Evaluation and Testing (IO5)

Louise Fryer

Utopian Voices Ltd.

ADLAB PRO

Ljubljana, 4th June 2019

Dept. of Legal, Language, Translation and Interpreting Studies, Section of in Modern Languages for Interpreters and Translators
University of Trieste, Via Filzi, 14 - 34144 Trieste, Italy
Project numberStudies: 2016-1-IT02-KA203-024311
www.adlabproject.eu
FUNDED BY THE ERASMUS + PROGRAMME OF THE EUROPEAN UNION

O5 details

- **Full title**: Evaluation and Testing
- **Leading partner**: Utopian Voices Ltd.
- **Contributing partners**: all
- **Other contributors**
  - All participants in ADLAB PRO events
  - AD students/trainers/lecturers/course deliverers/AD providers/users

A Guide to the Evaluation of Training Materials: ADLAB PRO a case study

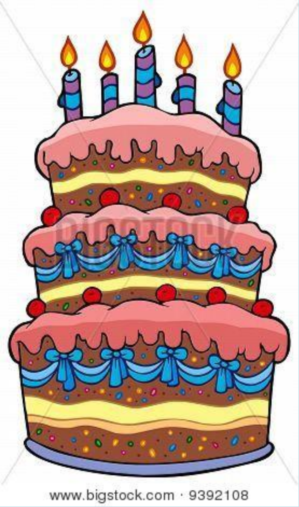
- a digest of the research: strategies, execution & results.
- **Chapter 1: methodological guidelines for evaluation**:
- Who? What? Why? When? & How? of evaluation.
- Chapter 2: Case Study 1: Evaluation of modules in IO3
  - Case Study 2: Evaluation of training materials in IO4
  - Appendix: all EFs used in ADLAB PRO.
  - 240pp. C.60,000 words

Why evaluate?

Evaluation provides evidence that justifies the value and viability of training programmes.

AIMS

- To improve a project. To check that it’s meeting its goals
- To better communicate its achievements
  - Serendipitous gains/unexpected insights
  - To show what we’ve done & to celebrate what we’ve achieved



There are 2 types of evaluation/Assessment

- **Formative** (assessing ongoing activities)
- **Summative** (assessing the end result)

- Formative: the chef tasting the soup as (s)he makes it
- Summative: everyone else tasting the soup when it’s ready

How: Quantitative versus Qualitative Measures

- Quant: Collection and analysis of numerical data.
- Numerical frequencies (percentages; averages – mean mode; standard deviation).
- Objective (large sample sizes overcome individual variation) can be generalised from participants to the population at large.
- Scientific.
- Replicable.
- Determine questions of cause and effect.

Qualitative Measures

- textual data from surveys, interviews, focus groups, (observation and ethnographies).
- Subjective.
- Less scientific.
- Less rigorous.
- Less replicable.
- Less easy to generalise.

But: Qualitative Measures

- Ask questions of real people in real situations (Plumb and Spyridakis, 1992).
- Produce richer data.
- By providing thick description of a specific context, the reader can apply the findings to their own situation.

Qual versus Quant: a comparison

- Quant: takes a hypothesis and tests it. The ideal is a randomised control trial.
- Qualitative uncovers what the hypothesis should be. It’s collaborative. It’s about listening.

