Determining Missing Data Rules for Patient-Reported Outcomes: Alpha-If-Item-Deleted

Cheryl D Coon, Valerie SL Williams, Lauren M Nelson, Mark A Price

RTI Health Solutions, Research Triangle Park, NC, United States

BACKGROUND

- Instrument developers must establish scoring algorithms, including rules for handling missing data.
- Often, instrument developers use arbitrary rules of thumb when creating scoring guidelines for missing data.
- Patient-reported outcomes research commonly uses the "50% rule," which indicates that an overall or subscale score can only be computed if at least half of the items of the instrument or subscale are nonmissing.
- The arbitrary application of missingness rules without regard for item properties can lead to scores with inadequate reliability.
- Cronbach's alpha is considered a lower bound estimate for reliability, and while other methods exist, this approach has been commonplace in instrument development for the last half century.[1,2,3]
- One simple approach for empirically determining the maximum tolerable number of missing item-level responses is a stepwise alpha-if-item-deleted method described by Mitchell and Bradley.[4]

OBJECTIVE

- To compare the reliability of scores based on the alpha-if-item-deleted method's missing data rule to the reliability of scores based on the rule proposed by the instrument developer.

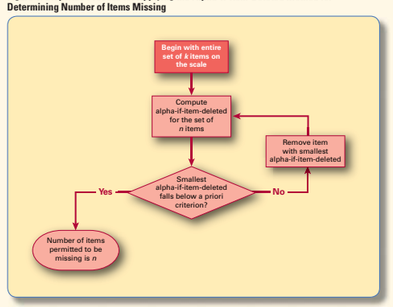
METHOD

- Using Cronbach's alpha[5] to compute reliability of a *k*-item scale:

$$\alpha = \left(\frac{k}{k-1}\right)\left[1 - \frac{\Sigma_i^k V(Y_i)}{V(\Sigma_i^k Y_i)}\right]$$

- We computed alpha three times for each instrument:
  1. All items on the scale to identify the upper-bound of alpha for this scale score when no data are missing.
  2. Only the number of items permitted to be missing according to the developer's guidelines, where the items excluded were those with the largest contribution to alpha (i.e., lowest alpha-if-item-deleted values).
  3. Only the number of items permitted to be missing according to the alpha-if-item-deleted approach (Figure 1), also basing exclusion criteria on the largest contribution to alpha.
- Stepwise alpha-if-item-deleted method:[4]
  - Cronbach's alpha is calculated sequentially as items are excluded from the item set one at a time.
  - The order in which items are removed from the item set is based on the magnitude of alpha-if-item-deleted, where the item with the smallest alpha-if-item-deleted (i.e., the item with the greatest contribution to the alpha) is removed at each step.
  - When the smallest alpha-if-item-deleted for a set of items falls below an a priori identified threshold (e.g., 0.70), the number of items considered at that step is determined to be the minimum number of items needed to calculate a reliable scale score for an individual.
  - In the present study, we used an alpha of 0.70, the minimum recommended by Streiner and Norman[6] for group-level comparisons. The alpha for the stepwise method should usually be just above 0.70 because the removal of one additional item would drop alpha below this standard of 0.70.

Figure 1. Stepwise Process for Applying the Alpha-If-Item-Deleted Method for Determining Number of Items Missing



- Table 1 illustrates the stepwise alpha-if-item-deleted method for the 9-item hypothetical Example Questionnaire (ExQ).
  - ExQ items were removed in successive steps beginning with the item with the smallest alpha-if-item-deleted (i.e., the item that contributes the most to alpha), until the alpha of the remaining items fell below the threshold of 0.70.

  Step 1: Shows the alpha for the full 9-item scale (0.87) and the alpha-if-item-deleted for each of the 9 ExQ items.
  - Item 7 had the smallest alpha-if-item-deleted (0.84), so it was removed.

  Step 2: The alpha for the 8 remaining items was 0.84 (the alpha-if-item-deleted for the excluded item at the previous step).
  - Item 5 had the lowest alpha-if-item-deleted (0.79) so it was removed.

  Step 3: The remaining 7 items had an alpha of 0.79.
  - Item 3 had the smallest alpha-if-item-deleted (0.73) so it was removed.

