Interrater Reliability of the Adapted Fresno Test across Multiple Raters

Lucylynn M. Lizarondo, MSPT;* Karen Grimmer, PhD;*† Saravana Kumar, PhD*†

ABSTRACT

Purpose: The Adapted Fresno Test (AFT) is a seven-item instrument for assessing knowledge and skills in the major domains of evidence-based practice (EBP), including formulating clinical questions and searching for and critically appraising research evidence. This study examined the interrater reliability of the AFT using several raters with different levels of professional experience. **Method:** The AFT was completed by physiotherapists and occupational therapists, and a random sample of 12 tests was scored by four raters with different levels of professional experience. Interrater reliability was calculated using intra-class correlation coefficients (ICC [2, 1]) for the individual AFT items and the total AFT score. **Results:** Interrater reliability was moderate to excellent for items 1 and 7 (ICC = 0.63–0.95). Questionable levels of reliability among raters were found for other items and for the total score. For these items, the raters were clustered into two groups—“experienced” and “inexperienced”—and then examined for reliability. The reliability estimates for rater 1 and rater 2 (“inexperienced”) increased slightly for items 2 and 5 and for the total score, but not for other items. For raters 3 and 4 (“experienced”), ICCs increased considerably, indicating excellent reliability for all items and for the total score (0.80–0.99), except for item 4, which showed a further decrease in ICC. **Conclusion:** Use of the AFT to assess knowledge and skills in EBP may be problematic unless raters are carefully selected and trained.

Key Words: Adapted Fresno Test; evidence-based practice; occupational therapy; physical therapy specialty; reproducibility of results.

RÉSUMÉ

Objectif : Le test adapté de Fresno (*Adapted Fresno Test*, AFT) est un instrument de mesure en sept points qui vise l'évaluation des connaissances et des compétences dans les principaux domaines de pratique fondée sur des faits probants, notamment la formulation de questions cliniques et la recherche ainsi que l'évaluation critique de preuves issues de la recherche. L'étude s'est penchée sur la fiabilité de l'AFT entre évaluateurs, en travaillant avec des évaluateurs ayant divers degrés d'expérience professionnelle. **Méthodologie :** Des physiothérapeutes et des ergothérapeutes ont procédé à un AFT. Un échantillon aléatoire de 12 tests a ensuite été analysé par quatre évaluateurs avec des degrés variés d'expérience professionnelle. La fiabilité entre évaluateurs a été calculée à l'aide de coefficients de corrélation intraclasse (CCI [2, 1]) pour les points individuels de l'AFT et pour le pointage total de l'AFT. **Résultats :** La fiabilité entre évaluateurs variait de modérée à excellente pour les points 1 et 7 (CCI = 0,63–0,95). Pour les autres points et pour la note totale toutefois, les niveaux de fiabilité sont sujets à caution. Pour les points en question, les évaluateurs ont été séparés en deux groupes—les « expérimentés » et les « inexpérimentés »—et leur fiabilité a ensuite été analysée. Les estimations de la fiabilité de l'évaluateur 1 et de l'évaluateur 2 (« inexpérimentés ») étaient légèrement supérieures pour les points 2 et 5 et pour le pointage total, mais ce n'était pas le cas pour les autres points. Pour les évaluateurs 3 et 4 (« expérimentés »), le CCI était beaucoup plus élevé, ce qui dénote une excellente fiabilité pour tous les points et pour la note totale (0,70–0,99), sauf pour le point 4, qui affichait une baisse plus marquée de son CCI. **Conclusion :** L'utilisation du test de Fresno adapté pour évaluer les connaissances et les compétences en pratique fondée sur les faits probants peut être problématique, à moins que les évaluateurs soient rigoureusement choisis et formés.

Current research suggests that physiotherapists and occupational therapists have positive attitudes toward evidence-based practice (EBP) and believe that their practice should be evidence based.^{1–4} However, while they recognize its importance and value, EBP is often not integrated into their day-to-day practice.^{2,5} The gap between practitioners' intentions and their actual practice has been attributed to a lack of the knowledge and skills needed to undertake EBP processes,^{1,2,4} including formulating clinical questions, searching for relevant evidence,

critically appraising evidence, implementing it into practice, and evaluating outcomes.^{6,7} While it is unclear whether addressing this issue will translate into changes in patient outcomes, we believe that addressing this knowledge and skills gap will enable practitioners to use evidence effectively to inform their health care decisions.

