When Do We Have Enough Evidence to Accept Migration of Patient-Reported Outcome Measures (PROMS) from Paper to Screen-based Formats without Additional Testing?

ISPOR European Congress 2018

Barcelona, Spain

November 14, 2018

CRF HEALTH

WILLIE MUEHLHAUSEN CONSULTING

PRO

Disclaimer

- The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies or the Critical Path Institute.
- These slides are the intellectual property of the individual presenters and are protected under the copyright laws of the United States of America and other countries. Used by permission. All rights reserved. All trademarks are the property of their respective owners.

Presenters

- *Bill Byrom, PhD*
  - Vice President of Product Strategy and Innovation, CRF Bracket
- *Sonya Eremenco, MA*
  - Associate Director, Patient-Reported Outcome (PRO) Consortium, Critical Path Institute
- *Willie Muehlhausen, DVM*
  - Managing Director, Muehlhausen Ltd

Presentation Outline and Objectives

When Do We Have Enough Evidence to Accept Migration of Patient-Reported Outcome Measures (PROMS) From Paper to Screen-based Formats without Additional Testing?

- Introduction
- Presentation of key works
- Faithful migration best practices
- Discussion: case examples

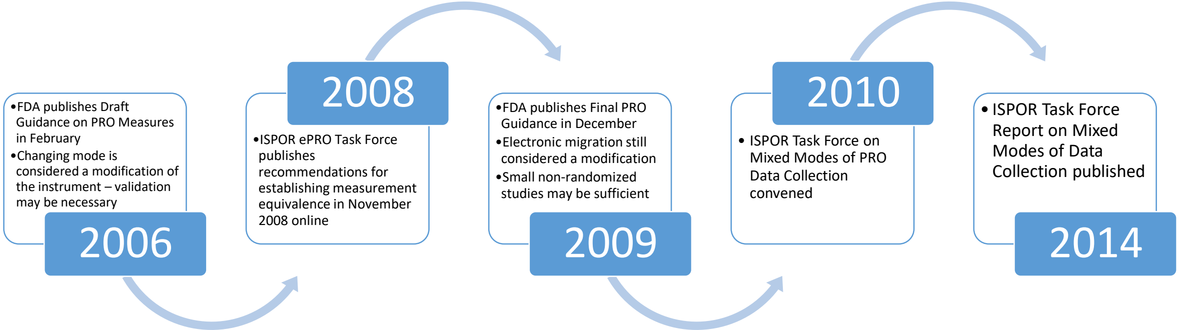
Introduction

Sonya Eremenco, PRO Consortium, Critical Path Institute

Set the scene regarding electronic migration

- How did we get here?
- What are the current recommendations for evaluating equivalence?
- Where do we go next?

Brief history of migration/equivalence recommendations



2006

- FDA publishes Draft Guidance on PRO Measures in February
- Changing mode is considered a modification of the instrument – validation may be necessary

2008

- ISPOR ePRO Task Force publishes recommendations for establishing measurement equivalence in November 2008 online

2009

- FDA publishes Final PRO Guidance in December
- Electronic migration still considered a modification
- Small non-randomized studies may be sufficient

2010

- ISPOR Task Force on Mixed Modes of PRO Data Collection convened

2014

- ISPOR Task Force Report on Mixed Modes of Data Collection published

ISPOR ePRO Task Force Report Recommendations

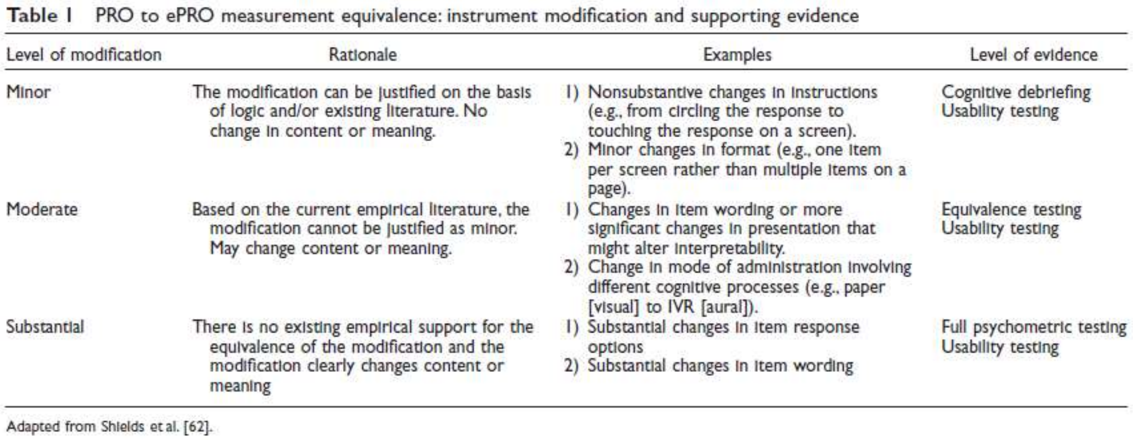


Table 1 PRO to ePRO measurement equivalence: instrument modification and supporting evidence

