Paper 426-2012

Are You in Need of Validation?

Psychometric Evaluation of Questionnaires Using SAS®

Eric Elkin, ICON Late Phase & Outcomes Research, San Francisco, CA, USA

ABSTRACT

Presentations at prior SAS user group meetings have focused on factor analysis and related topics in order to develop the scale structure of a questionnaire. Instead, this presentation will assume that the scale has already been developed but needs validation as a new scale or for use in a new population. The examples are taken from health-related quality-of-life research and will follow the "Guidance for Industry" published by the FDA (Dec. 2009). The focus will be on the classical test theory approach to psychometric validation including internal consistency, test-retest, and inter-rater reliability; construct and known-groups validity; and responsiveness. The discussion will include samples of SAS code as well as tips for interpreting and presenting results.

INTRODUCTION

Questionnaires should undergo evaluation of their psychometric properties before they are relied on for making decisions. Psychometric validation is used in various fields including education, psychology, and health research to define the properties of the scales which are included in these questionnaires.

In health research (especially in outcomes research), questionnaires completed by patients to report on various aspects of their health-related quality-of-life (QoL), functional status, well-being, or satisfaction are known as patient-reported outcomes (PROs).

For this discussion, questionnaires (or surveys) are the things that the respondent fills out (on paper, by interview, on a computer, etc.). Questionnaires may contain one or more instruments representing distinct concepts (or domains) and often include other information such as demographic data. Scale scores (or instrument scores) are derived using mathematical combinations of specific item(s) from the questionnaire. Scales may be single-item or multi-item. Note that the term "scale" may also refer to the range and interval of response options to items or the combined instrument score (example: a 1-to-5 scale or a 0-to-100 scale).

Scales combine items (or questions) into concepts that the researcher wants to measure. There are scientific and statistical advantages to combining multiple items into a scale. For example, different aspects of a concept can be combined into one domain. For a "sadness" scale, a researcher might ask how often the respondent is sad in addition to rating how sad they are. In addition, multiple items can confirm that the respondent is consistent in his responses, often by including positively and negatively worded questions. For example, a researcher might ask how sad the respondent is and also how happy the respondent is. Statistically, combining items creates more variability in a measure. If two items have responses that are each on a 5-point integer scale (1-5) then taking the average provides scores with decimals and summing extends the scores to a range of 2-10. Adding more items moves the single item measured as an integer onto a more continuous multi-item scale. The main disadvantage to combining items is that missing items may lead to missing scales. However, this can be mitigated by scoring rules which account for missing items. Single-item scales with a missing item are always missing (no scoring rule can help).

For this discussion, we assume that the scoring algorithm has already been developed. The scale may need validation as a brand new scale to support the hypothesized scale structure or it may be an existing scale which is being used in a new population. Several good papers at prior SAS Global Forums and WUSS conferences describe scale development using principal components, exploratory factor analysis, confirmatory factor analysis, and variable clustering (see http://support.sas.com/events/sasglobalforum/previous/online.html or http://lexjansen.com/).

The science (and art) of questionnaire validation is undertaken to provide evidence that the instrument is measuring what it is supposed to measure. It demonstrates the quantitative integrity of the instrument. (Questionnaire validation also includes qualitative evidence; however this is beyond the scope of this discussion.) Of particular importance to those of us working in the pharmaceutical, biotechnology, and medical device industry is the recent FDA Guidance for Industry (*Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*, U.S. Dept Health and Human Services, Food and Drug Administration, December 2009). The guidance explains how the FDA evaluates PRO instruments to support claims for medical product labels, including the evaluation of the psychometric properties of these instruments. The guidance (and this discussion) focuses on the classical test theory approach to psychometric validation including reliability (internal consistency, test-retest, inter-rater); validity (construct, known-groups); and ability to detect change (responsiveness).

SCALE DESCRIPTION

Prior to using more sophisticated statistical methods, a full description of the scale score can provide early information about how well the instrument performs in the study population. Besides the mean, standard deviation, median and other descriptive statistics, it is helpful to review the percentage of missing scores. A high percentage

missing may indicate that patients don't understand how to complete the survey instrument or that the items are not meaningful to them. It may also indicate logistical issues with the study such as lost survey pages or incorrect programming of electronic surveys. In addition, the percentage of scores at the maximum and minimum values should also be assessed. Extreme percentages (too high or too low) may indicate floor or ceiling effects of the survey instrument. That is, the wording of the questions and/or the response options has meant that respondents favor the highest or lowest response for all items.