Researchers have proposed providing EBP training programmes to health care practitioners as an effective way of facilitating an evidence-based approach to clinical practice.^{8,9} Educators who provide such training require

From the *International Centre for Allied Health Evidence and †School of Health Sciences, University of South Australia, Adelaide.

Correspondence to: Lucylynn Lizarondo, International Centre for Allied Health Evidence, University of South Australia, North Terrace, Adelaide, SA 5000 Australia; lucylynn.lizarondo@unisa.edu.au.

Contributors: All authors designed the study, collected the data, and analyzed and interpreted the data; drafted or critically revised the article; and approved the final draft.

Competing interests: None declared.

Physiotherapy Canada 2013; 65(2):135–140; doi:10.3138/ptc.2012-15

standard, robust instruments to evaluate the effectiveness of their programmes and to document changes in the competence of practitioners being trained. At present, however, research on EBP often relies on self-report data, which are subjective and potentially non-standardized. There is some evidence to suggest that individuals' self-reports of their own knowledge are often inaccurate in representing their actual knowledge,^{10,11} since assessments based on self-reporting are subject to response bias. Therefore, such assessments can be useful only if the reported perception of knowledge is valid relative to objective assessments of knowledge.

THE ADAPTED FRESNO TEST

To our knowledge, only one objective measure of EBP knowledge and skills has been tested and applied in allied health: the Adapted Fresno Test (AFT),¹² a seven-item instrument for assessing knowledge and skills in the major domains of EBP, such as formulating clinical questions and searching for and critically appraising research evidence. The questions revolve around two clinical scenarios relevant to allied health. Respondents are asked to write a focused clinical question to guide searching; list potential sources of information that will address their question; identify an appropriate study design; describe the search strategy; and identify characteristics of the study that will determine relevance, validity, magnitude of the impact, and clinical significance of the study.

The AFT is scored by comparing responses to a grading rubric.¹² For each question, the rubric specifies explicit grading criteria and gives examples of ideal responses. For example, the first item asks respondents to write a focused clinical question; responses are scored based on their inclusion of the PICO criteria (population, intervention, comparison, and outcome).¹³ Four grading categories are used (*not evident*, *limited*, *strong*, and *excellent*), each corresponding to a specific number of points. For example, a response that does not mention a patient population or that uses an irrelevant or inappropriate descriptor earns 0 points (not evident); use of a single general descriptor constitutes a limited answer (1 point); mentioning one appropriate but not specific descriptor is a strong answer (2 points); and using relevant and appropriate descriptors is excellent (3 points). Each criterion is scored according to these categories, and the sum of points for all criteria is the score for that item. The maximum possible AFT score is 156.

The AFT has been reported to have acceptable internal consistency and good to excellent interrater reliability.¹² McCluskey and Bishop (2009) identified a random sample of 20 AFT tests from a pool of 220 forms completed by occupational therapists, which were then scored by two raters with expert EBP knowledge; interrater reliability ranged from good to excellent for individual AFT items (ICC = 0.68–0.96) and was excellent for total AFT scores

(ICC = 0.91). The calculated Cronbach α was 0.74, indicating satisfactory internal consistency.¹²

Despite the availability of a scoring rubric, we have been concerned that the assignment of scores in the AFT is open to variability. Differences in scores are likely to be greater when multiple raters are involved in evaluating responses and when raters come from different professional backgrounds and have different experiences of EBP. Interrater reliability is the degree to which measurements of the same phenomenon by different raters will yield the same results, or the consistency of results across different raters.¹⁴ The previously reported evaluation of AFT¹² involved only two raters, both with expert knowledge of EBP, and may therefore have underestimated the potential for interrater variability.

The aim of our study was to examine the interrater reliability of the AFT using several raters with differing levels of professional experience.

METHODS

The study was approved by the Human Research Ethics Committee of the University of South Australia and by the Ethics Review Board of the University of Tasmania. There was no protocol violation, and all participants provided written informed consent.

The results presented here are part of a larger investigation into the effectiveness of a journal club in improving the EBP knowledge and skills of allied health professionals.