Level of modification	Rationale	Examples	Level of evidence
Minor	The modification can be justified on the basis of logic and/or existing literature. No change in content or meaning.	1) Nonsubstantive changes in instructions (e.g., from circling the response to touching the response on a screen). 2) Minor changes in format (e.g., one item per screen rather than multiple items on a page).	Cognitive debriefing Usability testing
Moderate	Based on the current empirical literature, the modification cannot be justified as minor. May change content or meaning.	1) Changes in item wording or more significant changes in presentation that might alter interpretability. 2) Change in mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]).	Equivalence testing Usability testing
Substantial	There is no existing empirical support for the equivalence of the modification and the modification clearly changes content or meaning	1) Substantial changes in item response options 2) Substantial changes in item wording	Full psychometric testing Usability testing

Adapted from Shields et al. [62].

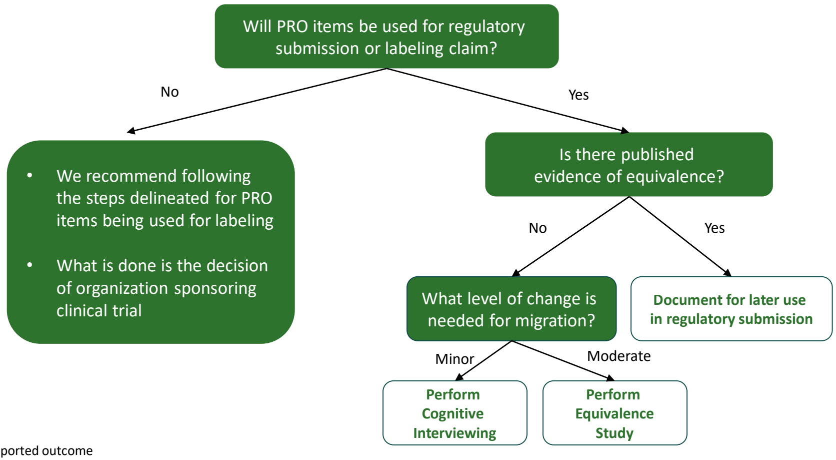
Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on Evidence Needed to Support Measurement Equivalence between Electronic and Paper-Based Patient-Reported Outcome (PRO) Measures: ISPOR ePRO Good Research Practices Task Force Report. *Value Health*. 2009;12(4):419-429.

ISPOR Mixed Modes Task Force Recommendations

1. Select appropriate mode(s) for trial
2. Perform a “faithful migration” (“migrate before you mix”)
   - Only necessary changes to the format and instructions are made and that the content of the items and responses has not changed.
   - Subjects ***interpret and respond*** to the questions/items the same way regardless of mode
3. Evaluate equivalence between the modes migrated and/or to be mixed
   - Use appropriate study design
4. If above conditions are met, implement the mode or modes in the trial
   - Avoid mixing paper and electronic diaries; assess risks of other combinations
   - If deciding to mix other modes
     - Plan and implement carefully; mix at country level or higher
     - Assess statistical issues and poolability of data

Eremenco S, Coons SJ, Paty J, et al. PRO data collection in clinical trials using mixed modes: report of the ISPOR PRO mixed modes good research practices task force. *Value Health*. Jul 2014;17(5):501-516.

Need to Establish Measurement Equivalence



Will PRO items be used for regulatory submission or labeling claim?

No

- We recommend following the steps delineated for PRO items being used for labeling
- What is done is the decision of organization sponsoring clinical trial

Yes

Is there published evidence of equivalence?

No

What level of change is needed for migration?

Minor

Perform Cognitive Interviewing

Moderate

Perform Equivalence Study

Yes

Document for later use in regulatory submission

PRO, patient-reported outcome

Qualitative Study Design: Cognitive Interview

- Purpose: to evaluate if the migration has impacted how subjects *interpret and respond* to the items
  - Not intended to revisit content validity of the original instrument
- Minor modifications to format or procedure
- Small sample size: 5 to 10 subjects
- Assess usability of instrument as a secondary goal
- Variations in study design include:
  - Whether patients complete both modes during interview
  - How responses can be compared
  - Whether multiple interview rounds are necessary to allow for revising/ retesting
- Key questions answered:
  - Why interpretation between modes may differ
  - Why responses between modes may differ

