# Evaluation of the effectiveness of a naturalness probe for assessing human and non-human generated COA translations in linguistic validation cognitive debriefing interviews

PCR86

Tim Poepsel, Chryso Hadjidemetriou, Rebecca Israel, Kwanele Masuku, Nkateko Ngalele, Rachael Browning

ISPOR Europe 2025 #ISPOREurope

## Introduction

The Wild et al. (2005) linguistic validation guidance establishes the goals of cognitive debriefing (CD) as probing comprehensibility and cognitive equivalence of translations, testing translation alternatives, and flagging conceptually inappropriate items, but does not mandate approaches or standard probes for conducting interviews.

CD subjects occasionally offer feedback that translations may be comprehensible and conceptually equivalent to source material without being natural or using patient-preferred wording. However, without standardized naturalness probing, this feedback is missed in an unknown proportion of CD interviews and may allow unnatural translations to go undetected or unresolved.

Emerging interest in AI-assisted LV translation amplifies this concern, with pilot studies suggesting patients and quality raters can detect AI-generated translations via divergences from naturalness, preferring human-generated, natural-sounding translations.

Accordingly, direct probing of naturalness may be essential for optimizing patient-centeredness and assessing quality of clinical outcome assessment (COA) translations, whether human-generated or not.

## Methods

We pilot tested a novel naturalness probe in CD interviews assessing translations of one 200-word chronic lung condition PRO in 12 languages for 15 countries, for a total of 21 language-country pairs (**see Table 1**). The instrument was composed of 42 unique sub-components (e.g., item; response option; instruction). For each instrument sub-component, patients provided feedback on translation naturalness, with unnaturalness defined as "language sounding 'translated' or 'machine-generated', using uncommon, awkward grammatical conventions, phrases or words". Patients were also rated on overall comprehension of the sub-component and paraphrasing success by the interviewer. **See Table 2** for an overview of the naturalness, comprehension, and paraphrasing probes and tasks.

There were 5 patients tested per language-country pair, for a total of 105 (61 females; 44 males; age range: 18-95 years; average educational attainment: 12.2 years, range 7-20 years – **see Table 1** for fine-grained demographic data for each sample).

The instrument itself consisted of 6 items covering different symptom areas, such as cough, fatigue, and breathlessness. Notably, the instrument lacked instructions or introductory text, and items themselves consisted of a symptom area (i.e., "cough") followed directly by a set of response options unique to each item.

### CD Sample Characteristics

| | Language-Country Pair | Age Range (years) | Male:Female | Average Education (years) | Education Range (years) |
|---|---|---|---|---|---|
| 1 | Arabic-Israel | 36-75 | 3:2 | 12.8 | 10-16 |
| 2 | Chinese-China | 36-75 | 3:2 | 11.4 | 9-15 |
| 3 | Chinese-Taiwan | 26-75 | 2:3 | 13 | 9-16 |
| 4 | Danish-Denmark | 18-85 | 2:3 | 14 | 10-18 |
| 5 | Dutch-Belgium | 18-85 | 2:3 | 12.2 | 10-16 |
| 6 | English-Australia | 56-85 | 1:4 | 14.4 | 8-19 |
| 7 | English-Canada | 56-85 | 2:3 | 13 | 11-16 |
| 8 | English-US | 26-75 | 1:4 | 11.6 | 10-14 |
| 9 | French-Belgium | 46-85 | 2:3 | 11.8 | 10-15 |
| 10 | French-Canada | 56-95 | 2:3 | 12.6 | 10-16 |
| 11 | French-France | 46-75 | 2:3 | 9.4 | 8-12 |
| 12 | German-Germany | 26-65 | 2:3 | 10.8 | 9-13 |
| 13 | German-Austria | 36-85 | 1:4 | 9.6 | 9-12 |
| 14 | German-Belgium | 26-65 | 2:3 | 10.2 | 9-12 |
| 15 | Hebrew-Israel | 36-75 | 3:2 | 12.6 | 10-16 |
| 16 | Italian-Italy | 46-85 | 2:3 | 12.2 | 10-16 |
| 17 | Russian-Israel | 36-75 | 3:2 | 12.6 | 10-15 |
| 18 | Spanish-US | 36-75 | 2:3 | 12 | 10-14 |
| 19 | Spanish-Argentina | 36-75 | 2:3 | 14 | 7-20 |
| 20 | Spanish-Spain | 36-75 | 2:3 | 15 | 12-18 |
| 21 | Vietnamese-Vietnam | 36-65 | 3:2 | 10.4 | 7-14 |