  Step 4: The remaining 6 items had an alpha of 0.73.
  - Item 2 had the smallest alpha-if-item-deleted (0.65) so it was removed.
- The deletion of Item 2 in Step 4 drops the overall alpha for the remaining items below the 0.70 threshold (0.65); thus, the minimum number of nonmissing responses from the 9-item ExQ is 6 (the number of items considered at the last step when overall alpha is greater than the threshold).

Table 1. Sequential Alpha-If-Item-Deleted for the ExQ Items at Each Step

Item	Alpha-If-Item-Deleted			
	Step 1	Step 2	Step 3	Step 4
1	0.88	0.85	0.79	0.74
2	0.85	0.81	0.74	0.65
3	0.84	0.80	0.73	—
4	0.85	0.82	0.75	0.67
5	0.83	0.79	—	—
6	0.86	0.82	0.76	0.69
7	0.84	—	—	—
8	0.86	0.83	0.76	0.70
9	0.88	0.85	0.79	0.71
Total alpha	0.87	0.84	0.79	0.73

RESULTS

- A total of 6 datasets were analyzed, including 9 different instruments for a total of 34 different subscales (Table 2).
- Instruments: Brief Pain Inventory-Short Form (BPI-SF),[7] Center for Epidemiological Studies Depression Scale (CES-D),[8] Irritable Bowel Syndrome Quality of Life Questionnaire (IBS-QOL)[9] Medical Outcomes Study Cognitive Questionnaire (MOS-Cog)[10] Medical Outcomes Study Sleep Scale (MOS-Sleep),[11] Paediatric Asthma Caregiver's Quality of Life Questionnaire (PACQLQ)[12] Patient Assessment of Constipation Quality of Life questionnaire (PAC-QOL),[13] Patient Assessment of Constipation Symptoms Questionnaire (PAC-SYM),[14,15] Medical Outcomes Study Short Form–36 (SF-36).[16]

Scenario 1

- Alphas computed using the developer's guidelines were greater than alphas computed using the stepwise rule.
- This result indicated that the developer's guidelines were conservative and reliable scores could be calculated using fewer items (i.e., in the presence of more missing data).
- Alphas computed using the developer's guidelines were smaller than the corresponding alphas computed using the full item set, but still very satisfactory.
- Instruments: BPI-SF Impact of Pain on Daily Function; CES-D; PACQLQ Overall and Emotional Functioning; PAC-QOL Overall, Worries and Concerns, and Psychosocial Discomfort; PAC-SYM Overall.

Scenario 2

- Alphas computed using the developer's guidelines were smaller than alphas computed using the stepwise rule (and much smaller than alphas computed using the full item set).
- This result indicated that the developer's guidelines were too liberal (i.e., allowed too many missing data) and scores calculated with the recommended allowable number of items missing did not meet our reliability threshold.
- Instruments: BPI-SF Severity of Pain; IBS-QOL Mental Health and Sleep; MOS-Sleep Sleep Disturbance, Daytime Somnolence, 6-item Sleep Problems Index, and 9-item Sleep Problems Index; PACQLQ Activity Limitations; PAC-QOL Satisfaction; PAC-SYM Abdominal Symptoms and Rectal Symptoms; SF-36 Vitality and General Health Perceptions.

Scenario 3

- The developer's guidelines and the stepwise alpha-if-item-deleted method guidelines were the same.
- Instruments: IBS-QOL Emotional Health, Physical Function, Physical Role, Diet, Sexual Relations, and Social Role; MOS-Cog; PAC-QOL Physical Discomfort; PAC-SYM Stool; SF-36 Physical Functioning, Role Limitations due to Physical Health, General Mental Health.