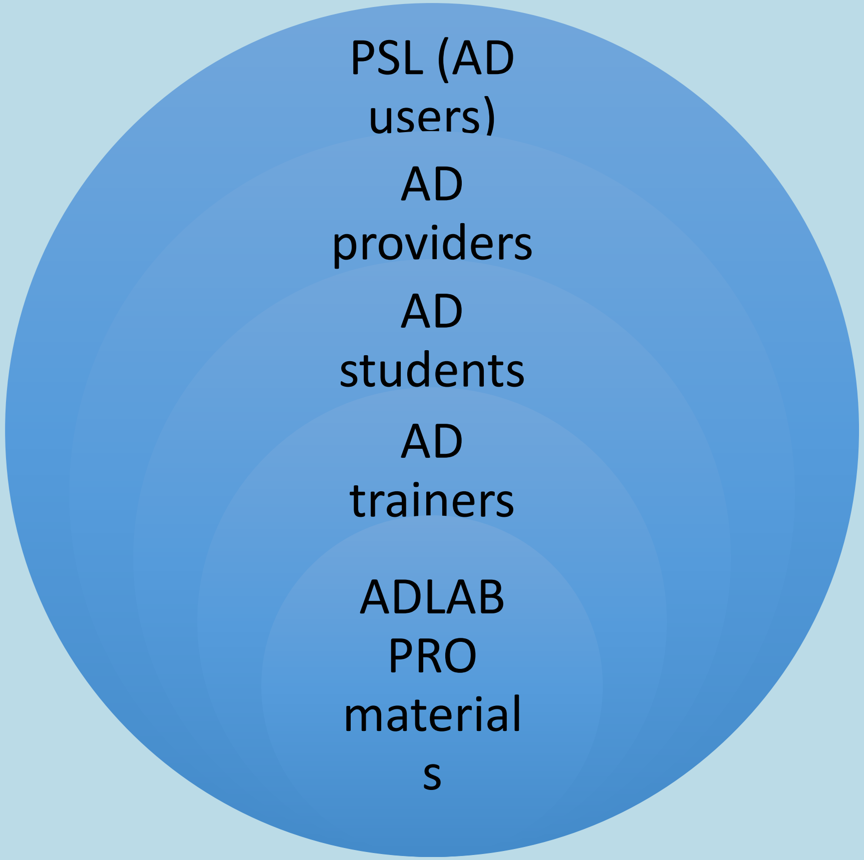
Mixed Methods

- Doesn’t have to be either or
- “Quiet revolution” O’Cathain (2009)
- Number of studies combining Quant & Qual approaches in Health research: 17% in the mid-1990s to 30% in the early 2000s
- Advantage: Intermethod discrepancies.
- Disadvantage: “more time, resources and effort to organize”
- (Collins, Onwuegbuzie & Jiao,2006)

Who? Stakeholders in ADLAB PRO

- “any group or individual who can affect or is affected by the achievement of the organization’s objectives” (Freeman,1984, p.46).
- Students of AD.
- Teachers/Tutors/Trainers/Lecturers.
- Providers of AD.
- Users of AD – principally PSL.

Hierarchy of Stakeholders



PSL (AD users)

AD providers

AD students

AD trainers

ADLAB PRO materials

Challenges

- Sheer number (60 core videos + associated ppt. slides and transcriptions; over 100 additional videos; 6 reading lists; 196 tasks).
- Timeline: Materials ready: March - June 2019.
- Project ends: Aug 2019.

Solution

- Evaluate a prototype for most material types; core videos; reading lists; tasks; introductory videos (formative, using key informants)
- Evaluate selected final examples (summative using stakeholders).

Example: Core videos: Process

- Core video prototype created by UAB
- 2 key informants (ext./formative)
- Accessibility by RNIB & SF (int/formative)
- Field testing (ext/summative)
- Mixed methods: “Shop window” evaluation (ME5) BCN
- Semi-structured interviews (qual.)
- Student evaluation (mixed)
- evaluation by AD Professionals (RTV-SLO)
- Focus Group with PSL (RNIB/UV) (qual.)

Method 1: Core videos: Key informants

- 2 AD trainers in academic environments: one from Poland, one from Spain.
- Experts in screen AD.
- test the acceptability and useability.
- user experience (UX) indicators: engagement and attention. Psychological indicators known to be linked to student success (Christenson, Reschly, & Wylie, 2012).

Quality Indicators (QI's)

1. The video makes a useful contribution to understanding the practice of AD. (/10)
2. The video gives a good overview of the module to students of AD.
3. The video held my attention.
4. The video was succinct.
5. The audio (voiceover) was engaging.
6. If I were running a training course on AD, I would include this video (please give your reasons).

QIs ctd.

7. The duration of the video was too short/too long/about right.

8. The video is well structured.

Text box for evidence/reason for each score.

Other comments.

QI’s marked /10 = total /70

Quantitative Results

- E01: 59/70
- E02: 67/70

- NB: E02 made more suggestions for improvement.

Acting on the evaluation

- “In min. 2 when talking about the examples, maybe add a screenshot of the Swedish TV or a picture from Inglorious Basterds?”

