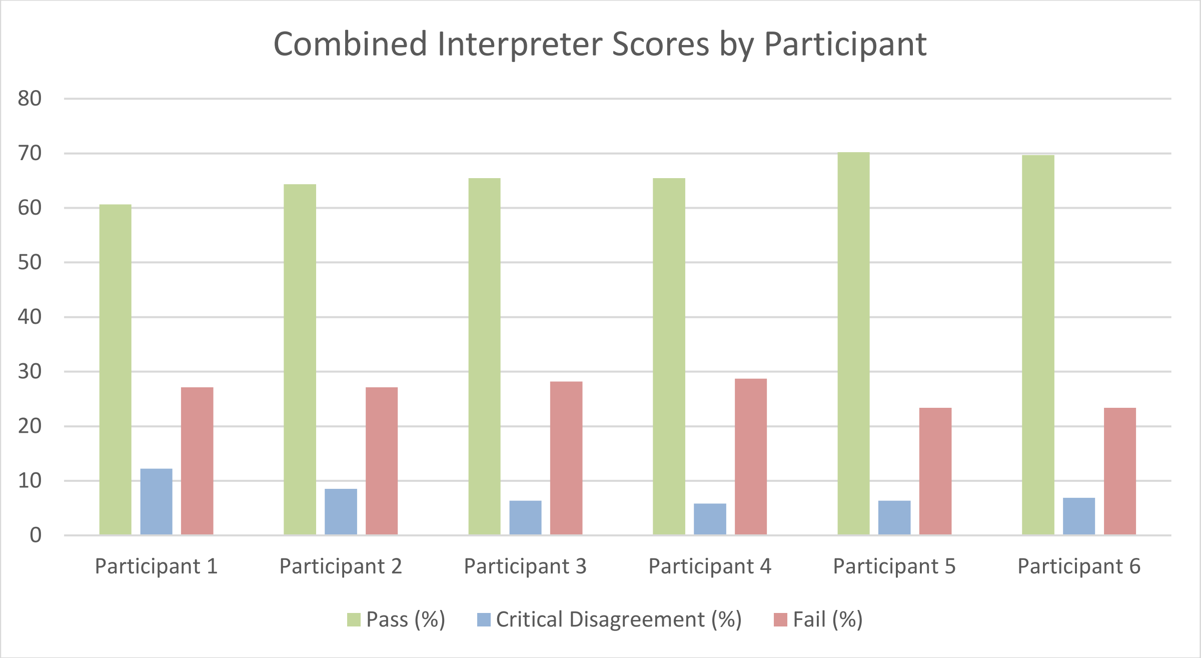


Fig. 4. Combined interpreter scores by participant.