Table 2. Comparison of Alphas and Number of Allowable Items Based on Developer's Guidelines and the Stepwise Alpha-if-Item-Deleted Rule

Scale	Number of Items	Alpha for Full-Scale	Developer's Guidelines		Alpha-If-Item-Deleted Rule	
			Number of Nonmissing Items Required	Alpha	Number of Nonmissing Items Required	Alpha
BPI-SF (n = 150)						
Severity of Pain	4	0.83	2	0.65	3	0.77
Impact of Pain on Daily Function	7	0.94	5	0.91	2	0.71
CES-D (n = 850)	20	0.91	19	0.90	9	0.73
IBS-QOL (n = 250)						
Emotional Health	4	0.90	2	0.77	2	0.77
Mental Health	4	0.86	2	0.62	3	0.79
Sleep	3	0.82	2	0.67	3	0.82
Physical Function	3	0.90	2	0.78	2	0.78
Physical Role	4	0.93	2	0.84	2	0.84
Diet	3	0.77	2	0.63	3	0.77
Sexual Relations	3	0.89	2	0.80	2	0.80
Social Role	4	0.90	2	0.79	2	0.79
MOS-Cog (n = 500)	6	0.89	3	0.76	3	0.76
MOS-Sleep (n = 950)						
Sleep Disturbance	4	0.77	1	-	4	0.77
Daytime Somnolence	3	0.75	1	-	3	0.75
6-item Sleep Problems Index	6	0.75	1	-	6	0.75
9-item Sleep Problems Index	9	0.83	1	-	6	0.71
PACQLQ (n = 200)						
Activity Limitations	4	0.86	2	0.63	3	0.79
Emotional Functioning	9	0.89	5	0.78	4	0.73
Overall	13	0.91	7	0.81	5	0.73
PAC-QOL (n = 650)						
Worries and Concerns	11	0.95	6	0.82	4	0.75
Physical Discomfort	4	0.86	2	0.72	2	0.72
Psychosocial Discomfort	8	0.92	4	0.81	3	0.75
Satisfaction	5	0.85	3	0.62	4	0.79
Overall	28	0.96	14	0.88	7	0.74
PAC-SYM (n = 650)						
Abdominal	4	0.86	2	0.66	3	0.77
Rectal	3	0.82	2	0.66	3	0.82
Stool	5	0.87	3	0.75	3	0.75
Overall	12	0.91	6	0.79	5	0.74
SF-36 (n = 150)						
Physical Functioning	10	0.89	5	0.74	5	0.74
Role Limitations due to Physical Health	4	0.93	2	0.85	2	0.85
General Mental Health	5	0.89	3	0.78	3	0.78
Role Limitations due to Emotional Problems	3	0.94	2	0.87	2	0.87
Vitality	4	0.83	2	0.56	3	0.76
General Health	5	0.81	3	0.61	4	0.74

Note: Rows highlighted in yellow signify Scenario 1. Rows highlighted in red signify Scenario 2. Rows highlighted in green signify Scenario 3. Only subscales with 3 or more items are presented.

- Rules of thumb cannot be applied blindly across scales because they don't account for scale properties (e.g., overall reliability, strength of individual items).
- Even within the same instrument, the application of the same rule (e.g., the 50% rule) may be too conservative for some subscales and too liberal for others. For example:
  - BPI-SF Impact of Pain on Daily Function (conservative) vs. BPI-SF Severity of Pain (liberal)
  - PACQLQ Overall and Emotional Functioning (conservative) vs. PACQLQ Activity Limitations (liberal)
  - PAC-QOL Overall, Worries and Concerns, and Psychosocial (conservative) vs. PAC-QOL Satisfaction (liberal)
  - PAC-SYM Overall (conservative) vs. PAC-SYM Abdominal Symptoms and Rectal Symptoms (liberal).
- Overall alphas are larger for scales where the developer guidelines are too conservative rather than too liberal. In most cases, the overall alpha for scales with conservative developer guidelines is greater than 0.90, and there's more room for alpha to decrease with item deletion before going below 0.70 (Table 2).

DISCUSSION AND CONCLUSIONS

- It is important to consider the reliability of the scale and the effect of item-level missing data on scale reliability.
- The application of the 50% rule and other rules of thumb may be too liberal for some subscales and too conservative for other subscales within the same instrument because item and instrument properties are not carefully considered.
  - When too many missing data are allowed, the available data are insufficient and scores are less reliable than expected using the overall scale. At minimum, the resulting score is full of error, and excluding scores with such error can increase the signal to noise ratio when considering study results.
  - When the developer's rule unnecessarily restricts meaningful missing item-level data, missing score values will be assigned to individuals where reliable scores could be calculated given even larger amounts of missing data. The reliability of the resulting score may be greater than 0.70.
- Calculation of reliable subscale scores when missing data are present is desirable because it provides information at the individual level and increases power at the study level.
- When the alpha-if-item-deleted method is applied, it is known that all scores calculated with a reduced item set when item responses are missing, no matter which items are missing, will meet a certain reliability level.