- Real world constraints: Budget, timeframe, copyright.

Another suggestion

- I would try to include a couple of short video examples (10-15 seconds?). In my lessons, some students don't understand the concept until they are shown a video with AST.”

Solution

- Create an additional video: much freer than the tightly structured core videos. More space for examples.

What does the Evaluation Process show?

- Intermethod discrepancies:
- “Is it succinct?”: assumed to be a positive indicator.
- 95% agreement that it was succinct. (Caveat: only two respondents)
- Qualitative comments showed it was not necessarily a positive indicator. Yes…but…
- “Maybe even too succinct. I'd prefer a 10-minute video including some video examples which allowed internalising at least the basic concepts.”

Evaluator Commitment

- Responded to the request.
- Responded in detail.
- Formative suggestions.
- Committed to improving the quality.

Method 2: Core videos ME5 "Shop window" evaluation

- M2_U5 time & space; M1_U6
- Evaluators: ME participants
- Session 1: UAB introduced IO4, the typology of materials and their main features. Examples of: reading lists; trainer's guides; an introductory video.
- Session 2: UAB presented : **core videos**, tasks, and additional videos.
- Data by hard copy questionnaire (Online also available).

Participants ME5 Session 2

- Practicing describers (3)
- AD user (1)
- Academic/researcher (8)
- AD Tutor/Teacher/Lecturer (1)
- Provider of AD content (3)
- Student (5)
- Multiple (14) (39%)
- Total = 36.

From

- Europe = 30
- Asia = 2
- Australia = 2
- North America = 1
- South America = 1

Results (Quant)

- Likert scale 1 – 5
- I find the core videos…
- …Interesting m = 4.58; s.d. = .649
- …well-structured m = 4.57, s.d. = .85
- …confusing m = 1.42, s.d. = .906
- …easy to understand m = 4.50, s.d. = .941
- …increased my understanding of AD m = 4.39, s.d. = .899

Results Qual

- Liked most/least:
- Positive Keywords
- “simple, easy to understand, useful, clear, accessible, short, visually polished.”

- Negative Keywords
- “pace (too quick) delivery (tone naïve), grammatical errors/typos, cheap graphics, too short.”

Method 3: Semi-structured Interviews (Qual)

- 9 interviewees: 3 from Europe/3 from N. America/2 from Australia/1 from Asia
- 4/9 audio describer & trainer.
- All would use the materials in a future training course and recommend them to other trainers (9/9)

Strengths

- “The opportunity to see how other people approach it, how other people frame it in terms of competencies is really useful.”
- “I'm working in Hong Kong. I want some references from outside the Chinese community.”
- “freely available under a creative commons licence.”
- **“**I find them hands on, varied, creative and based on experience - both teaching experience professional experience and connections with the people, with the users.”

