Measurement Comparability Between Paper and Alternate Versions: Recommended Assessment Steps Using the Lung Function Questionnaire as an Example

Anand Dalal,[1] Lauren Nelson,[2] Theresa Gilligan,[2] Lori McLeod,[2] Sandy Lewis,[2] Carla DeMuro[2]

[1]GlaxoSmithKline, Durham, NC, United States; [2]RTI Health Solutions, Research Triangle Park, NC, United States

BACKGROUND

- Providing participants with choices in how their data are collected may lead to greater participation, less missing data, improved data quality, and in some cases, decreased costs in data collection.

OBJECTIVE

- To provide recommended steps to assess measurement comparability among different versions of the same questionnaire using a crossover study design and a case-finding questionnaire, the Lung Function Questionnaire (LFQ), as an example.

METHODS

LFQ

- Five-item questionnaire developed using questions from the third National Health and Nutrition Examination Survey (NHANES III).
- The instrument measures patient perception of breathing problems and activity limitation.
- The five items are summed to create a total LFQ score, which can range from 5 to 25; lower scores indicate risk of obstruction.
- The LFQ was developed as a paper (P)-based tool and validated in a cross-sectional study.[1,2]
- To promote widespread use of the LFQ, three additional versions were developed: Web (W) based, interviewer (I) based, and IVRS based.
- Participants also completed demographic and health questions, and a short questionnaire regarding their administration preference.

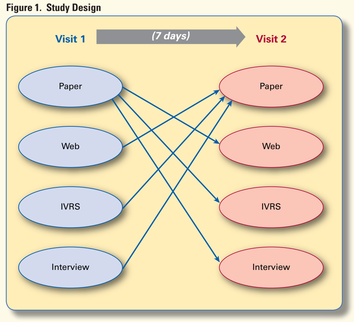
Design

- Participants were 40 years of age or older; self-reported current or former smokers (defined as ≥ 10 pack years); able to provide informed consent; able to read and understand English; and did not have a diagnosis of chronic obstructive pulmonary disease, emphysema, or asthma.
- A two-visit, crossover design was employed (Figure 1).
  - Participants were randomly assigned to one of six sequence groups based on the LFQ version completed and order of administration (i.e., P-W, P-IVRS, P-I, W-P, IVRS-P, and I-P) at two visits.

Sample Size Justification

- Crossover designs greatly reduce the sample size required, because subjects serve as their own controls, reducing variability.[3]
- Because the LFQ was developed as a case-finding tool for screening patients, the sample size for this example was based on the intraclass correlation coefficient (ICC).[3-5]

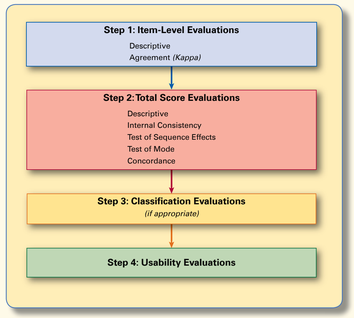
Figure 1. Study Design



Guideline for Evaluation

- The steps recommended for assessing comparability for minor to moderate levels of instrument modification under crossover designs are shown in Figure 2.
  - Adaptations from the P-based version to the W-based version were considered minor, because they were both self-administered and had identical items.[3]
  - Modifications from the P-based version to the I- and IVRS-based versions were considered moderate because different cognitive processes are required from the P-based version (i.e., visual versus auditory).[3]

Figure 2. Evaluation Guide: Recommended Steps for Assessing Measurement Comparability for Crossover Designs



Step 1. Item-Level Evaluations

- Descriptives
  - Response frequency distributions and descriptive statistics were examined for all LFQ items for each of the six sequence groups and combined across the order of administration.
- Item-level agreement
  - Weighted kappas were computed by the six sequence groups and for the three pair groups (combined over sequence). A kappa ranging from 0.21 to 0.41 is considered poor to fair, 0.41 to 0.60 is moderate, 0.61 to 0.80 is substantial, and over 0.81 is nearly perfect.[6,7]

Step 2. Total Score Evaluations

- Total score agreement was evaluated based on descriptive comparisons across all sequence groups.
- Internal consistency reliability: Because the LFQ is a cumulative risk index, items were not expected to be highly correlated; therefore, evaluating the internal consistency reliability is not as appropriate for the comparison across versions.
- Test for sequence effects (i.e., order or carry-over effect): A t-test was performed to compare the mean difference of LFQ scores by the sequence (e.g., P-W versus W-P). If there was no statistical evidence for sequence effects at $P < 0.05$, then the groups were combined.
- Test of mode: Following a nonsignificant sequence effect, paired t-tests provided further evidence that the score distributions between two measures were similar, using a P value ≤ 0.05 as evidence that the difference in the means were statistically different from zero.
- Concordance: Separate ICC estimates were computed for each combined pair group (i.e., P/W, P/IVRS, P/I) and compared with previous GSK estimates of the test-retest reliability for the P-based version of the LFQ.[8]

