



TOPRA 2006  
EU Legislation - Is it Delivering Innovation?

## Readability Testing: Patient and Testers' Perspective

Simon Andriesen, MediLingua BV



Since 1996, MediLingua's core business has been medical translation and all types of related services. Last year, we have added performing readability tests to our list of services and in the meanwhile this has become an important part of our activity.

# Topics



- Setting up and Performing Readability Tests
- Ten Months' of Testing:  
Experiences and Best Practices
- Concerns



I have 20 minutes and will rush you through my 20 or so slides.

First I will discuss what it takes to set up and do readability tests

Then, I will share with you our observations, based on the dozens of tests we have completed to date.

Also, I will share some of our worries and concerns – things that we feel are obstacles on the way to optimal readability.

## What do we (providers) need



- Expertise regarding text and design of patient information leaflets
- In-depth know-how of the requirements (of what readability testing is all about)
- Qualified interviewers
- Large pool of testers
- Rooms, recording equipment
- Coffee, cookies, gift certificates



A test provider needs to understand patient information and know what it is all about. Also important to have is in-depth knowledge of the official requirements, including the various differences between the Agencies (most of them are small, one is rather big). Of course we need qualified interviewers, some prefer to use psychologists, but as far as I am concerned this is not necessary. We use interviewers who are very experienced in interviewing patients. They surely need to have good observation skills. Further, we obviously need a large pool of volunteer test participants. For each test we ‘consume’ 25 testers, and we should not use the same tester more than once every 6 months or so. Office space, digital recorders... We ‘pay’ testers a 15 euro gift certificate.

# Test preparation



- Define methods, testers' profile, criteria, 15-18 questions (plus questions about layout and style),
- Carefully check (and sometimes heavily pre-edit) 'baseline' PIL! Text should conform QRD template and EC Readability Guideline 1998
- Recruit either well-balanced (sex, age, education) or disorder-specific test-panels



Before the actual start of a test, we describe (and agree with the registration holder) the test methods, profile and recruitment of testers, and test questions.

What is very important at this point is that we take a very close look at the existing leaflet. Using our expertise regarding medical texts in general and patient information in particular, combined with 12 months' of readability testing, we can very quickly see what will go wrong with a leaflet. We then fix the textual problems and as you will see in one of the next few slides, this is when most of the changes are implemented.

It is not only textual changes that we fix, we also make sure the leaflet is in the current QRD format. We understand that the first thing an Agency does is check if the PIL is in QRD format, and if not they don't even look at it.

We also invite groups of test participants, and form test panels of 10 persons each, in a balanced mix regarding age, gender and education.

## The Actual Test



- Perform pilot test (5 testers) and then two test rounds (10 testers each)
- During interviews, testers are observed how they find information and when they get confused
- Many comments not related to specific questions
- Define problems and revise PIL after each round
- Deliverables: report ('certificate') and revised PIL



We always start with a pilot of a few – no more than 5 – testers. Most of the changes that are implemented during the complete test phase will have been implemented at this stage: before and immediately after the pilot test.

The end product is the readability test report and the resulting leaflet to be registered.

## The first 10 months: Testers



- Sex
  - Men 38%
  - Women 62%
- Age
  - 18% between 17 and 25 years
  - 61% between 26 and 64 years
  - 21% between 65 and 80 years
- Education
  - Secondary (low-mid-high): 42% (11-9-22%)
  - Tertiary (low-mid-high): 58% (14-19-25%)



We have done an analysis of the tests we have done until mid-September.

Of all testers we have used over the past 10-12 months, around 40% were men, 60% women. Men are just a little bit harder to find.

60% of the testers is in the age group between 25 and 65, 20% are younger than 25, 20% are older than 65.

Less than half of the testers has had secondary education, the rest had tertiary education.

## Testers (cont'd)



- No difference in score between men and women
- Younger testers (<25 years) scored 19% better (= more correct answers) than elderly testers (>65 years)
- Younger testers also considerably faster (are more used to recognising structure of information in general, which makes it easier to find requested information in a PIL)
- Testers are informed that we do not test **them**, but that they test the **leaflet**.
  - However, men often act as if THEY are tested!



Our analysis has not shown a difference between men and women.

We did see a difference concerning age: younger testers scores 19% more correct answers than people over 65.

They were also considerably faster.

Although testers are informed that we do not test them but that they test the leaflet, men typically act as if it is a race and as if they are tested.

## The first 10 months: Revisions



Average no. of changes per test round:

- During edit and pilot test: 57 (67%)
- During test round 1: 18 (21%)
- During test round 2: 11 (12%)

Remarks:

- Many revisions not necessary to pass test
- Leaflet writers use results for learning curve



In average we have made 86 changes to the tested leaflets.