“New” Literature on Equivalence

- EuroQol 5-Dimension questionnaire (EQ-5D): IVR and Paper
  - Lundy JJ, Coons SJ. Measurement equivalence of interactive voice response and paper versions of the EQ-5D in a cancer patient sample. *Value Health*. 2011;14(6):867-871.
- EORTC: IVR and Paper
  - Lundy JJ, Coons SJ, Aaronson NK. Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. *Qual Life Res*. 2014;23(1):229-237.
- PROMIS Physical Function, Fatigue, Depression banks: personal computer (PC) vs. IVR, personal digital assistant (PDA), Paper, or PC
  - Bjorner JB, Rose M, Gandek B, et al. Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. *J Clin Epidemiol*. 2014;67(1):108-113.
- Reviews of paper vs. electronic studies
  - Campbell N, Ali F, Finlay AY, Salek SS. Equivalence of electronic and paper-based patient-reported outcome measures. *Qual Life Res*. 2015;24(8):1949-1961.
  - Rutherford, C., Costa, D., Mercieca-Bebber, R., Rice, H., Gabb, L. & King, M. Mode of administration does not cause bias in patient-reported outcome results: a meta-analysis. *Quality of Life Research*. 2016 Mar;25(3):559-74.
- Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE): Web, IVR and Paper
  - Bennett et al. *Health Qual Life Outcomes.* 2016; **14**:24. https://doi.org/10.1186/s12955-016-0426-6
- Bowel function instrument, linear analog scale assessment (LASA) quality-of-life (QOL) and Adapted Sydney Swallow Questionnaire (SSQ): Web, IVR and Paper
  - Bennett et al. *Qual Life Res*. 2016 May;25(5):1123-30. doi: 10.1007/s11136-015-1162-9.
- Bring your own device (BYOD)
  - Coons SJ, Eremenco S, Lundy JJ, et al. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. *Patient*. 2015;8(4):301-309.
  - Gwaltney C, Coons SJ, O’Donohoe P, O’Gorman H, Denomey M, Howry C, Ross J. “Bring Your Own Device” (BYOD): The future of field-based patient-reported outcome data collection in clinical trials? *Ther Innov Regul Sci.* 2015 Nov;49(6):783-791. doi: 10.1177/2168479015609104.

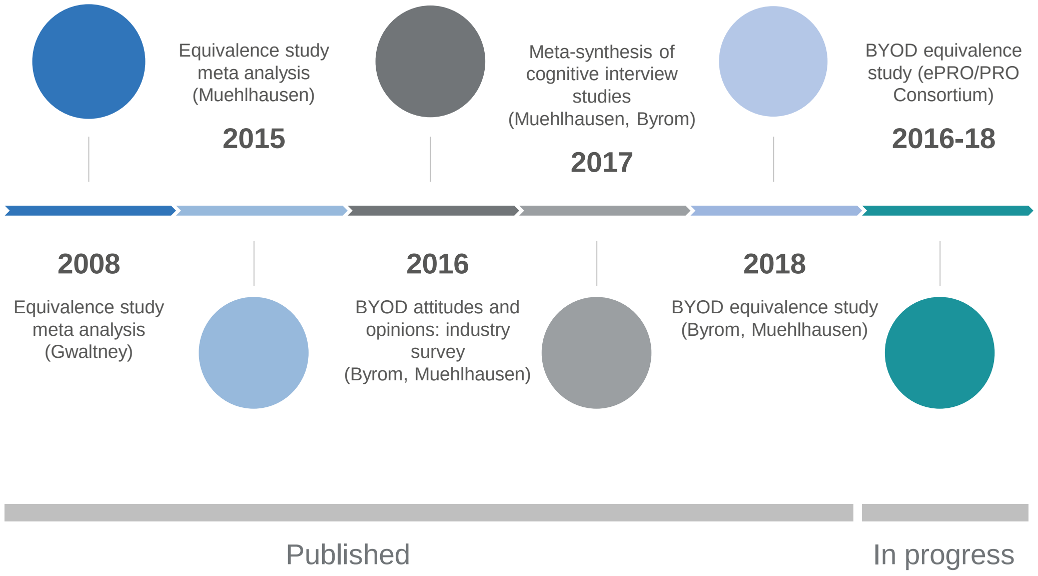
Where Do We Go Next?

- What has changed since 2014?
  - Hundreds of unpublished qualitative migration studies conducted confirming equivalence
  - Usability issues are the more salient results
- Bring Your Own Device (BYOD) is becoming mainstream
  - Mixing is inherent in BYOD implementations
  - Not feasible to conduct equivalence studies among all possible devices
- Industry views "equivalence studies" as a requirement when implementing clinical outcome assessments electronically because of regulatory uncertainty
- A new ISPOR ePRO Task Force will update the previous recommendations
  - Outline the evidence required to ensure a faithful migration and suggest when, in light of the accumulated evidence, additional testing is not required
  - Identify aspects of instrument migration or study design that may jeopardize compatibility between modes or impact the operational integrity of the study
  - Explore the role of feasibility testing

Key Works

Bill Byrom, CRF Bracket

Growing body of evidence



2008

Equivalence study meta analysis (Gwaltney)

Equivalence study meta analysis (Muehlhausen)

2015

2016

BYOD attitudes and opinions: industry survey (Byrom, Muehlhausen)

Meta-synthesis of cognitive interview studies (Muehlhausen, Byrom)

2017

2018

BYOD equivalence study (Byrom, Muehlhausen)