One way to implement these checks in SAS is to create indicator variables (0,1) to signify missing scores or scores at the maximum or minimum value. For a scale that ranges from 0 to 10, this might look like:

```
ind_miss_a = (scale_a eq .);
ind_max_a = (scale_a eq 10);
ind_min_a = (scale_a eq 0);
```

These indicators can be included in PROC MEANS along with the actual scale score. The mean of the indicator variables is equivalent to the proportion of observations with the value (missing, maximum, or minimum).

```
proc means data=temp01;
    var scale_a ind_miss_a ind_max_a ind_min_a;
    output out=temp02
       mean(scale_a ind_miss_a ind_max_a ind_min_a)=mn_scalea p_missa p_maxa p_mina
       n(scale_a)=n_scalea
       std(scale_a)=std_scalea
       median(scale_a)=med_scalea
       min(scale_a)=min_scalea
       max(scale_a)=max_scalea
       ;
run;
```

Table A. Description of scales used in the study

Scale	N	Mean	STD	Median	Actual Min	Actual Max	% Min	% Max	% Missing
A*	318	5.9	2.1	6.0	1	10	0.0	4.1	0.3
B*	318	5.4	2.4	5.1	0	10	0.6	3.8	0.3
C*	317	4.2	2.7	4.1	0	10	8.2	1.9	0.6
D†	297	2.6	0.8	3.2	1.1	4.7	0.0	0.0	6.9

* Scales A, B, C can range from 0 to 10 with higher scores indicating better functioning.
† Scale D can range from 1 to 5 with higher scores indicating worse symptom bother.

Table 1. Presentation of Results for Describing Scales

RELIABILITY

Reliability is the extent to which a measure is free from random error. The FDA guidance primarily considers three types of reliability when evaluating PRO instruments: internal consistency, test-retest, and inter-rater.

INTERNAL CONSISTENCY RELIABILITY

Internal consistency reliability evaluates the extent to which related items measure the same concept. It is measured using Cronbach's alpha which represents the degree to which items within a scale are inter-correlated with one another. Statistically, it is based on the sum of the variances of the items divided by the variance of the scale. Cronbach's alpha typically ranges from 0 to 1. Internal-consistency reliability is usually considered to be acceptable when Cronbach's alpha ≥0.70 (Special Advisory Committee of the Medical Outcomes Trust [SACMOT], 2002). Internal consistency is relevant only for multi-item scales.

SAS produces Cronbach's alpha as an option on PROC CORR. An example follows:

```
proc corr data=temp01 alpha;
    var item1 item2 item3 item4 item5;
run;
```

Sample output from SAS looks like the following:

Cronbach Coefficient Alpha	
Variables	Alpha
Raw	0.646000
Standardized	0.656554

Cronbach Coefficient Alpha with Deleted Variable				
	Raw Variables		Standardized Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
item1	0.477637	0.558306	0.481717	0.569833
item2	0.571099	0.501932	0.590448	0.514968
item3	0.476281	0.555717	0.474474	0.573365
item4	0.481494	0.555361	0.487106	0.567195
item5	0.070063	0.748415	0.069258	0.748124

Table 2. Output for Cronbach's Alpha

The standardized results use the values of the items after they are standardized to a standard deviation of 1. This can be especially helpful if the items have very different ranges (for example, some items range from 1 to 5 and some items range from 0 to 100). In the current example, all of the items are on the same scale so there is not much difference between the raw and the standardized values. In this example, Cronbach's alpha is 0.66 (standardized result), which does not meet the *a priori* criterion of alpha ≥0.70.

The SAS output contains further details which can help explain what is happening. The "correlation with total" shows the correlation of the item with the scale (all items combined). The "alpha" column shows what the Cronbach's alpha would be if the item were removed. One can see that item5 has a low correlation with the total and that alpha would increase if it were removed from the scale. The other items have reasonable correlations with the total and if any of them were removed the alpha would decrease.