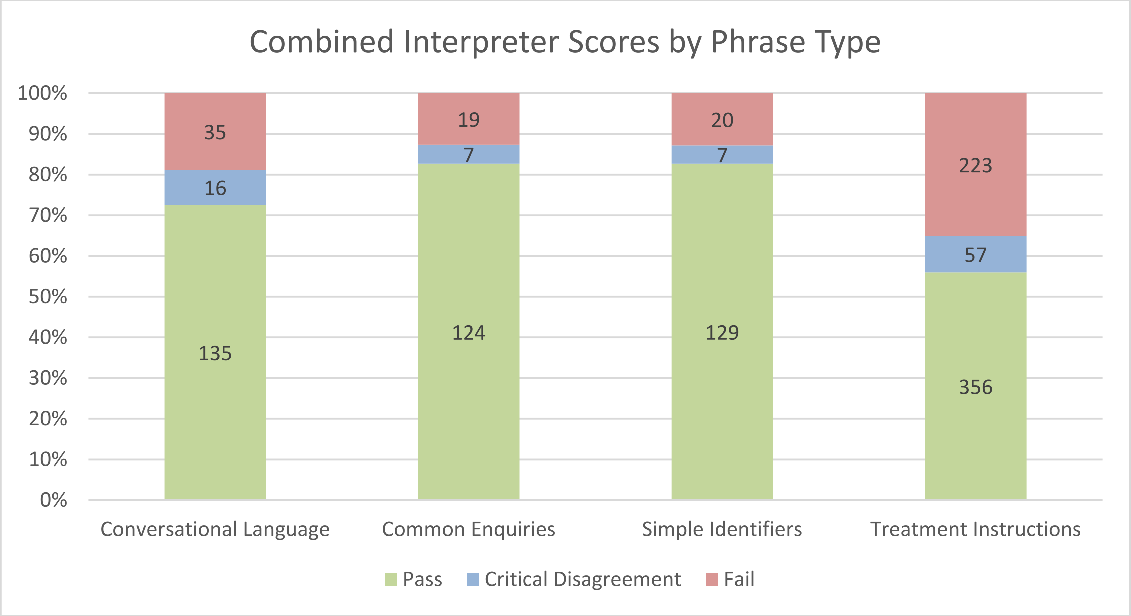


Fig. 5. Combined interpreter scores by phrase type.

Due to sampling limitations, educational levels of middle school and above were combined, yet remained the smallest cohort (n = 18 phrases, 108 trials). Flesch-Kincaid grade levels were analysed by preschool/kindergarten level (mean = 1.4, SD = 1.4), elementary school level (mean = 2.6, SD = 1.3) and middle school and above (mean = 10.4, SD = 3.9). The results were statistically significantly where, the chi-square statistic was 11.2 with a corresponding *p*-value of <0.024.

By contrast, FORCAST grade levels (Fig. 7) yielded varying results. No phrase satisfied preschool/kindergarten level. Pass rates were observed across elementary school (66.5%, n = 351), middle school (62.6%, n = 214), high school (66.7%, n = 132) and college (78.3%, n = 47), respectively. Mean and standard deviations were given by: elementary school (mean = 5.0, SD = 0), middle school (mean = 8.1, SD = 0.5), high school (mean = 10.7, SD = 1.0) and college (mean = 19.0, SD = 2.1). The results were, however, not statistically significant at $p < 0.05$ (chi-square statistic = 8.1, *p*-value = 0.23).

Considering the grade 8 threshold, data spanning Flesch-Kincaid and FORCAST modelling is represented in Fig. 8. Pass rates for phrases at or below the grade 8 threshold were 64.2% (Flesch-Kincaid, n = 659) and 64.7% (FORCAST, n = 505) respectively. Phrases that exceeded the grade 8 threshold yielded pass rates of 83.3% (Flesch-Kincaid, n = 85), and 68.7% (FORCAST, n = 239), respectively. Chi-square statistics demonstrated significance in both Flesch-Kincaid (chi-square = 15.3, *p*-value = 0.0004) and FORCAST (chi-square = 8.2, *p*-value = 0.0165).

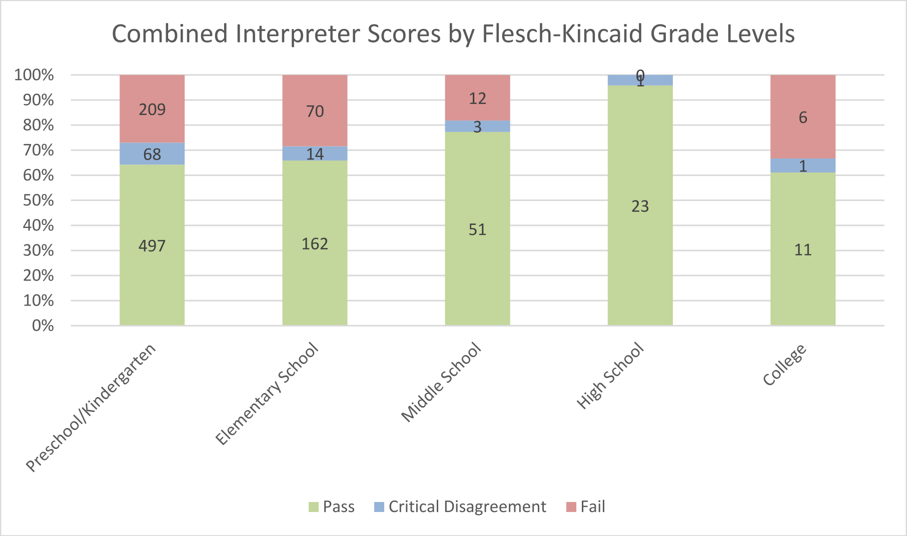


Fig. 6. Combined interpreter scores by Flesch-Kincaid grade levels.