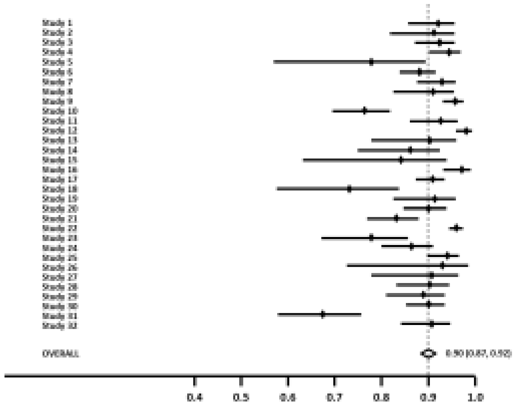
BYOD equivalence study (ePRO/PRO Consortium)

2016-18

Published

In progress

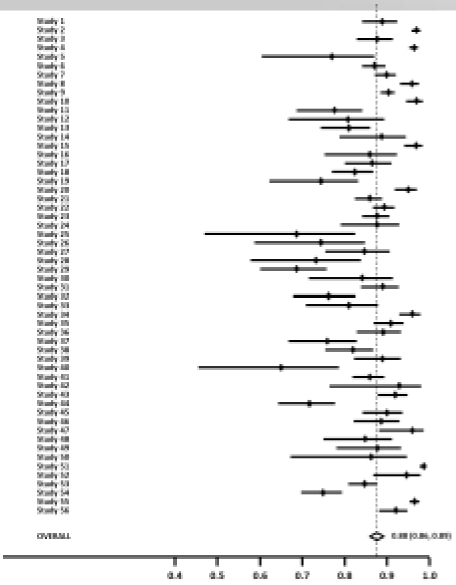
Meta-analysis: Gwaltney et al. 2008



Gwaltney et al. (2008)

- 46 studies
- 278 PROM comparisons
- Paper vs. PC/handheld device
- Study n: 10 – 189
- **Pooled correlation coefficient: 0.90 (95% CI: 0.87–0.92)**

Meta-analysis: Muehlhausen et al. 2015



Muehlhausen et al. (2015)

- 72 studies
- 152 PROM comparisons
- Paper vs. PC, handheld device, IVRS
- 23 patient populations
- Ages: 6 – 68 years
- **Pooled correlation coefficient: 0.875 (95% CI: 0.867 to 0.884)**

"PROMs administered on paper are quantitatively comparable with measures administered on an electronic device" across multiple scales and patient groups."

Meta synthesis of cognitive interview and usability studies

ScienceDirect

ELSEVIER

Standards for Instrument Migration When Implementing Paper Patient-Reported Outcome Instruments Electronically: Recommendations from a Qualitative Synthesis of Cognitive Interview and Usability Studies

– Muehlhausen, Byrom, Skerritt et al. (2017)
  - All studies conducted by ICON from 2012 to 2015
  - 53 studies
  - Wide range of patient populations including:
    - Respiratory, Gastrointestinal, Oncology, Central Nervous System disorders, Rheumatology, Cardiovascular disorders, Dermatology, Gynaecology, Infectious disease, Metabolic, Urology, Vaccines.
  - 68 instruments
  - 101 PROM comparisons
  - Response scale types included: visual analogue scale (VAS), verbal rating scale (VRS), numeric rating scale (NRS), EQ-VAS (from EQ-5D)

Meta synthesis of cognitive interview and usability studies

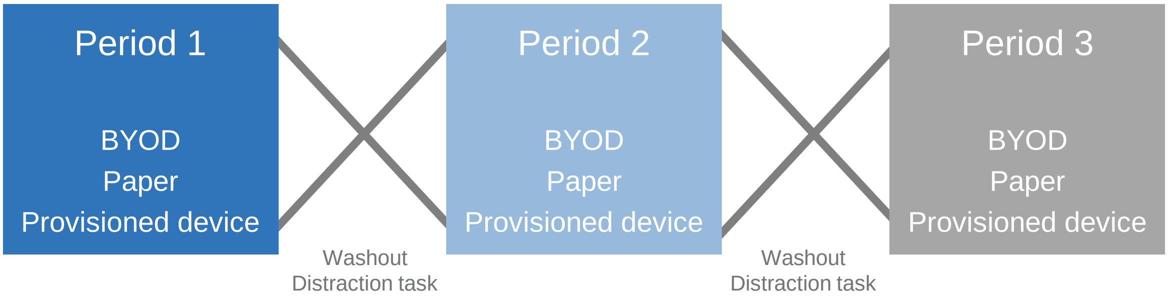
ELSEVIER

ScienceDirect

Standards for Instrument Migration When Implementing Paper Patient-Reported Outcome Instruments Electronically: Recommendations from a Qualitative Synthesis of Cognitive Interview and Usability Studies

"With the benefit of accumulating evidence, it is possible to **relax the need to routinely conduct** cognitive interview and usability studies when implementing minor changes during instrument migration. Application of **design best practice** and selecting vendor solutions with good user interface and user experience properties that have been **assessed for usability in a representative group** may enable many instrument migrations to be accepted without formal validation studies by instead conducting a structured expert screen review."