Removing item 5 produces the following results:

Cronbach Coefficient Alpha	
Variables	Alpha
Raw	0.748415
Standardized	0.748124

Cronbach Coefficient Alpha with Deleted Variable				
	Raw Variables		Standardized Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
item1	0.521674	0.702182	0.519412	0.702879
item2	0.668977	0.612959	0.672598	0.614393
item3	0.413912	0.760271	0.414690	0.758668
item4	0.581207	0.669772	0.575711	0.671328

Table 3. Output for Cronbach's Alpha, Removing Item 5

Some additional issues to consider when performing analysis of internal consistency:

- If a scale score is already defined as a simple sum or average of items with a similar range of possible values, report the raw results. If the weights for combining a scale have not been determined, it makes sense to report the standardized results (see discussion above) to be consistent across analyses. This will be more meaningful when items are on different measurement scales, while not making much difference when they are all on the same scale.
- Understand the extent of missing values for the items. The nomiss option on the PROC CORR statement results in casewise deletion if any item is missing. A conservative approach would always use this option. However, if there are few missing values then this is not necessary. The nomiss option should be considered when item(s) have a lot of missing values, at least as a sensitivity analysis.
- Understand the directionality (positive vs. negative wording of the question) of the items. Items should be reversed, if needed, prior to the evaluation of internal consistency. This can greatly impact the results. For example, item3 has been reversed so it is in the opposite direction of the other 3 items. This causes a major change in the results:

Cronbach Coefficient Alpha	
Variables	Alpha
Raw	0.268402
Standardized	0.272344

Cronbach Coefficient Alpha with Deleted Variable				
	Raw Variables		Standardized Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
item1	0.350278	-.069185	0.345328	-.073999
item2	0.459731	-.318242	0.473273	-.283447
item3 (reversed)	-.413912	0.760271	-.414690	0.758668
item4	0.449602	-.231768	0.441948	-.229978

Table 4. Output for Cronbach's Alpha, Removing Item 5 and Reversing Item 3

TEST-RETEST RELIABILITY

Test-retest reliability is used to understand how stable a respondent's answers are over time. In other words, if you gave the same questionnaire to the same patient at a different time (and nothing else had changed) how consistent would the patient's answers be? Test-retest reliability is measured by the intraclass correlation coefficient (ICC). The ICC is the proportion of the total variance explained by the between-person variance. In other words, if the between-person variance is much greater than the within-person variance over the two administrations then the instrument is considered reliable over the test-retest period (Deyo et al, 1991). The ICC theoretically ranges from 0 to 1. An ICC ≥ 0.70 is an acceptable level of test-retest reliability (SACMOT, 2002).

The data set should include two records per patient: one record for each patient with the score at the "test" time and one record with the score at "retest". The ICC can be calculated using a macro available from SAS support (http://support.sas.com/kb/25/031.html).

Some additional issues to consider when performing analysis of test-retest reliability:

- How far apart should the test and retest time points be from each other? This will depend upon the subject matter as well as the logistics of the study.
- Test-retest assumes that nothing else has changed except for time. Researchers can often evaluate this by asking the respondent at retest whether there have been any changes (either positive or negative) in their health status (disease or treatment changes) since the first questionnaire. Then a sensitivity analysis can be performed by calculating the ICC only for those patients who reported no changes between the two test periods.

INTER-RATER RELIABILITY

Inter-rater reliability is used to understand how consistent responses are when two different interviewers give the same questionnaire to the same respondent. It is also used when evaluating how similar proxy responses are to the patient's response. For example, a researcher may want to conduct a pilot study to test if caregivers should be allowed to complete the questionnaire for the patients in an arthritis study in which the patients may not be able to hold a pen. Inter-rater reliability is also measured by the ICC (see the section on test-retest reliability for additional details).

CONSTRUCT VALIDITY

Validity is the extent to which the instrument measures what it is intended to (and not measuring what it isn't intended to). The FDA guidance considers content and construct validity when evaluating PRO instruments. Content validity is based on qualitative evidence and will not be discussed further here. Construct validity is the examination of logical relationships between related scales or between scales and other patient or disease characteristics. The two types of construct validity reviewed by the FDA are convergent/discriminant validity and known groups validity.

CONVERGENT/DISCRIMINANT VALIDITY

Convergent and discriminant validity is based upon *a priori* hypotheses regarding which instruments should be associated with each other (convergent) and which should not be (discriminant). The hypotheses are developed based on understanding the underlying constructs of the instruments. They include not only direction but strength of the association. Statistically, Pearson correlation coefficients are calculated for each hypothesized relationship. For example, a researcher might hypothesize that a scale measuring depression would be highly correlated with another scale measuring emotional functioning and moderately correlated with a scale measuring optimism (convergent validity). In addition, the researcher may hypothesize that the depression scale will have a lower correlation with a physical functioning scale than with the emotional functioning scale (discriminant validity). Of course, the researcher must define high, moderate, and low correlation levels before performing the analysis.