weaknesses

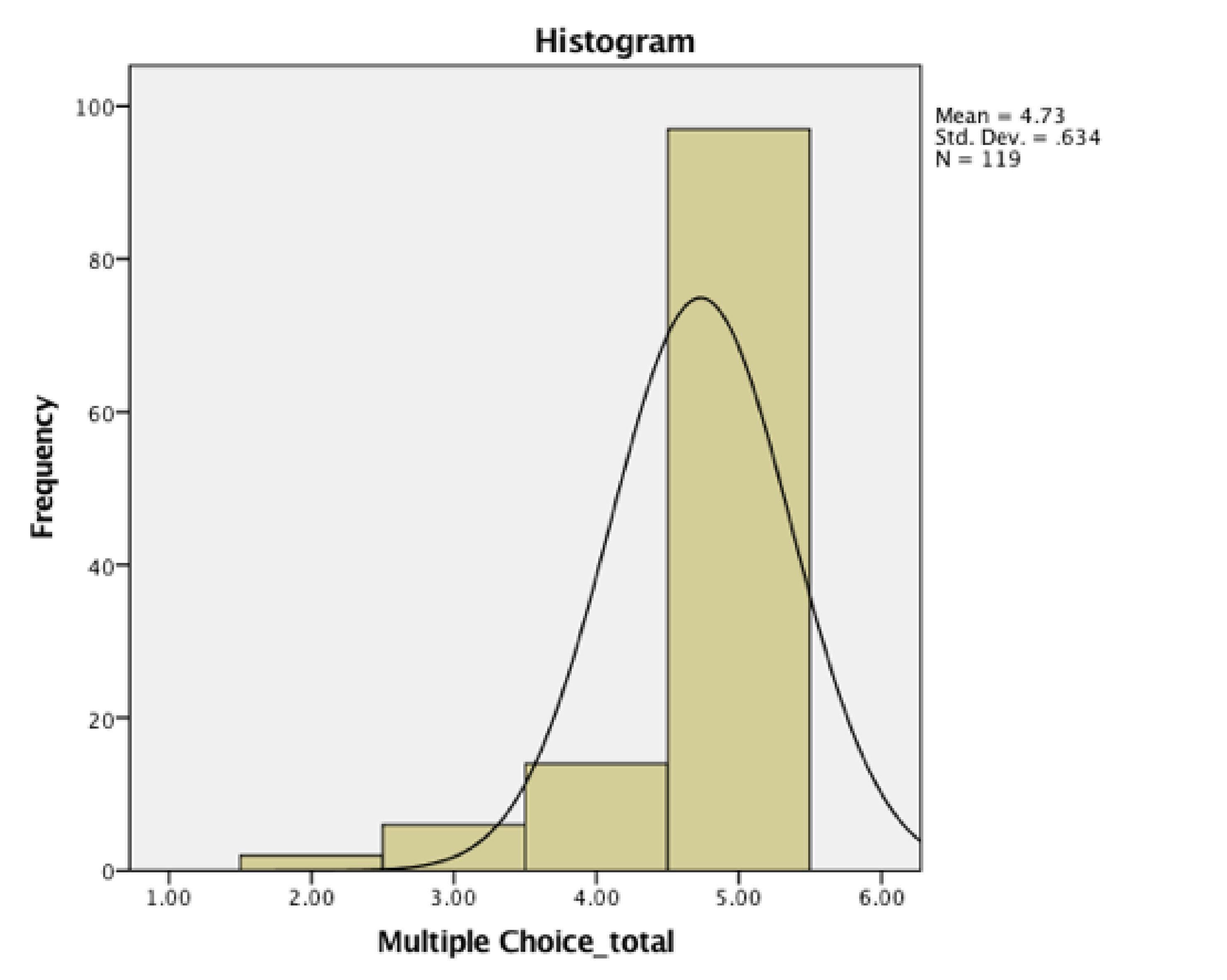
- "It seemed most suited for a university or academic kind of setting and I’m not coming from that place really.”
- “I as a broadcaster was a bit disappointed with the sound quality of the audio.”

Method 4: Course evaluation with students (UNITS)

- Core Video (Module 4 Unit 3: AD for static arts)
- Participants 119 Italian students studying Theory and history of translation with a very basic knowledge of AD.
- Method: watch the Core Video.
- Complete a Multiple Choice (MC) task.
- Fill in the evaluation form.

Results: The contents (are…)

- Clearly presented **m= 4, s.d.= .64, mode = 4**
- Effectively organised **m = 4, s.d. = .56, mode = 4**
- Easy to understand m = 3.9, s.d. = .68, mode = 4
- Provide appropriate guidance on the topic m = 3.78, s.d. = .78, mode = 4
- Increased my knowledge of the topic m = 3.9, s.d. = .71, mode = 4
- Developed my skills in this subject **m = 4.0, s.d. = .79, mode = 4**



Results (MC)

97 (81.5%) students got 5/5;
14 (11.8%) got 4 /5;
6 (5%) got 3 /5;
2 (1.7%) got 2 /5.
No one got 1 or 0 out of 5

Possible interpretations

- The Core Video is fairly comprehensible.
- The video is an excellent teaching tool.
- The multiple choice is too easy. Especially Q4 (100% correct).

Core Videos: Results Qual

- Liked most/least:
- Positive Keywords
- clear, good organization and/or structure, easy to understand, increase of knowledge, interesting content and/or topic, short, easy vocabulary; pleasant voice and/or presentation, good pace

- Negative Keywords
- boring, no pictures, not enough examples, too fast pace, graphics, layout, rather bad audio quality

Method 5: student/professional comparison

- Participants:
- Six students AMU, Poland: Range of familiarity with AD: 1 = 1 AD = completely new to me -5 = 2 (extremely familiar).
- Four experienced audio describers at RTV Slovenija.