Participants

The AFT was completed by 55 physiotherapists and occupational therapists who agreed to participate in the larger study before commencing EBP training. The majority of participants (62%) held a bachelor's degree; the remainder had completed postgraduate degrees in different clinical areas. Less than half had prior exposure to research (defined, for the purposes of our study, as participation in the conduct of a research project) or prior EBP training.

Raters

Four physiotherapists with different professional experiences served as raters for the responses to the AFT. Box 1 describes the four raters involved in the study.

Procedure

Of 55 completed AFT questionnaires, 12 were randomly selected for rating. Instructions specified that raters should score each AFT question independently and confidentially, without conferring or comparing ratings. Raters were given 2 weeks to score all questionnaires.

Before the study began, the four raters received training in the form of discussion about the AFT questionnaire and the scoring rubric, as well as collaborative scoring of a sample test. During a practice period, raters independently scored a second sample test, then compared and discussed discrepancies in scores.

Box 1 Professional Characteristics of Raters

	Rater			
	1	2	3	4
Profession	Physiotherapist	Physiotherapist	Physiotherapist	Physiotherapist
Academic background	Master's degree in sports and musculoskeletal physiotherapy	Master's degree in manual and sports physiotherapy	PhD (HS) candidate and health-related master's degrees (physiotherapy, clinical psychology)	PhD (HS) candidate and master's degree in physiotherapy
EBP training	Formal course in EBP	Formal course in EBP	Formal course in EBP	Formal course in EBP
Clinical experience	2 y in OP and hospital setting	5 y in OP setting	Internship/placement only	2 y in hospital setting
EBP-related research experience	1 y	2.5 y	7 y	8.5 y
Teaching experience	None	Occasional (clinical demonstration)	16 y undergraduate, 1 y postgraduate	14 y undergraduate, 9 y postgraduate
Other qualifications	None	Level 2 sports trainer certified by SMA	None	Director of research centre for 5 y

HS = health sciences; EBP = evidence-based practice; OP = outpatient; SMA = Sports Medicine Australia.

Table 1 Interrater Reliability for the Adapted Fresno Test Items and Total Score

Adapted Fresno Test item	ICC (95% CI)
1. Write a focused clinical question for one scenario to help you organize a search of the clinical literature.	0.85 (0.65–0.95)
2. Where might you find answers to these questions? Name as many possible sources of information as you can. List advantages and disadvantages.	0.85 (0.48–0.96)
3. What type of study (design) would best answer your clinical question and why?	0.70 (0.30–0.90)
4. Describe the search strategy you might use in Medline topics, fields, rationale and limits.	0.76 (0.35–0.92)
5. What characteristics of a study determine if it is relevant?	0.22 (–0.07 to 0.61)
6. What characteristics of a study determine its validity?	0.86 (0.50–0.96)
7. What characteristics of the study's findings determine its magnitude and significance?	0.87 (0.63–0.95)
Total score	0.55 (0.06–0.84)

Data analysis

We calculated interrater reliability for total AFT score and individual AFT items using intra-class correlation coefficients (ICC [2, 1]) and 95% CIs. The ICC (2, 1) is used when each item (subject) is measured by each rater and raters are considered representative of a larger population of similar raters.¹⁵ For interpretative purposes, ICC (2, 1) values ≥ 0.80 denote excellent reliability; values between 0.60 and 0.79, moderate reliability; and values < 0.60 , questionable reliability.¹⁶ Thus, higher ICC values suggest greater similarity in rater scores assessing the same test (subject).

RESULTS

Table 1 describes the interrater reliability for individual AFT items and the AFT total score. Interrater reliability was moderate to excellent for items 1 and 7 (see Table 1). While ICCs for items 2–4 and item 6 indicated excellent or moderate reliability, the lower boundary estimate

for the 95% CI suggests that questionable levels of reliability among raters are possible. The ICCs for item 5 and for the total AFT score were low, suggesting considerable variability among raters. For these two items, we plotted the scores to examine trends or patterns (see Figures 1 and 2).

Figure 1 summarizes the performance of the four raters for item 5. Here an obvious trend emerges: raters 1 and 2 tended to give lower scores than raters 3 and 4. Raters 3 and 4 were experienced researchers who held academic appointments, whereas raters 1 and 2 were relatively inexperienced EBP researchers and had recently completed clinical master's degrees.