Step 3. Classification Evaluations

- A cut score of 18 was applied to the scores on both the P-based and the alternate version, and the percentage agreement in classification (i.e., likely obstructed versus not likely obstructed) was computed.
- The kappa statistic was computed as a measure of agreement.[6]

Step 4. Usability Evaluation

- Each participant was asked to provide feedback on the P-based version of the LFQ as well as the alternate version they completed.
- Questions assessed any difficulty experienced when completing the questionnaire, ranging from 0 (Not at all) to 10 (Extremely), and a rating of overall experience completing the questionnaire, ranging from 0 (Terrible) to 10 (Excellent).

RESULTS

- A total of 149 participants were enrolled in the study, with 135 included in the comparison.*
- Characteristics of participants assigned to the W-, IVRS-, or I-based versions were comparable across all characteristics.

Item-Level Results

- There were no ceiling or floor effects.
  - In general, participants responded similarly at the two administrations.
- Table 1 contains the kappa statistic estimates and corresponding 95% confidence intervals (CIs) as measures of agreement.
  - The kappa statistics were highly satisfactory.

Total Score Results

- Table 2 shows that the descriptives of the total scores were comparable.
- Sequence effects: No significant differences between the P-based version and the alternate versions were found, irrespective of the order of administration. All further analyses were combined over sequence (i.e., P/W, P/IVRS, P/I).
- Test of mode: Paired t-tests were nonsignficant, further evidence that the LFQ versions are comparable at the total LFQ score level (Figure 3).
- Concordance: The ICCs were exceptionally higher than the threshold of 0.70, ranging from 0.81 to 0.93. The two highest ICCs were the W/P (0.93, 0.88-0.96) and the I/P (0.88, 0.79-0.93). The lowest, but still very acceptable ICC was IVRS/P (0.81, 0.68-0.89).

Classification

- Overall, the classifications were highly comparable across versions (Table 3).

Usability and Administration Version Preference Results

- Over 95% of participants reported no difficulties completing the P-based, I-based, or W-based versions; 87% reported no difficulties completing the IVRS-based version.
  - The remaining 13% assigned to the IVRS version reported just "slight" difficulty.
  - The most commonly reported complaint was that the IVRS system required respondents to wait until all answer choices were given for each question before the system would allow the selection of a response.
- When asked about their overall experience completing the questionnaire, approximately 98% of the participants reported having a good to excellent experience using the P- and W-based versions, 96% reported good to excellent for the I-based version, and 90% for the IVRS-based version.

Table 1. LFQ Item-Level Kappa Statistics, by Sequence Group

LFQ Item	Kappa (95% CI)					
	P-W	W-P	P-IVRS	IVRS-P	P-I	I-P
1	0.90 (0.80, 1.00)	0.78 (0.56, 0.99	0.73 (0.51, 0.94)	0.60 (0.32, 0.88)	0.72 (0.44, 1.00)	0.84 (0.63, 1.00)
2	0.83 (0.70, 0.96)	0.76 (0.56, 0.96)	0.79 (0.63, 0.95)	0.78 (0.59, 0.96)	0.73 (0.55, 0.92	0.88 (0.73, 1.00)
3	0.84 (0.71, 0.96)	0.67 (0.48, 0.85)	0.72 (0.53, 0.91)	0.52 (0.16, 0.89)	0.67 (0.45, 0.88)	0.79 (0.58, 0.99)
4	0.91 (0.83, 0.99)	0.97 (0.92, 1.00)	0.90 (0.81, 0.98)	0.85 (0.70, 1.00)	0.83 (0.67, 0.99)	0.85 (0.64, 1.00)
5	0.98 (0.95, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.97 (0.92, 1.00)	0.97 (0.92, 1.00)	1.00 (1.00, 1.00)

* Two participants were excluded because they did not complete two LFQ versions, and 12 participants were omitted because they reported active colds or infections at only one assessment, indicating a change in their respiratory state.