As I already mentioned, during preparation and during the pilot test most of the changes are made. Our analysis shows that two thirds of all changes are implemented then. As a result of test round 1 around 20% is implemented, and a bit over 10% comes out of test round 2 (mind you, that is when 10 of 25 testers – 40%! - are used, so with 60% of the testers we have made 90% of the changes).

Many of the changes are not absolutely necessary to pass the test but they are based on various tester remarks and on our own observations and serve to further improve a leaflet that would already pass the test.

Good news is that leaflet writers use the test results in future leaflets.

## Can they find it?



- Finding the information is hardly ever a problem:
  - Pilot tests: 96%
  - First test rounds: 97%
  - Second rounds: 98%



In a readability test we need to assess whether testers can find the information, understand it and apply it properly.

Finding it is hardly ever a problem. As you see, almost all information is found, already during the pilot test, provided that the text of the Pil has been pre-edited, or if you wish, polished up.

## Do they understand it?



- Understanding/acting appropriately more difficult

Score (understood/total):

- Pilot tests: 90%
- First test rounds: 96%
- Second test rounds: 98%

- We tested two PILs without pre-edit or pilot:

- Test round 1: A) 76% and B) 83%
- Test round 2: A) 94% and B) 90%



Understanding is a bit more difficult. Initially 90% of the found information is understood, and after test round 2 this has improved to 98%.

These numbers are influenced by the fact that we heavily pre-edit leaflets before the test. We have tested 2 leaflets without much or any pre-edit and without a pilot test. During test round 1 one of these scored 76% and the other 83%. During the next test round this increased to 94 and 90% respectively. The difference between 76-94 for leaflet A and 83-90 for leaflet B can be explained by the fact that the customer for leaflet B did not want to make as many changes as we recommended.

## Time sheet



- Preparation 40 hrs
- 25 interviews and data analysis 50 hrs
- Testers 25 hrs
- Reporting after the test rounds 24 hrs
- Final report and appendices 16 hrs
  
- Tests take 4 to 10 weeks



For your information, and in case you want to do your own readability tests, for every test we spend a minimum of around 140 hours, plus the 25 hours of the 25 testers. Not mentioned on this slide are the 8 hours we spend on recruitment and scheduling of the test participants.

Depending on how busy we are and on how willing the customer is in responding within 48 hours after receiving our reports of the various test rounds we need 4 to 10 weeks. Our speed record has been a bit less than 4 weeks.

# Report



- Details of test method, testers recruitment and profile, questions, reports of pilot, first and second tests with score per tester per question, description of problems and resulting revisions (all in English, of course)
- Appendices include start and end versions of leaflet in QRD format and in layout, all intermediate versions with tracked changes, questionnaire



The report.

Apart from a highly readable leaflet coming out of the test, the report is what's it is about! Without a report, no registration!

All in English, of course. It includes information about how the test was prepared, set-up, and performed. What the testers looked like, what questions were asked, how the questions were answered and what remarks were made that have resulted in revisions to the leaflet.

A typical report has around 20-25 pages (8-10,000 words), plus some 75 pages of appendices, including all tested versions both in layout and in QRD format with track changes to show the modifications since the previous version.



## Concerns



During the past year we have identified some problems and have some concerns.

# Concerns



- Different Agencies seem to use different criteria:
  - “16 out of 20” (=80%)
  - “90% of 90%”  
(By the way: the aggregate minimum score is not automatically 81%, but can range between 81% and 90%, depending on percentage of found information)
- Not many producers aim for the minimal score!
- Assume BfArM or CBG accepts a score lower than 90% of 90%. What if the MHRA is involved as CMS in an MRP, will they accept the report?



One of them is about the criteria.

Many Agencies use the rule that of 20 testers (that is 2 test panels of 10 testers each) at least 16 must be able to answer all questions correctly. You could say this is an overall score of 80%.

The British MHRA uses the rule that 90% of the testers need to be able to find the information and 90% of those need to be able to understand and apply it appropriately. (By the way: this not automatically 81%. If all testers find the information and 90% of those need to understand it, this means that 90% of all testers need to understand it and not 81%).

So far, hardly any – if any - of our customers were aiming for the minimal score.

What if...

## Concerns



- A leaflet with a score of 16 out of 20 or 90% of 90% may meet the criteria, but it still is a RATHER POOR leaflet!
- Aim for score of >95% after second test round



Whatever the criteria: a leaflet that scores 16 out of 20, or 90% of 90% may meet the criteria, but it still is a pretty poor leaflet.

We aim for - and usually reach - well above 95%.

I would invite anyone providing readability test services to get together at some point to discuss what can be improved, and to possibly come up with some common 'best practices'.

We could define a structured critique of the template in so far as it stands in the way to readability.

We should also discuss the test methods. As far as we are concerned, 10 plus 10 testers is not ideal. It would be better to do 5 plus 5 plus 5 testers. Less testers (and that is OK as we find most problems with 10 to 15 testers and by then the Pil is in the plus-90% range), but one extra revision opportunity, and perhaps a bit lower cost.