We observed a similar trend when the total scores were plotted, as shown in Figure 2.

Based on the trends we identified for item 5 and the total score, we grouped the raters into two groups—experienced and inexperienced—and examined interrater

Figure 1 Performance of raters for item 5 of the Adapted Fresno Test

Figure 2 Performance of raters for total Adapted Fresno Test Score

reliability within each group for all items with questionable reliability (items 2–6 and total score). ICCs for these items are shown in Table 2. Reliability estimates for raters 1 and 2 (inexperienced) increased slightly for items 2 and 5 and for total score, but not for other items. For raters 3 and 4 (experienced), ICCs increased considerably, indicating excellent reliability for the total score and for all items except item 4, for which ICC decreased further.

DISCUSSION

Our results suggest that care should be taken to reliably rate scores when using the AFT to determine EBP skills

and knowledge. We found excellent reliability between raters for only two AFT items, which indicates that using multiple raters is likely to increase variability in scoring, and thus potentially introduce Type II errors, when the AFT is used to estimate the effectiveness of EBP training. The AFT is the only known objective test of EBP knowledge and skills; unlike a traditional fixed-response assessment, however, it has features of a performance-assessment measure, including aspects related to the raters, rating scale (i.e., scoring rubric), and rating procedures.^{17,18} Because of these features, the AFT involves some degree of subjective judgment, which makes the assessment prone to variability, particularly when multiple raters are involved. This potential variability in rater responses attenuates the usefulness of the AFT.

Our study found that the interrater ICCs for items 1 and 7 were high, a strong indication that, irrespective of rater background and training, the assessments of these items were reliable. Item 1 asks respondents to write a clinical question about a scenario, which helps to organize a search for relevant literature. Asking a structured, focused clinical question is the essential first step in EBP, and is therefore a critical component of EBP teaching.⁷ There is a standardized format for teaching how to formulate a clinical question, which is not the case for other steps of EBP; learners are required to use the PICO framework, which represents the four elements of an answerable clinical question: population, intervention, comparison, and outcomes.¹³ All individuals who undergo EBP training are expected to know this format. Since all four raters had prior training in EBP, it is not surprising that they all scored this item in a similar way. Item 7, on the other hand, asks about the characteristics of study findings that respondents will consider in determining magnitude and significance (both clinical and statistical); it was not clear to us why this item was rated more reliably than other AFT items. On reviewing the scores, we found that of the seven AFT items, participants scored lowest on item 7, almost half scoring 0. We

Table 2 Reliability Estimates for Inexperienced and Experienced raters

Adapted Fresno Test item	ICC (95% CI)	
	Inexperienced*	Experienced†
2. Where might you find answers to these questions? Name as many possible sources of information as you can. List advantages and disadvantages.	0.96 (0.86–0.99)	0.97 (0.86–0.99)
3. What type of study (design) would best answer your clinical question and why?	0	0.94 (0.80–0.98)
4. Describe the search strategy you might use in Medline topics, fields, rationale and limits.	0.66 (–0.04 to 0.90)	0.70 (0.06–0.91)
5. What characteristics of a study determine if it is relevant?	0.24 (–0.34 to 0.70)	0.95 (0.83–0.99)
6. What characteristics of a study determine its validity?	0.70 (–0.02 to 0.91)	0.98 (0.92–0.99)
Total score	0.58 (–0.26 to 0.88)	0.92 (0.72–0.98)

*Inexperienced: Raters 1 and 2.

†Experienced: Raters 3 and 4.

then examined the completed questionnaires and observed that participants either wrote very short responses to this item or did not answer the question at all. These minimal responses could be expected to reduce variability in interpretation, and therefore increase reliability between raters. Based on our findings for most items in the AFT, therefore, we speculate that if item 7 is answered more fully, the reliability of ratings will be questionable for this item as well.