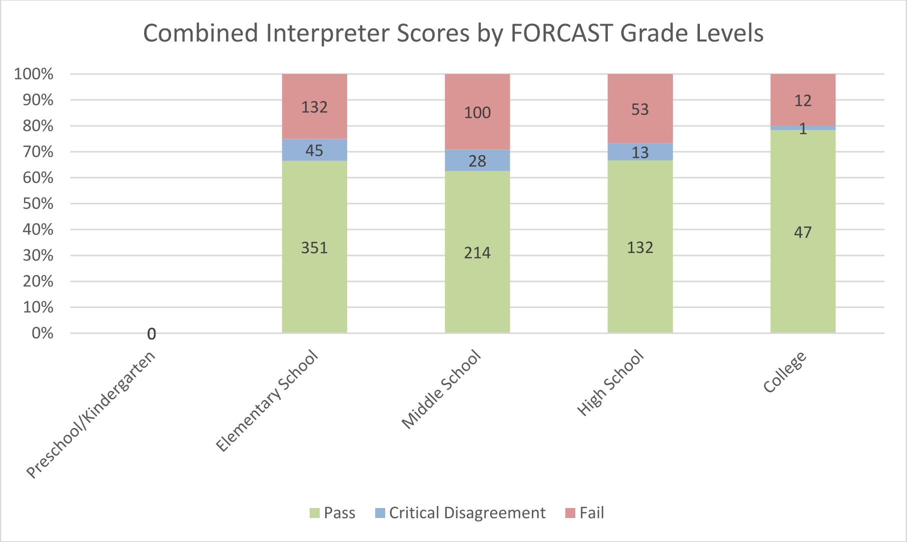


Fig. 7. Combined interpreter scores by FORCAST grade levels.

Testing the null hypothesis

A single proportion hypothesis test (z-test) was performed, considering a null proportion ($P_0 = 0.577$), sample proportion ($p = 0.660$ – as given by the aforementioned total pass rate), the sample size ($n = 1128$) and a significance of 5%. Considering the alternative hypothesis of $P > 0.577$, the z-test yielded a probability of $< 0.0001\%$, thus satisfying sufficient evidence to reject the null hypothesis. Thus, it was demonstrated that the results of this study were significantly greater than the prior study by Patil and Davies [8].

Discussion

Comparing validation methodologies

This study presents the only known validation methodology for a translator device intended for use within the context of RT. An overall pass rate of 66.0% demonstrates promise for the use of machine translation in a controlled RT setting – mitigating identified limitations to the device accuracy. Furthermore, the methodology sought to challenge the translation device – with a pass only afforded under strict conditions, agreed upon by both participating interpreters. However, the high pass rate does not appear to be consistent in related health applications. A 2016 study by Lear et al [28] employed online machine translation to convert 13 English sentences from clinical outcome assessments of undefined health disciplines into 13 different languages. The output was back-translated to English and the results were classified as incorrect (66%), conveyed with linguistic errors (26%) and correct (8%). No analysis of the foreign language output was performed, rather the back-translated English alone.

A similar process was undertaken to validate the US National Cancer Institute’s PRO-CTCAE outcome measurement. [29] Akin to the validation methodology employed in our study, two interpreters were utilized to perform independent translations. However, the PRO-CTCAE study employed forward and backward translations, as well as cognitive debriefing interviews to finalize the translated document for clinical use. [29] This process may offer further rigor to that of our study – as back translation provides an objective confirmation of output accuracy and an opportunity to ensure that phrase meaning is synonymous. However, the intention of this process was to establish documentation, rather than critically appraise output.