Equivalence with variable screen size (BYOD)

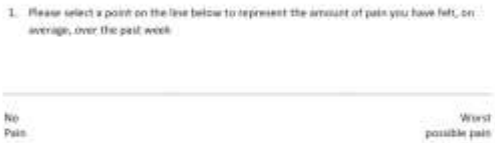


- **156 subjects**
  - 19 to 69 years old (48.6 ± 13.1)
  - Female: 83 (54%)
  - Male: 72 (46%)
  - Conditions resulting in chronic pain
  - Broad range of educational backgrounds

- **SF-20**
  - VRS
  - Y>3, Y<3, N
  - Likert
- **VAS and NRS-11 (pain)**

Equivalence with variable screen size (BYOD)

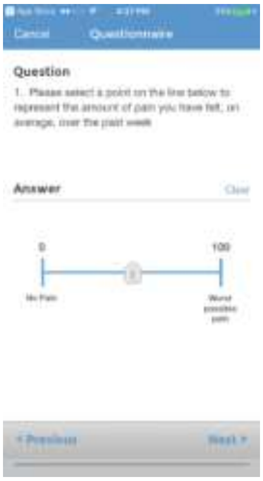
Paper



1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

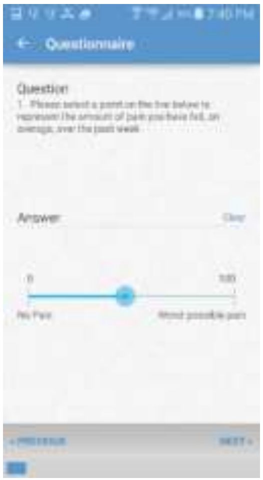
No Pain — Worst possible pain

iOS



1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

Android



1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

Equivalence with variable screen size (BYOD)

- Very high correlation between the **three modes** of administration:
  - ICCs: 0.816 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between **paper and BYOD**
  - ICCs: 0.806 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between **site device and BYOD**
  - ICCs: 0.791 to 0.966
  - Lower bound of the 95% confidence interval > 0.70

- Very high correlation between the **three modes** of administration for each response scale type:
  - VRS: ICC: 0.97 (0.96 – 0.98)
  - NRS: ICC: 0.98 (0.97 – 0.98)
  - VAS: ICC: 0.94 (0.91 – 0.95)

Good summary of all the evidence

DIA

Measurement Equivalence of Patient-Reported Outcome Measures Migrated to Electronic Formats: A Review of Evidence and Recommendations for Clinical Trials and Bring Your Own Device

Bill Byrom, PhD[1], Chad Gwaltney, PhD[2], Ashley Slagle, PhD[3], Ari Gnanasakthy, MBA[4], and Willie Muehlhausen, DVM[5]

Abstract

[illegible]

Keywords

[illegible]

Introduction

[illegible]

Byrom B, Gwaltney C, Slagle A, Gnanasakthy A, Muehlhausen W (2018)
Therapeutic Innovation and Regulatory Science.
https://doi.org/10.1177/2168479018793369

Thoughts:
Faithful Migration Best Practices

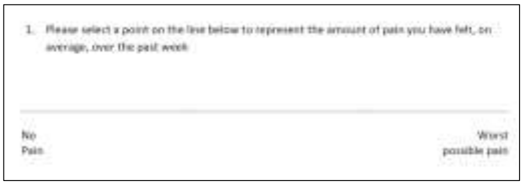
Willie Muehlhausen, Muehlhausen Ltd.

Why are we doing this?

- Clinical Trials are including more patient input (Patient-centric)
- Virtual Clinical trials
- Real World Evidence
- Patient Care and Remote Monitoring
- Bring Your Own Device (BYOD)

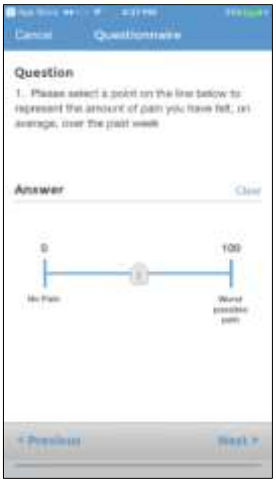
Why are we doing this?
Differences in presentation

Paper



1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

iOS

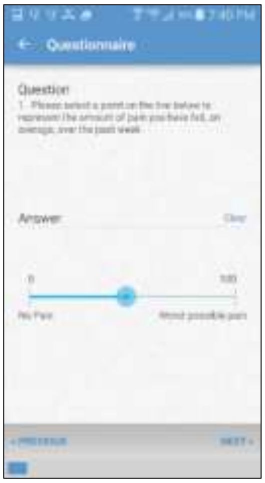


Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

Answer

Android



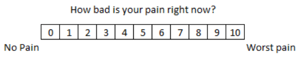
Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

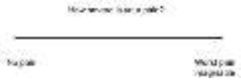
Answer

Instrument “Controls” / “Widgets”