Interrater reliability for all other items (items 2–6) and for the total score was moderate or poor, but the reliability estimates changed when raters were dichotomized into experienced and inexperienced groups: reliability estimates generally improved for experienced but not for inexperienced raters, which suggests that professional experience may play a role in scoring behaviour. The experienced raters, who were both educators, may have thought and behaved similarly because of the level of assessment experience they had gained through their exposure to teaching. Moreover, being exposed to different learner populations as teachers likely led them to form expectations that were different from those of raters without prior experience in scoring or marking examinations. Although the excellent reliability between the experienced raters does not necessarily mean that their ratings were accurate, we are inclined to think that the judgments of the experienced raters should take precedence over those of the inexperienced raters. The experienced raters, being educators in EBP, skilled in assessment, and highly knowledgeable in their own domains, can be considered experts.^{19,20} The inexperienced raters, on the other hand, can be classified as novices because of their limited exposure to assessment and less extensive experience in EBP teaching and research.^{19,20} To optimize reliability or to make novices perform like experts, the literature suggests making rating criteria and procedures explicit and training raters to apply these criteria.^{21,22} Our study, however, found a lack of interrater reliability despite explicit rating criteria and rater training. While the scoring rubric is helpful in guiding raters, it appears that the training package was not sufficient to achieve reliability in rating the AFT.

The important finding of this study is that all raters are not equal, and therefore EBP educators or researchers should not assume that AFT scores are always accurate. Our study simulated the scenario in multicentre trials that examine EBP interventions, in which outcomes may be assessed by multiple raters from a variety of professional backgrounds. While the use of AFT by two individuals with similar training and background may provide an adequate measure of EBP knowledge and skills, conventional use of the AFT in larger trials is likely to reduce the study's power to truly find an effect of the intervention.

Overall, the variability of scores appears to be related to the raters' level of professional experience: raters with extensive experience in EBP teaching and research are

more likely to be reliable in their scoring behaviour than those who have less exposure to EBP teaching and research. Our results suggest that training alone, without regard to professional experience, is not sufficient to ensure reliability in scoring the AFT.

LIMITATIONS

Our study has some limitations that need to be considered. First, all raters involved in our research were from the same professional discipline (i.e., physiotherapy), and we do not know whether reliability findings would be similar with raters from other allied health disciplines. Second, time and resource constraints permitted us to include only four raters; there are no clear guidelines on how many raters are required for reliability testing, and we do not know whether our results would change with a different number of raters. The sample size may have been too small; larger studies are required to confirm these results. Despite these limitations, our results may provide useful information on rating trends that may exist when the AFT is used to measure EBP knowledge and skills.

CONCLUSION

Implications for practice

Our findings suggest that using the AFT to assess knowledge and skills in EBP may be problematic unless raters are carefully selected and trained. EBP educators and researchers who intend to use the AFT as an outcome measure should identify raters who share a similar professional background (i.e., researchers and educators in EBP) and qualify as experts in EBP. Careful training should then be provided, and reliability should be tested using several sample answers, before raters participate in the actual assessment.

Implications for research

Our findings underscore the need for further research on outcome evaluation instruments in the area of EBP in allied health. EBP researchers should continue to explore evaluation approaches, perhaps developing and testing an instrument that will measure not only knowledge and skills in formulating questions and in searching for and critically appraising evidence, but also the ability to apply evidence to decision making for individual patients or clients. Measurement instruments for EBP domains should be sufficiently robust to account for differences in the interpretation of scoring criteria.

KEY MESSAGES

What is already known on this topic

The AFT is the only objective measure of EBP knowledge and skills that has been tested and applied in allied health. The AFT has acceptable internal consistency and good to excellent interrater reliability. Scoring involves comparing responses to a grading rubric, which is not

similar to a traditional fixed-response assessment. The AFT has features similar to a performance-assessment measure, including aspects related to the raters, rating scale, and rating procedures. Therefore there is some degree of subjective judgment involved in scoring the AFT.

What this study adds

This study suggests that raters for the AFT should be carefully selected and trained to obtain reliable assessment of EBP knowledge and skills. When the AFT is used as an outcome measure, raters should not only share a similar professional background but must also qualify as experts in EBP. There is still a need for further research on outcome evaluation instruments in the area of EBP in allied health.

