Available online at www.sciencedirect.com
ScienceDirect
International Journal of Nursing Studies 45 (2008) 1238–1246
INTERNATIONAL JOURNAL OF NURSING STUDIES
www.elsevier.com/ijns

Interpreting interrater reliability coefficients of the Braden scale: A discussion paper

Jan Kottner*, Theo Dassen

Department of Nursing Science, Centre for Humanities and Health Sciences, Charité Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

Received 21 June 2007; received in revised form 8 August 2007; accepted 15 August 2007

Abstract

There are many studies investigating psychometric properties of the Braden scale, a scale that predicts the risk for pressure ulcers. The main focus of these studies is validity as opposed to reliability. In order to estimate the degree of interrater reliability a literature review revealed that numerous statistical approaches and coefficients were used (Pearson's product-moment correlation, Cohen's kappa, overall percentage of agreement, intraclass correlation). These coefficients were calculated for the individual items and the overall Braden score and were used inconsistently. The advantages and limitations of every coefficient are discussed and it is concluded that most of them are inappropriate measures. Therefore, estimating the degree of the Braden scale interrater reliability is limited to a certain extent. It is shown that the intraclass correlation coefficient is an appropriate statistical approach for calculating the interrater reliability of the Braden scale. It is recommended to present intraclass correlation coefficients in combination with the overall percentage of agreement.

Keywords: Interrater reliability; Intraclass correlation; Pressure ulcer; Risk assessment

What is already known about the topic?

- Pearson's product–moment correlation coefficient and the overall proportion of agreement are the most frequently used coefficients for indicating the degree of the overall Braden score interrater reliability.
- Interrater reliability of the individual items was calculated using Cohen's kappa.
- It is assumed that interrater reliability for the Braden scale is high.

What this paper adds

- Pearson's product–moment correlation and Cohen's kappa are inappropriate measures for the interrater reliability of the Braden scale.
- Published interrater reliability coefficients are not comparable and interpreting the degree of the Braden scale interrater reliability is limited.
- The intraclass correlation coefficient in combination with the overall proportion of agreement is recommended for calculating the degree of interrater reliability for single items and the overall Braden score.

1. Introduction

Pressure ulcers are a serious health problem. In 2006, the prevalence of pressure-related wounds in patients at

*Corresponding author. Tel.: +49 30 450 529 054; fax: +49 30 450 529 900.
E-mail address: jan.kottner@charite.de (J. Kottner).

0020-7489/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijnurstu.2007.08.001

risk in Germany was 8.9% for hospitals and 4.6% for nursing homes (Kottner et al., 2007). Pressure ulcers have a significant impact on morbidity and mortality (Allman, 1997) and on the quality of life of the persons affected (Spilsbury et al., 2007). The treatment of pressure ulcers causes high economic costs (Bennett et al., 2004; Gethin et al., 2005). Furthermore, the occurrence of pressure ulcers is regarded as an important indicator for the quality of care (Bates-Jensen et al., 2003). Therefore, pressure ulcer prevention is of utmost importance for daily nursing practice. The first step in pressure ulcer prevention is the identification of individual risks (European Pressure Ulcer Advisory Panel, 1998). Pressure ulcer risk assessment scales were designed to assist nurses in determining whether the person is at risk or not.

So far, more than 20 risk assessment scales have been known worldwide, but only few have been psychometrically tested (Papanikolaou et al., 2007). The risk assessment scale most extensively studied is the Braden scale (Bergstrom et al., 1987a). It is used in different nursing care settings worldwide (Bergstrom et al., 1998). The Braden scale is composed of six items: sensory perception, moisture, activity, mobility, nutrition and friction and shear. The first five items are rated from 1 (most impaired) to 4 (least impaired) and the item "friction and shear" is rated from 1 (problem) to 3 (no problem). The individual item scores are added up, so that the overall Braden score can range from 6 to 23, low scores hereby indicating a high risk for developing pressure ulcers.

As any assessment instrument used in nursing practice or research, the Braden scale should be reliable and valid. Reliability refers to the degree to which measurement error is non-existent in the obtained scores (Polit and Beck, 2004). The larger the amount of error variance the lower the degree of reliability. Data containing a large amount of measurement error will fail to reflect the criterion of interest. Therefore reliability is a prerequisite for validity.

Interrater reliability is a specific aspect of reliability referring to the degree of measurement error due to bias caused by different raters or observers rating the same person or object. With regard to the Braden scale a high degree of interrater reliability is necessary to obtain valid Braden scores in nursing practice and research. Each time the Braden scale is used in research the degree of interrater reliability must be measured, because one has to determine the extent to which variability in the observed outcomes can be attributed to true variation or interrater biases. Recently, Capon et al. (2007) explored the association between the Braden score and several explanatory variables. They discussed a possible lack of reliability in their study but failed to make any attempts to measure it.

Today, there are many studies investigating psychometric properties of the Braden scale (Brown, 2004; Pancorbo-Hidalgo et al., 2006; Papanikolaou et al., 2007). In general, the focus of these studies was put on the predictive validity, whereas the issue of reliability often was of minor interest only. When looking closely at published data focusing on the Braden scale interrater reliability, a variety of statistical methods for computing interrater reliability coefficients can be identified. In a review, Pancorbo-Hidalgo et al. (2006) reported 12 coefficients of Pearson's r ranging from 0.83 to 0.99 and concluded that the interrater reliability is high. Apart from Pearson's r further coefficients are reported: The overall proportion of agreement (p_o) of the Braden score was calculated by several authors (Bergstrom et al., 1987a; Langemo et al., 1991; Watkinson, 1996). The same coefficient was also calculated for an individual item only (Powers et al., 2004). A coefficient frequently used for calculating the interrater reliability for individual items is Cohen's kappa (κ). It was computed by Bours et al. (1999) and Halfens et al. (2000). At the same time κ is used to calculate the interrater reliability of the overall Braden score (Halfens et al., 2000; Vanderwee et al., 2005). In addition, Ooka et al. (1995) and Carlson et al. (1999) calculated the intraclass correlation coefficient (ICC) for the overall score. Bours et al. (1999) computed whether there was a statistically significant difference in the mean scores between pairs of raters for individual items and for the overall Braden score. Obviously, Oot-Giromini (1993) confused two different statistical methods when reporting that "the percentage of agreement … in this study was $r = .83$ to .96" (p. 27). Lewicki et al. (2000) reported an interrater reliability of .90 without any additional explanations.

When looking at the variety of published coefficients representing the interrater reliability of the overall Braden score or individual items, the question is raised which coefficient or statistical method is appropriate.

In order to answer the question whether the interrater reliability of the Braden scale is high or low, this paper presents a review of currently used statistical approaches in measuring the interrater reliability of the Braden scale. The properties of each reported coefficient are discussed and its advantages and limitations are presented. Recommendations are also given for calculating the interrater reliability of the Braden scale and its individual items. Although this discussion focuses on the Braden scale, the statements are also applicable to other assessment scales with a similar structure.

2. Review of interrater reliability coefficients

2.1. Search strategy

The international databases MEDLINE, CINAHL, EMBASE and the German database CARELIT were searched using the terms Braden scale and/or pressure sore