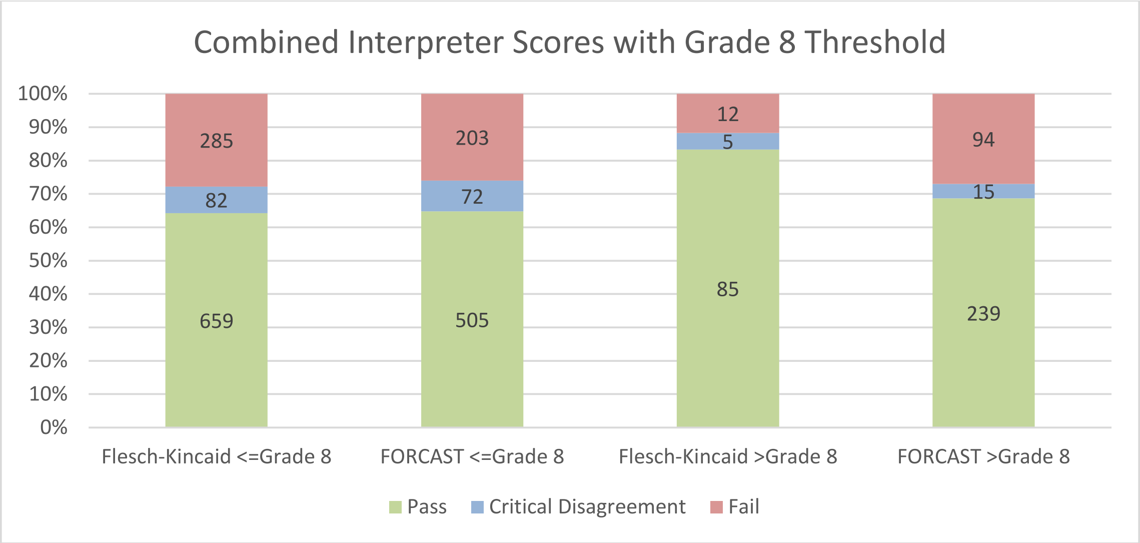


Fig. 8. Combined interpreter scores with grade 8 threshold (Flesch-Kincaid & FORCAST).

Health information is infrequently delivered at an appropriate readability level to meet the needs of the general population. [15] Our study of 188 phrases yielded a mean Flesch-Kincaid grade level of 0.63 (range −3.4–20.2). Compared to previous studies, an inverse relationship was observed, where higher readability scores correlated with improved translations. However, colloquialisms (such as 'shuffle' and 'wriggle') were identified key contributors to poorer translation outcomes. A comparable study utilizing a sample of six sentences (relating to diabetes education) were scored on the Flesch-Kincaid grade level, and then subjected to both machine and interpreter translation. [11] The mean Flesch-Kincaid grade level was calculated at 5.4 (range 2.8–9.0). Higher readability score correlated with poorer translation outcomes, although the sample size was very small. [15] Our results align with findings by Khoong et al [12], who claim significantly poorer translation outcomes with advanced readability (>8th grade), medical terminology and use of colloquial English. [12].

It is important to highlight that Flesch-Kincaid and FORCAST grade levels was observed to yield surprising results with advancing education and should be considered with caution. Calculation of readability with these metrics considers total words and syllables. Where multisyllabic single words existed, the calculation afforded higher scores. Examples include the terms 'fantastic' and 'wonderful' (Flesch-Kincaid = 20.2, FORCAST = 20) – corresponding with a College level readability. When considering scores beyond the grade 8 threshold, this phenomenon was more pronounced in the Flesch-Kincaid data (pass rate = 83.3%, n = 85), as compared with FORCAST (68.7%, n = 239). One must also consider the relationship between the US and Australian educational systems when drawing comparisons.