**Table 1:** Language-country pairs in the sample, along with demographic information for each patient sample

### Descriptions of CD Probes and Tasks

| CD Probe / Task | Description |
|---|---|
| 1 Naturalness | By 'natural', we mean the degree to which the text sounds like common and fluent language a speaker of this language in this locale would use or be able to easily understand. Unnatural language may be comprehensible, but sound 'translated' / 'machine generated", or as if it uses uncommon or awkward grammatical conventions and arrangements of phrases, or concepts that don't fit the target language, culture, or locale<br><br>If the language seems unnatural to a participant, ask them which parts of the translation seem unnatural, and if there are any translation changes that would improve naturalness in the target language |
| 2 Comprehension | By 'understood', we mean the extent to which the concepts in the translation were clear to the participant. Enter "No" if there were any difficulties interpreting the concepts, or if the source content as a whole was difficult for the participant to understand |
| 3 Paraphrasing | Ask the participant to paraphrase the translation and record with a "Yes" or "No" whether the participant's paraphrase of the translation was accurate |

**Table 2:** Descriptions of and wording for the different CD probes / tasks

## Results

Naturalness probes flagged a total of 60 source sub-components in 16/21 (76%) languages in the project (**see Table 3**). In 39/60 (65%) cases, the same sub-component was not flagged by standard probes for "comprehension" or "paraphrasing success" (**see Figure 1**). Translation updates were made in 37/60 (62%) cases where "naturalness" was flagged, and 23/37 (62%) of these updates were uniquely motivated by feedback to the "naturalness" probe (that is, patients only provided feedback to the naturalness probe in those instances, while comprehending and accurately paraphrasing the source content – **see Figure 2 and Table 4**). At an instrument level, of the 42 unique sub-components that composed the instrument, 24 (55%) were flagged by the naturalness probe across the entire study. Across all samples, 67/105 (64%) provided feedback on the naturalness of at least one source component, flagging naturalness a total of 184 times (**see Table 3**). **See Table 5** for an example of feedback received from the naturalness probe and action taken as a result.

## Conclusion

Pilot testing of a novel naturalness probe was well-tolerated and understood by patients, and provided unique, actionable CD feedback relative to standard comprehension and paraphrasing probes, leading to a significant number of patient-centered translation improvements. We suggest standardization of naturalness probes in CD interviews, with data supporting their ability to consistently flag a dimension of translation quality not consistently captured by standard probing approaches, with potential extension to quality assessment of non-human (i.e., AI) generated COA translations.

### Naturalness Probe Hits by Locale

| | Language-Country | Instrument sub-components flagged as unnatural | Total Naturalness probe hits | # of patients who responded to naturalness probe |
|---|---|---|---|---|
| 1 | Spanish-Spain | 15 | 28 | 4 |
| 2 | Dutch-Belgium | 8 | 26 | 5 |
| 3 | Chinese-Taiwan | 5 | 25 | 5 |
| 4 | Danish-Denmark | 5 | 14 | 4 |
| 5 | Spanish-Argentina | 4 | 16 | 5 |
| 6 | Vietnamese-Vietnam | 4 | 12 | 5 |
| 7 | Arabic-Israel | 3 | 15 | 5 |
| 8 | English-Australia | 3 | 8 | 4 |
| 9 | Chinese-China | 2 | 6 | 3 |
| 10 | French-France | 2 | 6 | 5 |
| 11 | German-Germany | 2 | 4 | 3 |
| 12 | Russian-Israel | 2 | 7 | 4 |
| 13 | Spanish-US | 2 | 7 | 5 |
| 14 | French-Belgium | 1 | 5 | 5 |
| 15 | French-Canada | 1 | 3 | 3 |
| 16 | Italian-Italy | 1 | 2 | 2 |
| 17 | English-Canada | 0 | 0 | 0 |
| 18 | English-US | 0 | 0 | 0 |
| 19 | German-Austria | 0 | 0 | 0 |
| 20 | German-Belgium | 0 | 0 | 0 |
| 21 | Hebrew-Israel | 0 | 0 | 0 |
| | TOTAL | 60 | 184 | 67 |