REFERENCES

1. Metcalfe C, Lewin R, Wisher S, et al. Barriers to implementing the evidence base in four NHS therapies: dietitians, occupational therapists, physiotherapists, speech and language therapists. *Physiotherapy*. 2001;87(8):433–41. [http://dx.doi.org/10.1016/S0031-9406\(05\)65462-4](http://dx.doi.org/10.1016/S0031-9406(05)65462-4)
2. Jette DU, Bacon K, Batty C, et al. Evidence-based practice: beliefs, attitudes, knowledge, and behaviors of physical therapists. *Phys Ther*. 2003;83(9):786–805. Medline:12940766
3. Bennett S, Tooth L, McKenna K, et al. Perceptions of evidence-based practice: a survey of Australian occupational therapists. *Aust Occup Ther J*. 2003;50(1):13–22. <http://dx.doi.org/10.1046/j.1440-1630.2003.00341.x>
4. Iles R, Davidson M. Evidence based practice: a survey of physiotherapists' current practice. *Physiother Res Int*. 2006;11(2):93–103. <http://dx.doi.org/10.1002/pri.328>. Medline:16808090
5. Stevenson K, Lewis M, Hay E. Does physiotherapy management of low back pain change as a result of an evidence-based educational programme? *J Eval Clin Pract*. 2006;12(3):365–75. <http://dx.doi.org/10.1111/j.1365-2753.2006.00565.x>. Medline:16722923
6. Cordell WH, Chisholm CD. Will the real evidence-based medicine please stand up? *Emerg Med*. 2001;23(6):11–4.
7. Dawes M, Summerskill W, Glasziou P, et al; Second International Conference of Evidence-Based Health Care Teachers and Developers. Sicily statement on evidence-based practice. *BMC Med Educ*. 2005;5(1):1. <http://dx.doi.org/10.1186/1472-6920-5-1>. Medline:15634359
8. Ilic D. Teaching evidence-based practice: perspectives from the undergraduate and post-graduate viewpoint. *Ann Acad Med Singapore*. 2009;38(6):559–5. Medline:19565109
9. Flores-Mateo G, Argimon JM. Evidence based practice in post-graduate healthcare education: a systematic review. *BMC Health Serv Res*. 2007;7(1):119. <http://dx.doi.org/10.1186/1472-6963-7-119>. Medline:17655743
10. Tracey JM, Arroll B, Barham P, et al. The validity of general practitioners' self assessment of knowledge: cross sectional study. *BMJ*. 1997;315(7120):1426–8. <http://dx.doi.org/10.1136/bmj.315.7120.1426>. Medline:9418092
11. Khan KS, Awonuga AO, Dwarakanath LS, et al. Assessments in evidence-based medicine workshops: loose connection between perception of knowledge and its objective assessment. *Med Teach*. 2001;23(1):92–4. <http://dx.doi.org/10.1080/01421590150214654>. Medline:11260751
12. McCluskey A, Bishop B. The Adapted Fresno Test of competence in evidence-based practice. *J Contin Educ Health Prof*. 2009;29(2):119–26. <http://dx.doi.org/10.1002/chp.20021>. Medline:19530195
13. Richardson WS, Wilson MC, Nishikawa J, et al. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123(3):A12–3. Medline:7582737
14. Amelang A. Inter-rater reliability of the clinical practice assessment system used to evaluate pre-service teachers at Brigham Young University [master's thesis]. Provo (UT): Brigham Young University; 2009.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–8. <http://dx.doi.org/10.1037/0033-2909.86.2.420>. Medline:18839484
16. Richman J, Makrides L, Prince B. Research methodology and applied statistics. *Physiother Can*. 1980;32:253–7.
17. Scott S. Practical implications of reliability and performance-based assessments. *General Music Today*. 2003;16:18–22.
18. Stecher B. Performance assessment in an era of standards-based educational accountability. Stanford (CA): Stanford Centre for Opportunity Policy in Education, Stanford University; 2010.
19. Royal-Dawson L, Baird J. Is teaching experience necessary for reliable scoring of extended English questions? *Educ Meas*. 2009;28(2):2–8. <http://dx.doi.org/10.1111/j.1745-3992.2009.00142.x>
20. Graham M, Milanowski A, Miller J. Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Center for Educator Compensation Reform; 2012 [updated 2012 Feb; cited 2012 May 14]. Available from: http://cecr.ed.gov/pdfs/Inter_Rater.pdf
21. Barrett S. The impact of training on rater variability. *Int Educ J*. 2001;2(1):49–58.
22. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm*. 2008;65(23):2276–84. <http://dx.doi.org/10.2146/ajhp070364>. Medline:19020196