# Concerns



- QRD template causes readability problems
  - Some headings not clear
    - Take special care with X
    - Taking X with food and drink
    - Important information about some of the ingredients of X
    - If any of the side effects gets serious ...
  - (Some national versions of template were poorly translated, but these have recently been improved!)
- Template is OK if PIL is for patients. Not ideal in cases of vaccines, IV, contrast media, oxygen
- Who wins: Template or Readability?



An interesting concern is about the QRD template and how during readability tests some of the mandatory headings and phrases are identified as problems.

Some examples are: 1. What does 'special care' mean?

2. This heading suggests that the patient will find instructions about how to take the medicine: with water, after a meal, under the tongue. However, this section is about possible interaction with non-medicinal substances.

3. This heading attracts extra attention because it is the longest in the leaflet while it concerns information about the relatively unimportant 'other ingredients' but the heading mentions 'important information'. This is confusing; is the rest of the information less important? Usually, it only mentions that the product contains lactose or alcohol or something like that. When that is the case, we usually change this heading into 'This product contains alcohol/ or lactose'.

4. First of all, to the best of my knowledge, and many British testers seem to agree, it is 'get' not 'gets'. And they would even prefer to use 'become'. Also, testers complain that there are many side effects that may get serious, but there is no way of knowing for a patient if it happens (blood pressure, liver function).

QRD template OK for real PILs. For products such as oxygen, contrast media et cetera, it is often difficult to squeeze the information into the template. The question is what prevails: template or readability?

## Concerns



- Some Agencies require verbatim answers by tester
  - For all  $2*15*10 = 300$  questions, this means 300 to 600 lines, or 12 to 24 pages, in each report. How useful is it to read how testers formulate the 95% correct answers if these answers are also provided in Questionnaire?
  - Of course, ‘raw data’ are relevant to illustrate incorrect answers and/or when answers or remarks lead to revisions
  - **BUT PLEASE NOTE:** Revisions can’t always be backed up by quotes. Many changes are made during preparation and many others during the evaluation of test round results, by interviewers, writer, designer



I have some concerns here: First of all, and that is the last bullet on this screen, many revisions to the PIL are made on the basis of what many people say, or how they behave, and after evaluation between the writer of the leaflet and the person who has interviewed the testers. So not all revisions can be backed-up by quotes.

Of course, if the Agencies want all answers, they can get them, but this means an extra dozen or two dozen pages in the report. Who is going to read all that, and what purpose would it serve? It is the end result that counts, and isn't it true that we also don't know why a leaflet writer has used certain phrases?

Almost all questions are answered correctly, and as far as we are concerned, it makes no sense to write out all the different ways a testers gives a correct answer. The correct answer is part of the report anyway (in the instruction to the interviewers), so the Agency can see what we felt the correct answer was. Of course, in cases of incorrect answers, it is very relevant to see what the testers said, and also in cases were an answer leads to a revision.

# Concerns



- Has tester experience with product or disease?
  - **NO PROBLEM:** For disease-specific products we recruit patients who use the product or have the disease
  - **PROBLEM:** For most products testers are recruited from the general population (as long as ‘they can imagine that they ever have to use the product’). Asking these **volunteers** about medicine experience or health is less relevant, can be inappropriate and may jeopardize test atmosphere. (And how should we check the information, if at all provided?)



# Test in one language only



- **Good news**
  - Not 21 times the test cost in case of an EU-wide launch
  - Possibility to do test in a country with an adequate test infrastructure and sufficient capacity, to reduce cost and/or to escape wait lines
- **Bad news**
  - Readability problems are often caused by shortcomings in the wording. Readability of a tested version can easily be destroyed during the translation phase



## So, who's side are we on?



- Industry? (Have to pay our bill...)
- Agencies? (Have to accept our report...)

We are right in the middle, but:

- Hey, we are potential patients ourselves!
- And, quality PILs have positive effect on therapy compliance, public health and health care cost!



A recent survey by the Netherlands Institute for Responsible Medicine Use (DGV) shows that just 8% of respondents feel the PIL is OK. 54% feels the PIL can be improved and that explanation by doctor or pharmacist is necessary, 21% feels that without a medical dictionary PILs can not be understood, and 17% of doctors feel that PILs are even for them hard to comprehend. Also, compliance can be hurt by a poor, or unnecessarily frightening PIL.



Back to the Theme of this Conference:  
“EU Legislation - Is it Delivering Innovation?”

**Innovation** is a strong word, but yes, readability testing certainly improves the quality and level of PILs and we see a learning curve in companies, where readability test revisions are incorporated into their standard phraseology



Thank you.



Any questions?

**Contact information:**

[simon.andriesen@medilingua.com](mailto:simon.andriesen@medilingua.com)

[www.medilingua.com](http://www.medilingua.com)