LIMITATIONS

- There are a number of methods available for assessing reliability that consider levels of reliability across the range of scale scores (e.g., standard error of measurement in item response theory).
  - This stepwise Cronbach's alpha-based method is presented as the most common and straightforward method to consider when scoring scales with missing data, although researchers are encouraged to consider reliability as calculated through model-based scoring algorithms.
- The stepwise alpha-if-item-deleted method can lead to the deletion of items whose removal entails a loss of composite reliability, despite an improvement in alpha.[17,18]
  - This method should be considered a first step for determining the number of items that can be missing from a subscale score, and resulting alphas should be considered a lower bound for reliability.
- Missing data scoring guidelines, especially the methods examined here, do not consider the content validity of the scale.
  - Allowing for a certain number of missing items can lead to scale scores that no longer measure the full spectrum of the intended construct.

REFERENCES

1. Cortina JM. What Is Coefficient Alpha? An Examination of Theory and Applications. Journal of Applied Psychology 1993;78:98-104.
2. Zimmerman DW, Zumbo BD, Lalonde C. Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and Psychological Measurement 1993;53:33-49.
3. Osburn HG. Coefficient alpha and related internal consistency reliability coefficients. Psychological Methods 2000;5:343-55.
4. Mitchell J, Bradley C. Psychometric evaluation of the 12-item Well-being Questionnaire for use with people with macular disease. Qual Life Res 2001;10:465-73.
5. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:294-334.
6. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. New York: Oxford University Press; 1995.
7. Hahn BE, Kirchdoerfer LJ, Fullerton S, Mayer E. Evaluation of a new quality of life questionnaire for patients with irritable bowel syndrome. Aliment PharmacolTher 1997;11:547-52.
8. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Journal of Applied Psychological Measurement 1977;1:385-401.
9. Mendoza T, Mayne T, Rublee D, Cleeland C. Reliability and validity of a modified Brief Pain Inventory short form in patients with osteoarthritis. Eur J Pain 2006;10:353-61.
10. Stewart AL, Ware JE, Sherbourne CD, Wells KB. (1992). Psychological distress/well-being and cognitive functioning measures. In A.L. Stewart and J.E. Ware (Eds.), Measuring Functioning and Well-being: The Medical Outcomes Study Approach. Durham NC: Duke University Press.
11. Hays RD, Stewart AL. Sleep measures. In: AL Stewart, JE Ware, editors. Measuring functional and well-being: The Medical Outcomes Study approach. Durham, NC: Duke University Press; 1992. p. 235-59.
12. Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in the parents of children with asthma. Qual Life Res 1996; 5:27-34.
13. Marquis P, De La Loge C, Dubois D, McDermott A, Chassany O. Development and validation of the patient assessment of constipation quality of life questionnaire. Scand J Gastroenterol 2005 May;40:540-51.
14. Frank L, Kleinman L, Farup C, Taylor L, Miner P Jr. Psychometric validation of a constipation symptom assessment questionnaire. Scand J Gastroenterol 1999 Sep;34:870-7.
15. Slappendel R, Simpson K, Dubois D, Keininger DL. Validation of the PAC-SYM questionnaire for opoid-induced constipation in patients with chronic low back pain. Eur J Pain 2006;10:209-17.
16. Ware JJ, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
17. Raykov T. Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. The British Journal of Mathematical and Statistical Psychology 2007;60:201-16.
18. Raykov T. Alpha if item deleted: A note on loss of criterion validity in scale development if maximizing coefficient alpha. The British Journal of Mathematical and Statistical Psychology 2008;61:275-85.

CONTACT INFORMATION

Cheryl D Coon
Director, Psychometrics

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

Phone: +1.702.818.4249
Fax: +1.702.818.4249
E-mail: ccoon@rti.org

Presented at: ISPOR 15th Annual International Meeting
May 15-19, 2010
Atlanta, GA, United States