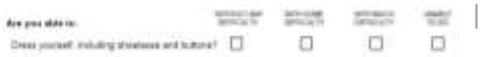
- Most instruments are composed of a small number of controls/widgets
  - Numeric Rating Scale (NRS)



  - Visual Analogue Scale (VAS)



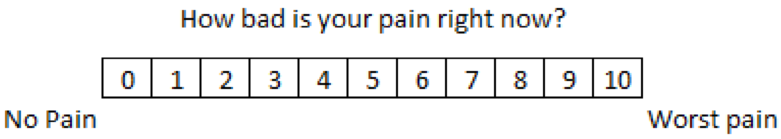
  - Verbal Rating Scale (VRS)



Instrument “Controls” / “Widgets”

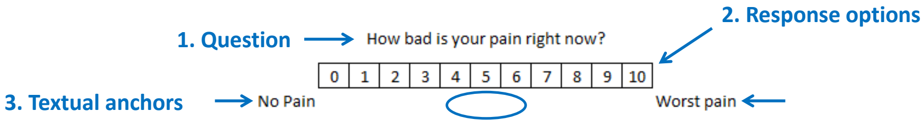
Definition:

- A **graphical control element** or **widget** is an element of interaction in a graphical user interface (GUI), such as a button or a scroll bar.
- A “Control” or “Widget” is an interface element (e.g., numeric rating scale)



Instrument “Controls” / “Widgets”

- Can we validate a control/widget independent of the context it applies to?
- Example: numeric rating scale components



1. Question

How bad is your pain right now?

0	1	2	3	4	5	6	7	8	9	10

2. Response options

3. Textual anchors

No Pain

Worst pain

- Can we appeal to previous validation of this widget when considering the need to perform future equivalence testing?

Meta synthesis of cognitive interview and usability studies

ScienceDirect

Standards for Instrument Migration When Implementing Paper Patient-Reported Outcome Instruments Electronically: Recommendations from a Qualitative Synthesis of Cognitive Interview and Usability Studies

- Muehlhausen, Byrom, Skerritt et al. (2017)
  - All studies conducted by ICON from 2012 to 2015
  - 53 studies
  - Wide range of patient populations including:
    - Respiratory, GI, Oncology, CNS, Rheumatology, CV, Dermatology, Gynaecology, Infectious disease, Metabolic, Urology, Vaccines.
  - 68 instruments
  - 101 PROM comparisons
  - Response scale types included: VAS, VRS, NRS, EQ-VAS

Equivalence with variable screen size (BYOD)

- Very high correlation between the **three modes** of administration:
  - ICCs: 0.816 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between **paper and BYOD**
  - ICCs: 0.806 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between **site device and BYOD**
  - ICCs: 0.791 to 0.966
  - Lower bound of the 95% confidence interval > 0.70

- Very high correlation between the **three modes** of administration for each response scale type:
  - VRS: ICC: 0.97 (0.96 – 0.98)
  - NRS: ICC: 0.98 (0.97 – 0.98)
  - VAS: ICC: 0.94 (0.91 – 0.95)

ScienceDirect

PATIENT-REPORTED OUTCOMES

Measurement Equivalence of Patient-Reported Outcome Measure Response Scale Types Collected Using Bring Your Own Device Compared to Paper and a Provisioned Device: Results of a Randomized Equivalence Trial

Best Practices

Best Practices for Electronic Implementation of Response Scales for Patient-Reported Outcome Measures

Best Practices for Maximizing Electronic Data Capture Options during the Development of New Patient-Reported Outcome Measures

Best Practices for Migrating Existing Patient-Reported Outcome Measures to a New Data Collection Mode

ELECTRONIC PATIENT-REPORTED OUTCOME MEASURES: AN IMPLEMENTATION HANDBOOK FOR CLINICAL RESEARCH

Bill Byrom & Willie Muehlhausen

Recommendations

- Keep it simple!!
- Text Art only when proven beneficial (i.e., bold, italic, underline, capitals)
- Colour only when proven beneficial
- One item per screen (even if there is space for more on the tablet)
- Use of neutral verbiage during development (i.e., “Select” vs. “Circle”)
- Use basic widgets and avoid creative combinations

- Then:
  - Conduct an Expert Screen Review and possibly Usability Testing only

Recommendations

ePRO Design Best Practice *

- Provide robust instructions on use of the application.
- Ensure font size is suitable, clear, and readable.
- Present a single question and response scale option per screen.
- Take care not to modify the original instrument text beyond minor changes.
- Precede question with instructional text screen if cannot be presented together.
- Ensure equal screen area, font and line spacing used for each response option.
- Use indicator arrows to identify the location of anchor text if needed.
- Present the recall period with each item as opposed to only in initial instrument instructions.