Study limitations & recommendations

The lack of published literature on validated instruments/tools required the researchers to develop a common phrase list. The curated list of phrases represents common language within the host RT department, though this is unlikely to differ greatly across Australian RT departments. There is merit in exploring options to formalize the phrase list on a national or international level for any future research. In curating a national/international RTT phrase list, a finalized list of appropriate English phrases for use with machine translation should consider:

(1) Validated and accurate translation
(2) The input English grade level (Grade 8 or below)
(3) The output Mandarin grade level (Grade 8 or below)
(4) Cultural nuances and sensitivity (i.e. polite and culturally-appropriate terminology)

The sample population of six RTTs could have also been expanded to observe the 'glass ceiling effect' – i.e. the threshold of translation accuracy with repetitive use. It was apparent the observed results demonstrated improved accuracy with sequential participants – further use may have continued to yield superior results. Furthermore, the authors recognize that the use of two interpreters could have introduced outlier results. A larger sample of interpreters may help to facilitate a reconciled agreement of scores, therefore aiding greater reliability and confidence in device output.

Finally, it is important to acknowledge that this research reflects the availability of natural language processing translation devices as at October 2017. The authors recognize considerable growth in machine translation technology over the subsequent years. As such, there is merit in conducting the market scan once again, identifying newer alternatives, which may prove suitable for repeated validation testing.

Whilst this study has demonstrated positive outcomes for language translation, one must consider the need for cultural competency training, should there be scope to pilot within the clinical environment. Similarly, in the interest of mitigating risk of known translation errors, avoidance of colloquialisms may prove suitable. However, technological advancement is continuing to drive improvements in dialects and regional languages. It is important to note that English (Australian) was added to the list of available languages on the Travis translator device in 2020.

Conclusions

This paper has presented a validation methodology for use in determining accuracy and reliability of machine translation. One-hundred and eighty-eight phrases were identified by the research team as commonly used by Australian RTTs. The interpreter device, Travis Translator, successfully translated in 66.0% of phrases. Colloquial Australian English reduced translation reliability and should be used with caution. The results warrant further research by way of clinical application testing, ensuring user training aligns with known limitations. Further validation could replicate the methodology across other languages and/or health disciplines, or use a national RTT common phrase list.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to acknowledge the Victorian Medical Radiation Practitioners Education Trust (VMRPET) for financial support of this research, as per award of the 2017 grant. Furthermore, the authors would like to extend their gratitude to Prof Michal Schneider, who contributed to the early development of this research program.

Appendix 1:. List of common RTT phrases

	Phrase	Flesch-Kincaid Grade Level	FORCAST Grade Level
Part 1: Conversational Language	Good morning	2.9	12.5
	Good afternoon	8.8	12.5
	Good evening	8.8	12.5
	Good day	−3	5
	G'day	−3.4	5
	Hello	8.4	20
	Hi	−3.4	5
	My name is…	−2.6	5
	I am a radiation therapist	10	11
	How are you?	−2.6	5
	Did you have a good weekend?	0.5	7.5
	Did you have a good night?	−1.5	5
	Have you had a good day?	−1.5	5
	How has your day been (so far)?	−1.8	5
	Have you had a nice day?	−1.5	5
	How was your morning?	0.7	8.8
	How was your afternoon?	3.7	8.8
	What's news?	−3	5
	What's the latest with you?	0.5	8
	What's been happening?	5.2	10
	What have you been up to?	−1.5	5
	What's goss(ip)?	−3	5
	What's going on?	1.3	5
	See you later	1.3	10
	See you tomorrow	5.2	10
	Goodbye	8.4	20
	Goodnight	8.4	20
	See ya (you)	−3	5
	Bye for now	−2.6	10
	Take care	−3	5
	Take care of yourself	0.7	8.8
Part 2: Common Enquiries	How do you feel today?	0.5	8
	How have you been?	−2.2	5
	Are you well?	−2.6	5
	Are you okay?	1.3	10
	Are you alright?	1.3	10
	How were you over the weekend?	2.5	10
	How were you overnight?	3.7	8.8
	How have you been today?	0.5	8
	How have you been recently?	2.9	8
	Have you been well?	−2.2	5
	Have you been well today?	0.5	8
	Have you been sick?	−2.2	5
	Have you been unwell?	0.7	8.8
	Are you in pain?	−2.2	5
	Is everything okay?	13.1	15
	Would you like to see a doctor?	0.6	7.1
	Would you like to see a nurse?	−1.1	5
	Can I help you?	−2.2	5
	What's wrong?	−3	5
	What's the matter?	1.3	10
	Do you need a hand?	−1.8	5
	Do you need help?	−2.2	5
	Do you need anything?	3.7	8.8
	Can you hear me?	−2.2	5
	Are you ready to go?	−1.8	8
Part 3: Simple Identifiers	What is your name?	−2.2	5
	Please tell me your name	−1.8	5
	What is your address?	0.7	8.8
	Please tell me your address	0.5	8