When presenting convergent or discriminant validity, it is often easiest to provide a full correlation matrix of all the scale scores and then highlight which ones did or did not meet the *a priori* hypotheses. A tip for presenting large correlation matrices is to use a numbering system for column labels and define the scale using a row label. This can keep the columns narrow and allow more columns on the page. See the example below:

Table B. Correlation matrix of scales used in the study

Scale	Scale #	1	2	3	4	5	6	7	8	9	10	11	12
Physical Functioning	1	1.00	--	--	--	--	--	--	--	--	--	--	--
Emotional Functioning	2	0.34	1.00	--	--	--	--	--	--	--	--	--	--
Social Functioning	3	0.46	0.49	1.00	--	--	--	--	--	--	--	--	--
Depression	4	-.63	-.27	-.36	1.00	--	--	--	--	--	--	--	--
Coping	5	0.33	0.45	0.51	-.20	1.00	--	--	--	--	--	--	--
Anxiety	6	-.30	-.52	-.44	0.26	-.39	1.00	--	--	--	--	--	--
Work Satisfaction	7	-.45	-.46	-.50	0.44	-.28	0.53	1.00	--	--	--	--	--
Family Relationships	8	0.40	0.41	0.55	-.38	0.39	-.41	-.55	1.00	--	--	--	--
Financial Problems	9	-.39	-.46	-.53	0.42	-.35	0.54	0.65	-.61	1.00	--	--	--
Life Changes	10	-.38	-.42	-.49	0.39	-.31	0.52	0.58	-.61	0.70	1.00	--	--
Treatment Burden	11	-.41	-.53	-.54	0.42	-.41	0.62	0.67	-.64	0.78	0.72	1.00	--
Symptom Bother	12	-.27	-.25	-.34	0.28	-.31	0.35	0.47	-.35	0.46	0.43	0.56	1.00

Table 5. Presentation of Results for Correlation Matrix

Another technique for compact display of information about scales is to replace the correlation of 1.00 along the diagonal with the internal consistency reliability as measured by Cronbach's alpha.

KNOWN GROUPS VALIDITY

Known groups validity also relies upon *a priori* hypotheses which are created based on understanding the underlying constructs of the instruments and the other data collected as part of the study. The mean instrument score is compared between groups known to differ on the construct being measured – very often some measure of severity (e.g., lab finding, physician assessment). The hypothesis is evaluated by t-test or analysis of variance (ANOVA). In the case of ANOVA it is important to use a multiple comparison test to distinguish which groups (such as low, medium, and high severity patient groups) are different from each other.

Results may be presented as in the example below:

Table C. Study scale scores by disease severity

Scale	Low Severity Mean (SD)	Medium Severity Mean (SD)	High Severity Mean (SD)	Overall P value‡	Group Comparisons§
A*	8.6 (1.9)	5.8 (2.1)	3.7 (2.0)	<0.01	a, b, c
B*	6.7 (2.1)	5.6 (2.5)	4.7 (2.4)	0.04	b
C*	4.4 (2.5)	4.1 (2.4)	3.7 (2.8)	0.12	---
D†	1.6 (0.6)	2.8 (0.6)	3.8 (0.7)	<0.01	a, b, c

* Scales A, B, C can range from 0 to 10 with higher scores indicating better functioning.

† Scale D can range from 1 to 5 with higher scores indicating worse symptom bother.

‡ P values from ANOVA.

§ Evaluation of pairwise comparisons of the mean using Tukey's studentized range test.

Notation key:

a = low and medium severity patients are different

b = low and high severity patients are different

c = medium and high severity patients are different

Table 6. Presentation of Results for Known Groups Validity

ABILITY TO DETECT CHANGE

In the FDA guidance, the ability to detect change (often referred to as responsiveness) is based upon a comparison of change in the instrument score to change in other variables. These other variables should indicate that the

patient's health state has changed with respect to the construct under consideration. For example, if patients are "cured" then how does their scale score change? What if, instead, patients develop more side effects – how does their scale score change in this case? As illustrated by the examples, change may be positive or negative and the instrument should be able to detect either type of change. The criterion for change in health could, for example, be a physician's evaluation of response to treatment, a change in a lab test, or reporting of a new side effect of treatment. Assessment of responsiveness usually requires longitudinal data.

WITHIN PERSON CHANGE OVER TIME

Based on the criterion for change, patients are very often grouped by those who improved, those who did not change, and those who worsened. The instrument's change score for each patient is calculated by subtracting the baseline score from the follow-up score for the time period under consideration. The mean change scores are compared for the patients in the three groups by ANOVA with a multiple comparison test. This analysis is very much like known groups validity but with a time component added to define the known groups.