**Table 3:** Descriptions of and wording for the different CD probes / tasks

### Naturalness Probe Hits and Translation Updates

| Locale | Translation Sub-components Flagged as Unnatural | Translation Sub-components with only Naturalness Flagged | # Translation Updates when Naturalness was Flagged | # Translation Updates due only to Naturalness Probe Feedback |
|---|---|---|---|---|
| Arabic-Israel | 3 | 3 (100%) | 3 | 3 (100%) |
| Chinese-China | 2 | 0 (0%) | 2 | 0 (0%) |
| Chinese-Taiwan | 5 | 2 (40%) | 4 | 2 (50%) |
| Danish-Denmark | 5 | 5 (100%) | 3 | 3 (100%) |
| Dutch-Belgium | 8 | 2 (25%) | 4 | 2 (50%) |
| English-Australia | 3 | 3 (100%) | 2 | 2 (100%) |
| French-Belgium | 1 | 0 (0%) | 1 | 0 (0%) |
| French-Canada | 1 | 0 (0%) | 1 | 0 (0%) |
| French-France | 2 | 1 (50%) | 1 | 0 (0%) |
| German-Germany | 2 | 2 (100%) | 0 | NA |
| Italian-Italy | 1 | 1 (100%) | 0 | NA |
| Russian-Israel | 2 | 2 (100%) | 2 | 2 (100%) |
| Spanish-Argentina | 4 | 3 (75%) | 3 | 2 (67%) |
| Spanish-Spain | 15 | 12 (80%) | 5 | 4 (80%) |
| Spanish-US | 2 | 0 (0%) | 2 | 0 (0%) |
| Vietnamese-Vietnam | 4 | 3 (75%) | 4 | 3 (75%) |
| English-Canada | 0 | 0 | 0 | NA |
| English-US | 0 | 0 | 0 | NA |
| German-Austria | 0 | 0 | 0 | NA |
| German-Belgium | 0 | 0 | 0 | NA |
| Hebrew-Israel | 0 | 0 | 0 | NA |
| Total | 60 | 39 | 37 | 23 |

**Table 4:** Instances of naturalness flagged by locale, along with number of translation changes and their motivation (due only to naturalness; naturalness plus other probe feedback)

### Example Patient and Interviewer Feedback to Naturalness Probe and Action Taken

| | |
|---|---|
| Source sub-component | "I feel tired but can still do the things I would like to do" |
| Original translation (Spanish-Argentina) | Me siento cansado; pero todavía puedo hacer las cosas que me gustaría hacer |
| Patient and Interviewer feedback | Patient 1 and Patient 3 think this translation is too literal and deviates from the source; "todavía" and "I would like" are not used here with an appropriate and idiomatic meaning.<br><br>I agree with the subjects. "Todavía" (still) never has in Spanish the meaning that "still" has here in English, and when making a request, placing an order, etc., the idiomatic wording in Spanish would be "I would want", "I would wish", "I wish", etc., not "I would like". |
| Action | Translation update recommended |
| Updated translation | Me siento cansado, pero aún así puedo hacer lo que quisiera hacer |

**Table 5:** An example of patient and interviewer feedback to the naturalness probe and action taken as a result

### Unique Contribution of Naturalness Probe

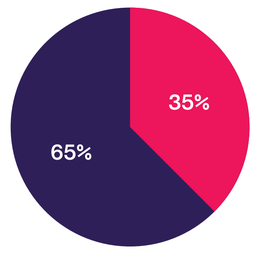


**Figure 1:** In 65% of cases where naturalness flagged an instrument sub-component, no other probes indicated a problem with the translation. In the remaining 35% of cases, the naturalness probe and one other (comprehension or paraphrasing) also indicated a problem.

### Translation Changes Driven Only by Naturalness

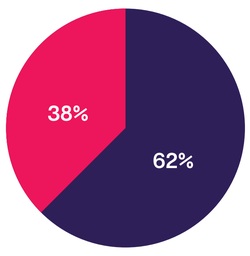


**Figure 2:** Across all 60 instances where naturalness probes flagged a potential source issue, 37 translation updates were made. In 62% of those cases (N=23), only the naturalness probe indicated a translation problem. In the remaining 38% of cases (N=14), the naturalness probe and at least one other (comprehension; paraphrasing) also flagged a potential problem.

rws