(continued on next page)

(*continued*)

	Phrase	Flesch-Kincaid Grade Level	FORCAST Grade Level
	Where do you live?	−2.2	5
	Where are you from?	−2.2	5
	What is your date of birth?	−1.5	5
	Please tell me your birthday	0.5	8
	Please tell me your date of birth	−1.1	5
	When were you born?	−2.2	5
	When is your birthday?	0.7	8.8
	How old are you?	−2.2	5
	What part of your body are we treating?	2.3	8.8
	Which part of your body are we treating?	2.3	8.8
	Whereabouts are we treating (on you)?	6.6	12.5
	Please point to where we are treating (on you)	0.6	7.1
	Can I see your ID?	−1.8	8
	Can I see your identification?	10	8
	Do you have identification?	12.5	8.8
	Do you have identification on you?	8.4	7.5
	Do you have identification with you?	8.4	7.5
	Do you have a driver’s license?	4.5	10
	Do you have your driver’s license?	4.5	10
	Please show me your driver’s license	4.5	10
	Please show me your ID	−1.8	8
	Please show me your identification	10	8
Part 4: Treatment Instructions	Please move up	−2.6	5
	Please move up the bed	−1.8	5
	Please move down	−1.8	5
	Please move down the bed	−1.8	5
	Please move to your left	−1.8	5
	Please move to your right	−1.8	5
	Please move away from me	0.5	8
	Please move closer to me	0.5	8
	Please shuffle up	1.3	10
	Please shuffle up the bed	0.5	8
	Please shuffle down	1.3	10
	Please shuffle down the bed	0.5	8
	Please shuffle to your left	0.5	8
	Please shuffle to your right	0.5	8
	Please shuffle away from me	2.9	11
	Please shuffle closer to me	2.9	11
	Please wriggle up	1.3	10
	Please wriggle up the bed	0.5	8
	Please wriggle down	1.3	10
	Please wriggle down the bed	0.5	8
	Please wriggle to your left	0.5	8
	Please wriggle to your right	0.5	8
	Please wriggle away from me	2.9	11
	Please wriggle closer to me	2.9	11
	Please stay still	−2.6	5
	Please don’t move	−2.6	5
	Please don’t help	−2.6	5
	Please stop	−3	5
	Keep going	−3	5
	A little more	1.3	10
	A little further	5.2	15
	A little less	1.3	10
	A bit more	−2.6	5
	A bit further	1.3	10
	Nearly there	2.9	12.5
	Not so far	−2.6	5
	A little bit	1.3	10
	Please breathe in	1.3	5
	Please breathe out	1.3	5
	Please hold your breath	−2.2	5
	Please breathe	2.9	5
	Inhale	8.4	20
	Exhale	8.4	20
	Please raise your chin	−2.2	5
	Please drop your chin	−2.2	5
	Please look up	−2.6	5
	Please look down	−2.6	5
	Please close your eyes	−2.2	5
	Please open your eyes	0.7	8.8
	Please turn your head away	0.5	8
	Please turn your head to me	−1.5	5
	Please sit down	−2.6	5

(*continued on next page*)

(*continued*)