Sample results are shown below:

Table D. Change in study scale scores by clinical change

Scale	Improved Mean (SD)	No Change Mean (SD)	Worsened Mean (SD)	Overall P value‡	Group Comparisons§
A*	+7.3 (5.3)	+0.8 (4.1)	-5.6 (5.4)	<0.01	a, b, c
B*	+5.7 (5.6)	+1.1 (4.4)	-1.8 (4.6)	0.03	a
C*	+2.3 (5.2)	+1.8 (4.6)	-0.4 (6.3)	0.21	---
D†	-2.9 (2.7)	+0.4 (3.2)	+3.2 (3.1)	<0.01	a, b, c

* Change scores for scales A, B, C can range from -10 to 10 with higher scores indicating improvement in functioning.

† Change scores for scale D can range from -4 to +4 with higher scores indicating worsening of symptom bother.

‡ P values from ANOVA.

§ Evaluation of pairwise comparisons of the mean using Tukey's studentized range test.

Notation key:

a = improved and no change patients are different

b = improved and worsened patients are different

c = no change and worsened patients are different

Table 7. Presentation of Results for Within Person Change over Time

EFFECT SIZE

The FDA guidance points out that the ability of an instrument to detect change can be incorporated into the sample size calculations needed for using the instrument as an outcome in a clinical study. Therefore, the effect size of the instrument can also be used as evidence for its ability to detect change. The effect size is the mean of the change score divided by the standard deviation (SD) of the baseline score. The baseline SD represents the amount of variability in the instrument score for the population. The effect size, then, describes how much the instrument score shifted relative to the variability of the population (i.e., the score moved X number of standard deviations). The effect size is easy to calculate once the analysis shown above (in Table D) is completed. All that is needed is the SD of the baseline score rather than the SD of the change score (which is presented in Table D).

Sample results are shown below:

Table E. Effect size of study scale scores by clinical change

Scale	Baseline SD	Improved	No Change	Worsened
A*	2.1	+3.5	+0.4	-2.7
B*	2.4	+2.4	+0.5	-0.8
C*	2.7	+0.9	+0.7	-0.1
D†	0.8	-3.6	+0.5	+4.0

* For scales A, B, C positive effect sizes indicate improvement in functioning.

† For scale D positive effect sizes indicating worsening of symptom bother.

Table 8. Presentation of Results for Effect Size

An additional issue to consider when performing analysis of responsiveness:

- How far apart should the baseline and follow-up time points be from each other? This will depend upon the subject matter as well as the logistics of the study. The amount of time between the two can affect the degree of change in the instrument for those patients in the "no clinical change" group. This can happen due to external forces in patient's lives that affect their well-being. For example, responsiveness over one week may be very different from responsiveness over 6 months.

CONCLUSION

This paper discusses the classical test theory approach to psychometric validation including internal consistency, test-retest, and inter-rater reliability; construct and known-groups validity; and responsiveness. The methods described here should fulfill the set of recommendations set out by the FDA in the "Guidance for Industry." Several aspects of questionnaire instruments (e.g., number of items, response options, recall period, time between administrations, respondent burden, and scoring algorithm) can affect their psychometric properties. Therefore, it is recommended that the statistician work collaboratively with knowledgeable professionals in the field for study planning and interpretation of results.

REFERENCES

- Deyo RA, Diehr P, Patrick DL (1991). "Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation." *Controlled Clinical Trials*. 12(Suppl): 142S-158S.
- U.S. Dept Health and Human Services, Food and Drug Administration (December 2009). *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*.
- Special Advisory Committee of the Medical Outcomes Trust (2002). "Assessing health status and quality-of-life instruments: attributes and review criteria." *Quality of Life Research*. 11(3): 193-205.
- Stewart AL (1990). "Psychometric considerations in functional status instruments." In: WONCA Classification Committee (ed.), *Functional Status Measurement in Primary Care*. New York: Springer-Verlag.

ACKNOWLEDGMENTS

I would like to thank David J. Pasta (Vice President of Statistical and Strategic Analysis, ICON Late Phase & Outcomes Research) for his consultation and review of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Eric Elkin
ICON Late Phase & Outcomes Research
188 The Embarcadero, Suite 200
San Francisco, CA 94105
Phone: 415-371-2153
Fax: 415-856-0840
E-mail: eric.elkin@iconplc.com
Web: www.iconplc.com