Published in final edited form as:

J Trauma Acute Care Surg. 2017 February ; 82(2): 426. doi:10.1097/TA.0000000000001275.

Considerations of a test-retest reliability study in injury prevention

Francisco J. Bonilla-Escobar, MD, MSc¹, Catalina Restrepo-Lopera, MD², and Juan Carlos Puyana, MD³

¹Institute for Clinical Research (ICRE), University of Pittsburgh, Pittsburgh, PA, USA; Global Health, Division of Trauma and Surgery, University of Pittsburgh, Pittsburgh, PA, USA; Cisalva Institute, Faculty of Health, Universidad del Valle, Cali, Colombia

²Global Health, Division of Trauma and Surgery, University of Pittsburgh, Pittsburgh, PA, USA

³Critical Care Medicine and Surgery; Global Health, Division of Trauma and Surgery, University of Pittsburgh, Pittsburgh, PA, USA

To the Editor

Reliability measures the capacity of a test to be reproducible. It is expressed by the degree of influence of random errors in a measurement. It is also a necessary but not a sufficient component of accurate measurements. It is critical to identify “operational scales” for a specific outcome to be measured. When doing research, scientists are always searching for measurable outcomes. Observable situations (i.e. injuries, deaths, strokes) are relatively easy outcome to measure; however, more difficult outcomes, such as risk of injuries, require a deeper inquiry and analysis of information, and consequently, reliable and valid instruments are needed.

Recently, Gittelman et al. published the test-retest results of an injury prevention screening tool for children younger than 1 year of age.(1) Reliability is an important component to the designing process of a survey and to be able to get there, researchers should follow different steps that preferably should first include: face validity, content validity, internal consistency, intra and inter-rater agreement, and test-retest agreement. After this, they can move forward to criterion, discriminant, convergent and construct validity.(2)

In a quality improvement program for injury prevention for pediatricians, six screening tools containing 10–20 questions were developed for each well-child visits through 1 year of age (birth, 2 months, 4, 6, 9 months, and 1 year). Questions were adapted from the American Academy of Pediatrics (AAP) Injury Prevention Program (TIPP) program (<http://patiented.solutions.aap.org/Handout-Collection.aspx?categoryid=32033>). Each tool was

Corresponding author: Francisco J. Bonilla-Escobar, Address: 200 Lothrop street, Division of Trauma and Surgery, University of Pittsburgh, Pittsburgh, PA 15213, Office F1263, USA. Telephone: +1-4126470421, fjb9@pitt.edu, francisco.bonilla@correounivalle.edu.co.

Conflict of interest statement: Nothing to declare.

Disclosures of funding received for this work: Nothing to declare.

reviewed by a group of members of AAP's Council on Injury, Violence, and Poison Prevention (content validity), and they were pretested in approximately 10 parents (face validity).(3) These screening tools were subjects of face and content validity even if not defined as such in the manuscript. Nevertheless, in the Gittleman et al. study, the six screening tools were merged into one by removing questions based on repetition and decided only by the authors criteria, resulting in a new screening tool of 30 questions.(1)

The methods for the design of the new screening tool and for the evaluation of the reliability, including the schedules of the assessments, the time interval for the test-retest assessment, and the statistical analysis raise several methodological concerns about the validity of the results.

Firstly, the new screening tool should be subject to a content validity process, through the revision by a group of external experts that are asked to classify the tool items in order to identify their agreement when analyzing the questions, the coverage of the full construct, and to include all the required domains proposed by the AAP.

Secondly, reliability attempts to reduce random error. In this case, the error could come from interviewees (parents), interviewer or contextual factors. The surveys in the discussed study were carried out in the available time of the research assistants. The authors did not mention how the interviewers' schedules could affect the measurement process and the results. The schedules of the interviews could affect also the contextual factors and then the reliability. It is different to do an interview in rush hours than in more flexible hours.

Choosing the optimal time interval for test-retest reliability could be difficult.(4) The test-retest carried out in Gittleman et al. study were separated by a difference of 4 hours. The authors justified this time interval based on a systematic review of cancer patients under palliative care. In fact, the authors of the review, concluded that the time intervals may be short to assess test-retest reliability due to the quick changes in health status of patients under palliative care.(5) Therefore, there is not a valid reason for this short interval and the reliability could be falsely elevated.

Finally, some missing items in the analysis should be explained in the discussion. An internal consistency analysis was not carried out. It is helpful to understand the association between the items in the scale through the Cronbach's coefficient alpha.(4) Measures such as Phi (Φ) for inter-rater reliability could be used as a more robust measure of inter-observer reliability than percent of agreement due to the inclusion of agreement by chance, and the correlation coefficient to identify the association between the scores could give more information about the data. Finally, the Kappa is an important statistic measure when analyzing reliability; however, it possess some limitations such as that it assumes independent ratings but the raters and the rating process were not independent. Kappa is also sensitive to the prevalence of the measured event and trauma is a prevalent but differentially distributed event, therefore the external validity of the measure could be affected. In addition, weights for weighted Kappa are subjective and they could lead to higher values of reliability. That is why the discussion could include some of these points to improve the readers' understanding of the impact of the new screening tool.

In conclusion, the methods carried out to build the proposed screening tool for injury prevention in children should include additional and important aspects of reliability procedures to enhance the results and its internal and external validity. Lack of strategies such as face and content validity, internal consistency, and the use of additional statistical tools to assess the results could lead to wrong interpretations of the results and to lower efficacy values of a tool attempting to measure risk of injuries and our ability to prevent them.

References

1. Gittelman MA, Kincaid M, Denny S, et al. Evaluating the Reliability of an Injury Prevention Screening Tool: A Test-Retest Study. *J Trauma Acute Care Surg.* 2016; 81(4 Suppl 1):S8–S13. [PubMed: 27488487]
2. DeVellis, RF. *Scale Development: Theory and Applications.* SAGE Publications; 2016.
3. Gittelman MA, Denny S, Anzeljc S, et al. A pilot quality improvement program to increase pediatrician injury anticipatory guidance. *J Trauma Acute Care Surg.* 2015; 79(3 Suppl 1):S9–14. [PubMed: 26131790]
4. Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health.* 2007; 10(Suppl 2):S94–S105. [PubMed: 17995479]
5. Paiva CE, Barroso EM, Carneseca EC, et al. A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. *BMC Med Res Methodol.* 2014; 14:8. [PubMed: 24447633]