Phrase	Flesch-Kincaid Grade Level	FORCAST Grade Level
Please sit	−3	5
Please lie down	−2.6	5
Please take a seat	−2.2	5
Please stand	−3	5
Please lie on your back	−1.8	5
Please lie on your stomach	0.5	8
Please lie face down	−2.2	5
Please lie on your side	−1.8	5
Please lie on your left side	−1.5	5
Please lie on your right side	−1.5	5
Please lift your feet	−2.2	5
Please lift your head	−2.2	5
Please lift your leg	−2.2	5
Please lift your arm	−2.2	5
Great	−3.4	5
Perfect	8.4	20
Spot on	−3	5
Well done	−3	5
Good	−3.4	5
Wonderful	20.2	20
Fantastic	20.2	20
Back soon	−3	5
See you soon	−2.6	5
Back in a minute	0.7	8.8
Back shortly	2.9	12.5
Back in a moment	0.7	8.8
See you shortly	1.3	10
See you in a few minutes	0.5	7.5
See you in a few moments	0.5	7.5
Follow me	2.9	12.5
Just this way	−2.6	5
Please get changed	−2.6	5
Please take your jumper off	0.5	8
Please take your shirt off	−1.8	5
Please take your trousers off	0.5	8
Please take your pants off	−1.8	5
Please take your top off	−1.8	5
Please take your dress off	−1.8	5
Please take your shoes off	−1.8	5
Please take your socks off	−1.8	5
Please take your glasses off	0.5	5
Please keep your underwear on	2.9	8
Please take your hearing aids out	0.5	7.5
Please turn off your phone	−1.8	5
Please take the basket with you	0.5	7.5
Please take your teeth out	−1.8	5
Please take your dentures out	0.5	8
Would you like water?	0.7	8.8
Have you drank your water?	0.5	8
Have your emptied your bowels?	2.9	11
Have you been to the bathroom?	0.5	7.5
Have you been to the toilet?	0.5	7.5
Would you like a glass of water?	0.6	7.1
Can I get you anything?	2.9	8

References

[1] Australian Institute of Health and Welfare. Australia's Health 2018: 5.3 – culturally and linguistically diverse populations; 2018. <https://www.aihw.gov.au/getmedia/f3ba8e92-afb3-46d6-b64c-ebfc9c1f945d/aihw-aus-221-chapter-5-3.pdf.aspx> [accessed 06 December 2021].

[2] Australian Bureau of Statistics. 2016 Census Data Summary: Cultural Diversity in Australia; 2017. <https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/14E20FFC53277975CA25814D002405D4/$File/cultural%20diversity,%202016%20census%20data%20summary%20(updated).pdf> [accessed 06 December 2021].

[3] Napier AD, Ancarno C, Butler B, Calabrese J, Chater A, Chatterjee H, et al. Culture and health. Lancet 2014;384(9954):1607–39.

[4] Department of Health and Human Services (Victoria). Language Services Policy; 2017. <https://dhhs.vic.gov.au/publications/language-services-policy> [accessed 06 December 2021].

[5] Gargan N, Chianese J. A review of the literature surrounding the provision of interpreters in health care, focusing on their role in translating information for non-English-speaking cancer patients and issues relating to informed consent. J Radiother Pract 2007;6:201–9. https://doi.org/10.1017/S1460396907006152.

[6] Haith-Cooper M. Mobile translators for non-English-speaking women accessing maternity services. Br J Midwifery 2014;22(11):795–803.

[7] Dew KN, Turner AM, Choi YK, Bosold A, Kirchhoff K. Development of machine translation technology for assisting health communication: a systematic review. J Biomed Inform 2018;85:56–67. https://doi.org/10.1016/j.jbi.2018.07.018.

[8] Patil S, Davies P. Use of Google Translate in medical communication: evaluation of accuracy. BMJ 2014;349:g7392. doi: 10.1136/bmj.g7392.

[9] Panayiotou A, Gardner A, Williams S, Zucchi E, Mascitti-Meuter M, Goh AMY, et al. Language Translation Apps in Health Care Settings: Expert Opinion. JMIR Mhealth Uhealth 2019;7(4):e11316. doi: 10.2196/11316.

[10] Sciarra AMP, Batigalia F, de Oliveira MAB. Technological devices improving system of translating languages: what about their usefulness on the applicability in medicine and health sciences? Braz J Cardiovasc Surg 2015;30(6):664–7. https://doi.org/10.5935/1678-9741.20150087.