- AMU: CV_M2_U2: (Process)
- RTV-SLO: CV_M3_U3 (what to describe in a live performance.)
- CV_M6_U2 (the technology used to deliver AD).

QIs

- A 7-point Likert scale was used to measure 8 QIs: attention; comprehension; accuracy of recall (general); accuracy of recall (specific); how easy the core videos were to follow and views on pace; interest; difficulty.

- All the points were labelled. A 9-point unlabelled scale was used for mental effort. This was to conform with the evaluation of cogniive load (CL) by Paas et al. (2003)

RTV-SLO Results

Did you think your comprehension was… (1 = “very poor”; 7 = “very good”): Min = 5; Max = 7; **mode = 6 (good)**; mean = 6; SD = .63

- How accurately can you remember general information? (1 = “not at all”; 7 = “extremely”): Min = 3; Max = 7; **mode = 5 (very)**
- How accurately can you remember specific information ? (1 = “not at all”; 7 = “extremely”): Min = 4; Max = 7; **mode = 5 (very)**

RTV-SLO Results

- How easily were you able to follow? (1 = “not at all”; 7 = “extremely”): Min = 4; Max = 5; mode = 4 (neither with difficulty nor easily)
- **Pace** (1 = “very fast”; 7 = “very slow”): Min = 4; Max = 4; mode = 4 (neither fast nor slow)
- **Interest** (1 = “very boring”; 7 = “very exciting”): Min = 4; Max = 4; mode = 4 (neither exciting nor boring)
- **Difficulty** (1 = “very difficult”; 7 = “very easy”) Min = 4; Max = 4; mode = 4 (neither with difficulty nor easily)

Cognitive Load: RTV-SLO Results

- How much mental effort did you put into following the core videos? (1 = “minimal effort”; 9 = “extreme effort”):
- Min = 5; Max = 6; mode = 5

Comparison AMU – RTV-SLO

- independent samples t-tests
- Significant differences were found for:
- pace F (4,6) = 1.265, p= .001
- mental effort F (4,6) = 6.741, p= .032
- professional describers finding the pace slower, yet feeling they had to expend more mental effort than the students.

Liked best/least RTV-SLO

The thing I liked best …

- Theoretical concepts are explained on practical examples
- Structure, content
- Systematically organized and structured materials
- Very good materials to use in different courses

There were no negative comments.

Method 6: Qualitative Focus Group (RNIB)

- 5 PSL attended. All were either employees of RNIB, or worked there as a volunteer. All identified as blind or partially sighted.

- Two identified as trainers, giving talks to community groups and to staff at Transport for London and Transport for All.

- All engaged in this type of advocacy e.g. with local societies for PSL.

Process

- CV_M1_U6 (Target audience for AD)
- Participants watched the video.
- Invited to make comments.
- Audio recording.
- Transcribed by app. Refined by the interviewer.
- Full transcript in the Evaluation Guide.

Results: Target Audience

- I think personally it's a really, it was a really good video that describes what audio description is and all who can benefit. (P02)
- I think the emphasis should be more, well not *should* be but some more emphasis *could* be put on blind people. (P02)

Memorability: Results

- Memorability: Measure of success of training materials and as a direct variable in the evaluations given by students

- I'm sitting here all evening and that figure of two hundred and fifty three million across the world. Yeah I just thought that was… that's unbelievable. I just, I know there's not as many people getting the support that we get. Yeah. And it's just really sort of quite tragic. (P01)

To sum up

- Formative & summative evaluation
- Mixed methods
- Intermethod discrepancies
- Quite consistent between evaluators & methods.

- Positives for Core videos: clarity; easy to understand; quite easy to remember – especially the general information. Useful to trainers. As training materials, they're pretty good.
- Negatives: as AV-productions they're not great.

A Guide to the Evaluation of Training Materials: ADLAB PRO a Case Study

Available to download from the ADLAB PRO website.

https://www.adlabpro.eu

Final comment from one of the ME5 interviewees

- This is good stuff for anybody anywhere.