* Consolidated from Critical Path Institute’s ePRO Consortium Recommendations and ICON research

Usability

- Usability should cover the app and all common widgets
- Usability evidence from representative groups is sufficient
  - Patients or healthy volunteers
    - Similar age range to target population, range of educational and socioeconomic backgrounds
  - Additional representative groups may include (as needed)
    - children/adolescents,
    - dexterity-challenged subjects
    - technology-naïve subjects (e.g., very elderly subjects)
    - cognitively challenged subjects
    - partially sighted subjects.

Expert Screen Review

1. Overall instructional information
   - Instrument and application instructions
   - Recall period representation
   - Author-specific requirements
2. Usability, including font size and navigation
   - Clarity, colour and font size per screen
   - Consistency, visibility and size of controls
   - Device-orientation changes
   - Back and forward navigation
   - End of questionnaire review (where included)
3. Item-by-item migration review
   - Single item fully visible
   - Recall period per item understood
   - No changes to core wording
   - Consistent use of bold and underlining where needed
   - Question skipping capability consistent with instrument requirements
   - VRS/NRS equally spaced and sized
   - VAS sufficient space at sides
   - Anchor text location clear

Next steps

- Request for help:
  - Screenshots of instruments in published/unpublished
    - Equivalence studies
    - Cognitive interview/usability studies
    - Expert Screen Review
- Share Experience with Regulators
  - Bring Your Own Device
- Distribute the C-Path Best Practice documents and use them!
  - c-path.org/programs/epro

Discussion: Case Examples

All presenters

Example 1: SF-36

- Current evidence
  - 7 studies within Gwaltney et al. 2008
  - 9 studies within Muehlhausen et al. 2015
  - 3 studies in meta-synthesis, 2018
  - Equivalence of VRS (BYOD study, 2018)
  - Meta-analysis of 25 studies: SF-36 only (White et al., 2018)
  - Unpublished CI/UT studies on various vendor platforms

1. Is there enough evidence to not require additional testing?
2. If so, what conditions would be required for this to be the case?
3. How would we package the evidence to support migration comparability?

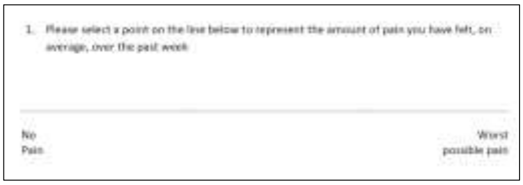
Example 2: Instrument in different population

- Current evidence
  - Demonstrated equivalence in population 1

- Required evidence
  - Evidence to support measurement equivalence in population 2

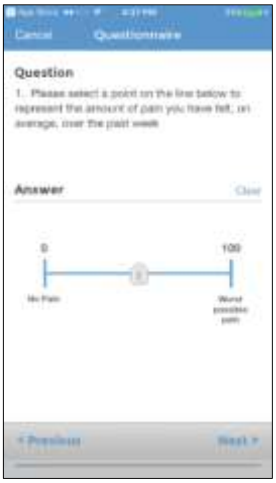
1. Examples of populations that would not require additional evidence
2. Examples of populations that would require additional evidence

Example 3: Visual analogue scale Baseline



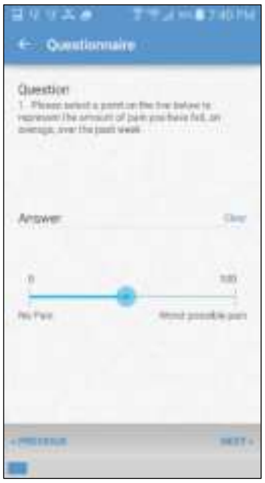
1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

No Pain
Worst possible pain



Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

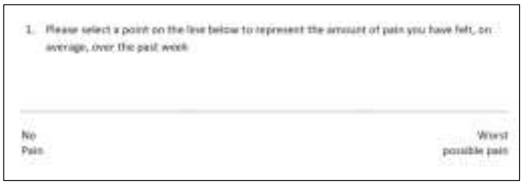


Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

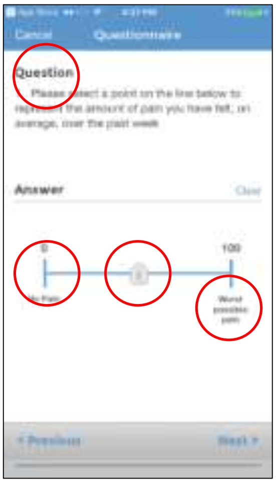
- Keep it simple and consistent
- Spot the difference?!