[11] Ogundokun RO, Awotunde JB, Misra S, Segun-Owolabi T, Adeniyi EA, Jaglan V. An android based language translator application. J Phys Conf Ser 2021;1767:012032. https://doi.org/10.1088/1742-6596/1797/1/012032.

[12] Khoong EC, Steinbrook E, Brown C, Fernandez A. Assessing the use of google translate for Spanish and Chinese translations of emergency department discharge

instructions. JAMA Intern Med 2019;179(4):580–2. https://doi.org/10.1001/jamainternmed.2018.7653.

[13] Taira BR, Kreger V, Orue A, Diamond LC. A pragmatic assessment of google translate for emergency department instructions. J Gen Intern Med 2021;36(11):3361–5.

[14] Turner AM, Dew KN, Desai L, Martin N, Kirchhoff K. Machine translation of public health materials from English to Chinese: a feasibility study. JMIR Public Health Surveill 2015;1(2):e17.

[15] Chen X, Acosta S, Barry AE. Evaluating the accuracy of google translate for diabetes education material. JMIR Diabetes 2016;1(1):e3.

[16] Rodriguez JA, Fossa A, Mishuris R, Herrick B. Bridging the language gap in patient portals: an evaluation of google translate. J Gen Intern Med 2020;36(2):567–9. https://doi.org/10.1007/s11606-020-05719-z.

[17] Khander A, Farag S, Chen KT. Identification and evaluation of medical translator mobile applications using an adapted applications scoring system. Telemed e-Health 2018;24(9):594–603. https://doi.org/10.1089/tmj.2017.0150.

[18] Vogel DA. Medical device software verification, validation, and compliance. 1st ed. Norwood MA: Artech House; 2011.

[19] Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9(3):330–8. https://doi.org/10.1016/j.sapharm.2012.04.004.

[20] Flesch R. A new readability yardstick. J Appl Psychol 1948;32:221–33. https://doi.org/10.1037/h0057532.

[21] Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report, Millington, TN: Naval Technical Training Command; 1975. p. 8–75.

[22] Byun J, Golden DW. Readability of patient education materials from professional societies in radiation oncology: are we meeting the national standard? Int J Radiat Oncol Biol Phys 2015;91(5):1108–9.

[23] Balushi MA, Ghosh S, Debenham B. The readability of online Canadian radiotherapy patient educational materials. J Med Imag Radiat Sci 2020;51:617–23. https://doi.org/10.1016/j.jmir.2020.09.007.

[24] Perni S, Rooney MK, Horowitz DP, Golden DW, McCall AR, Einstein AJ, et al. Assessment of Use, Specificity, and Readability of Written Clinical Informed Consent Forms for Patients With Cancer Undergoing Radiotherapy. JAMA Oncol 2019;5(8):e190260. doi: 10.1001/jamaoncol.2019.0260.

[25] Rosenberg SA, Francis DM, Hullet CR, Morris ZS, Brower JV, Anderson BM, et al. Online patient information from radiation oncology departments is too complex for the general population. Pract Radiat Oncol 2017;7(1):57–62.

[26] Rooney MK, Sachdev S, Byun J, Jagsi R, Golden DW. Readability of patient education materials in radiation oncology – are we improving? Pract Radiat Oncol 2019;9:435–40. https://doi.org/10.1016/j.prro.2019.06.005.

[27] Caylor JS, Stricht TG, Fox LC, Ford JP. Development of a simple readability Index for job reading material. Paper presented at the meeting the annual meeting of the American Educational Research Association, New Orleans, LA; 1973.

[28] Lear A, Oke L, Forsythe C, Richards A. "Why can't I just use Google Translate?" A study on the effectiveness of online translation tools in translation of COAs. Value Health 2016;18:A387. https://doi.org/10.1016/j.jval.2016.09.232.

[29] Kkf C, Mitchell SA, Chan N, Ang E, Tam W, Kanesvaran R. Linguistic validation of the simplified Chinese version of the US National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). BMC Cancer 2020;20:1153. https://doi.org/10.1186/s12885-020-07631-5.