Table 2. LFQ Total Score Descriptive Statistics by Sequence Group

Sequence	Version	n	Mean	SD	Median	Min	Max
P-W	P	25	16.8	2.7	17.0	10.0	22.0
	W	25	16.8	2.4	17.0	11.0	21.0
W-P	P	23	17.1	2.0	17.0	13.0	20.0
	W	23	16.8	2.0	17.0	13.0	20.0
P-IVRS	P	23	17.0	2.2	17.0	12.0	22.0
	IVRS	23	17.4	2.4	17.0	14.0	22.0
IVRS-P	P	22	17.1	2.8	17.0	13.0	23.0
	IVRS	22	16.9	2.9	17.5	12.0	22.0
P-I	P	22	15.4	2.7	15.0	11.0	20.0
	I	22	15.7	2.4	15.0	11.0	20.0
I-P	P	20	16.1	2.9	16.0	8.0	20.0
	I	20	15.6	2.9	15.0	8.0	20.0

Figure 3. Boxplots of LFQ Total Scores by Combined Pair Group

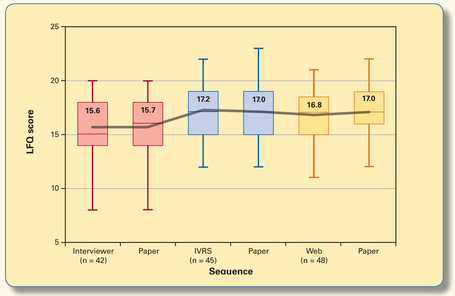


Table 3. Percentage Agreement in Obstruction Risk Between P-Based and Each Alternate Version, by Combined Pair Group

P-Based With Alternate Version	Obstruction Risk Both Versions n (%)	No Obstruction Risk Both Versions n (%)	Agreement n (%)	Obstruction Risk P Mode Only n (%)	Obstruction Risk Alternate Mode Only n (%)	Kappa
W	34 (70.8%)	12 (25.0%)	46 (95.8%)	0 (0.0%)	2 (4.2%)	0.89
IVRS	28 (62.2%)	12 (26.7%)	40 (88.9%)	5 (11.1%)	0 (0.0%)	0.75
I	31 (73.8%)	5 (11.9%)	36 (85.7%)	3 (7.1%)	3 (7.1%)	0.54

DISCUSSION

- An example using the LFQ, a case-finding tool originally designed for paper administration and adapted for three alternative versions using Internet, interview administration, and telephone technology provides a step-by-step illustration of the evaluation guide.
- Taken together, the evidence indicated high comparability between the item-level responses and the total scores of the LFQ, regardless of administration version. As a final evaluation, participants indicated that, although they had a preferred version, they had few difficulties with the versions they were assigned.
- Further psychometric evaluation within each version of the LFQ could help investigate and understand the lower ICC observed in the P/IVRS combined pair group, and the higher rate of disagreement (14%) in the P/I combined pair group.
- Study limitations:
  - Because the LFQ was developed as a case-finding tool, we based our ICC thresholds on 0.70 and not a higher bar of 0.90 (a threshold used to compare one individual's score with another individual's score).
  - Additionally, this study did not include spirometry, the "gold standard" to determine true airway obstruction risk; hence, new candidate cut points across versions could not be estimated or compared.

REFERENCES

1. Hanania NA, Mannino DM, Yawn BP, Mapel DW, Martinez FJ, Donohue JF, et al. Predicting risk of airflow obstruction in primary care: validation of the lung function questionnaire (LFQ). Respir Med. 2010 Aug;104(8):1160-70.
2. Yawn BP, Mapel DW, Mannino DM, Martinez FJ, Donohue JF, Hanania NA, et al. Development of the Lung Function Questionnaire (LFQ) to identify airflow obstruction. Int J Chron Obstruct Pulmon Dis. 2010 Feb;18;5:1-10.
3. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, 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 Jun;12(4):419-29.
4. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17:101-10.
5. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002 May 15;21(9):1331-5.
6. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. New York: Oxford University Press; 2003.
7. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.
8. GlaxoSmithKline (GSK). Data on file. Prevalence of chronic obstruction in subjects with a history of cigarette smoking in a primary care setting. Project ADC111116 COPD Screening Study. 2009. GSK, Research Triangle Institute, NC 27709.

CONTACT INFORMATION

Lauren M Nelson, PhD
Director of Psychometrics

RTI Health Solutions
200 Park Offices Drive
Research Triangle Park, NC 27709

Phone: +1.919.541.6590
Fax: +1.919.541.7222
E-mail: lnelson@rti.org

Presented at: ISPOR 16th Annual International Meeting
May 21-25, 2011
Baltimore, MD, United States