Baseline:



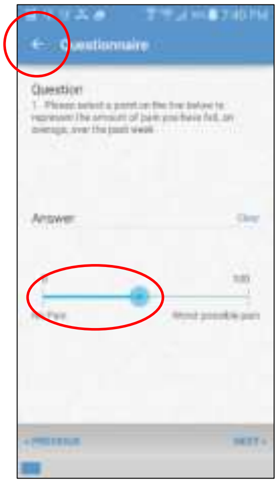
1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

No Pain
Worst possible pain



Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

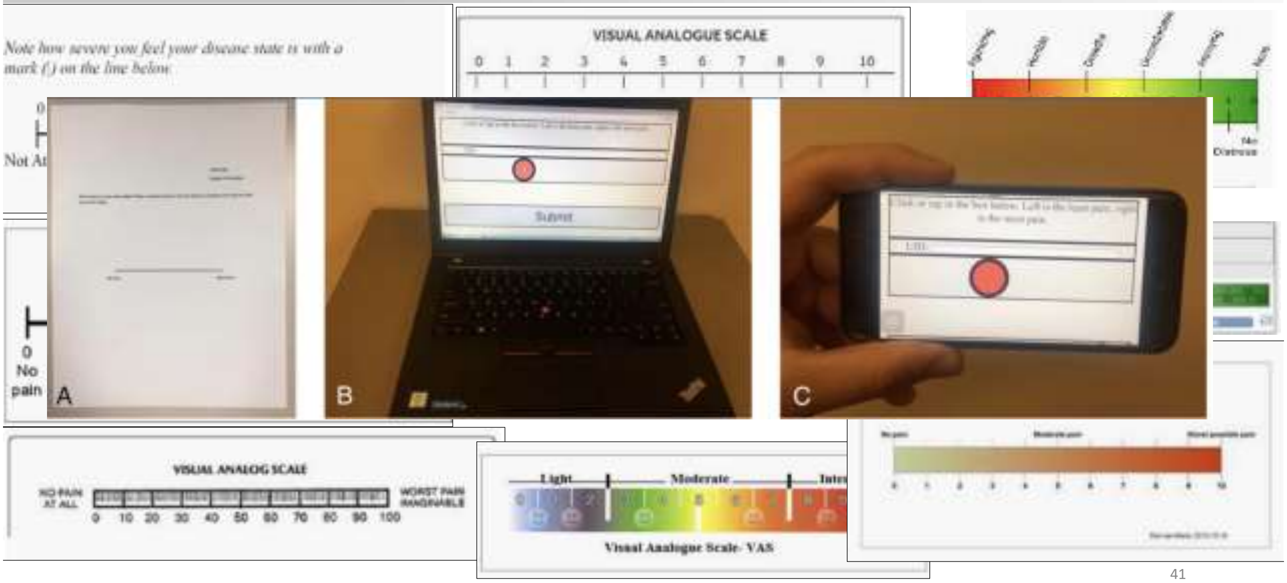


Question

1. Please select a point on the line below to represent the amount of pain you have felt, on average, over the past week

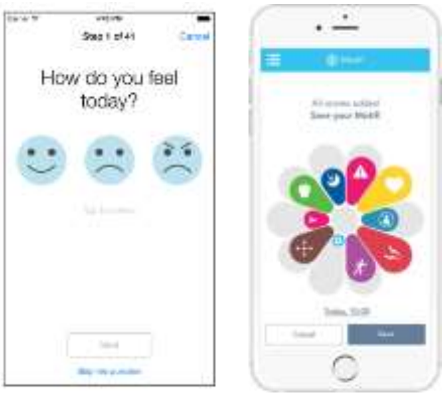
- Keep it simple and consistent
- Spot the difference?!

How do you think I feel about these?



Example 4: New instrument w/without standard widgets

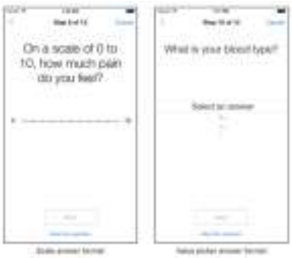
Examples:



1. What evidence would be needed to demonstrate migration equivalence?

2. How could this evidence be generated and reported to enable its re-use to support other studies?

Example 5: Apple Research Kit or non-standard response scale types



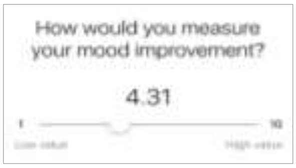
On a scale of 0 to 10, how much pain do you feel?

What is your blood type?

Select an answer

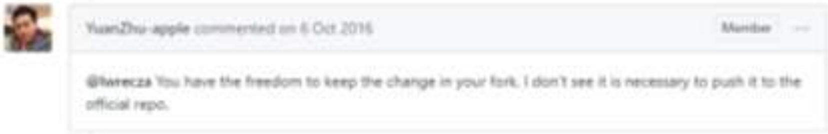
Scale answer format

Value picker answer format



How would you measure your mood improvement?

4.31



YuanZhu-apple commented on 6 Oct 2016

@lwrecza You have the freedom to keep the change in your fork. I don't see it is necessary to push it to the official repo.

1. What evidence would be needed to demonstrate migration equivalence?

2. How could this evidence be generated and reported to enable its re-use to support other studies?

Example 6: New vendor platform

- What you would need to do if it was completely a new platform
- Usability: screen review vs. data entry
- Confirm best practices are being followed
- Button vs. selecting answer

Thank you

@billbyrom
@crfhealth

@WMePRO

@CPathInstitute

CRF
HEALTH

WILLIE
MUEHLHAUSEN
CONSULTING

PRO