Collegian xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Collegian

journal homepage: www.elsevier.com/locate/coll

Hawks and doves: The influence of nurse assessor stringency and leniency on pass grades in clinical skills assessments

Miranda Daly (RN, MNurs) (Associate Lecturer)[a,*], Yenna Salamonson (RN, PhD) (Associate Professor)[a], Paul J. Glew (RN, BN, BEd, GradCertClinSc(ICN), MNurs, MAAppLingTESOL, EdD) (Senior Lecturer)[a], Bronwyn Everett (RN, PhD) (Associate Professor)[b]

[a] *Western Sydney University, School of Nursing and Midwifery, Centre for Applied Nursing Research, Ingham Institute for Applied Medical Research, Locked Bag 1797, Penrith, NSW 2751, Australia*
[b] *Centre for Applied Nursing Research, Ingham Institute for Applied Medical Research, Western Sydney University, School of Nursing and Midwifery, South Western Sydney Local Health District, Locked Bag 7017, Liverpool BC, NSW 1871, Australia*

ARTICLE INFO

Article history:
Received 19 February 2016
Received in revised form
14 September 2016
Accepted 22 September 2016
Available online xxx

Keywords:
Leniency
Stringency
Skills assessment
Nursing student

ABSTRACT

Background: Rater variability in performance-based assessments is well-documented, yet limited research has explored this in undergraduate nursing programs.
Methods: A prospective follow up study design was used to determine the extent of assessor stringency in clinical skills assessments in an undergraduate nursing program at a large multi-campus university in Sydney, Australia. Grades for students' clinical skills assessments in three units (semesters one, three and five) were extracted from an administrative database. Results were matched to student demographic data (age, gender, language spoken at home, country of birth) and the assessor.
Results: A total of 2339 graded clinical skills assessments of students in the undergraduate nursing program were available for analysis, representing 75% of students enrolled in three nursing skills units. Overseas-born students had lower pass grades than Australian-born students (78% vs. 85%; $p < 0.001$). Significant variability was seen in pass grades across units (χ^2: 32.32, *df*: 2, $p < 0.001$), campuses (χ^2: 17.81, *df*: 2, $p < 0.001$) and assessors, with pass grades ranging from 36% to 100%. In multivariate analysis, students assessed by the most lenient assessors were over seven times more likely to pass than students assessed by the most stringent assessors (AOR: 7.76; CI: 5.64-10.67; $p < 0.001$).
Conclusions: The strongest predictor of a student passing their nursing skill assessment was the leniency of the assessor. A proactive approach to detecting and correcting variability in clinical skills assessments, including reviewing assessor training and support, is needed in light of the high-stakes nature of these assessments.

1. Introduction

Rater variability in grading of theoretical and practical assessments for health professionals has long been identified as problematic, with research spanning more than four decades (Engvik, Kvale, & Havik, 1970; Havik, 1980; Lin et al., 2013; Sebok & Syer, 2015). Numerous factors have also been shown to explain some of the variance in assessment scores. These include: professional background (e.g. clinical versus academic); type of assessment task (e.g. theoretical or practical) (Havik, 1980); rater training (Lin et al., 2013); impact on rater variability; characteristics such as sex/ethnicity of the student (Awaisu, Mohamed, & Al-Efan, 2007); and those of the assessor such as personality (Finn, Cantillon, & Flaherty, 2014). Saal, Downey, and Lahey (1980) identified five major categories of error which may contribute to rater variability; (i) severity or leniency, which refers to the general tendency of the rater to consistently rate students higher or lower than is justified on the basis of their responses; (ii) the halo effect, which appears when the rater takes a holistic approach to the assessment, and fails to distinguish between essential or non-essential content (or correct and incorrect content); (iii) central tendency,

* Corresponding author.
E-mail addresses: miranda.daly@westernsydney.edu.au (M. Daly), y.salamonson@westernsydney.edu.au (Y. Salamonson), p.glew@westernsydney.edu.au (P.J. Glew), bronwyn.everett@sswahs.nsw.gov.au (B. Everett).

http://dx.doi.org/10.1016/j.colegn.2016.09.009
1322-7696/

which occurs when raters are reluctant to use the extremes of a rating scale, resulting in mean scores clustering around the scale midpoint; (iv) restriction of range, which is related to central tendency but also measures the extent to which ratings discriminate among different students with different abilities and (v) inter-rater reliability or agreement, which relates to consistency of agreement between raters, sometimes misconstrued as rating validity, when this could be due to rating bias.

In a study by Barrett (2001) of undergraduate business students, issues concerning marker variability and the possible impact on course results were expressed, with students identifying inconsistencies both between and within markers as concerning. Extreme raters are sometimes described as being either a 'hawk' or a 'dove', indicating some staff are renowned for being a hard (malignant) or an easy (benign) marker (Finn et al., 2014; McManus, Thompson, & Mollon, 2006). Discrepancies between the grading of written assessments by permanently employed or tenured staff and casually employed (also known as sessional) staff have also been reported (Salamonson, Halcomb, Andrew, Peters, & Jackson, 2010), while differences across testing sites have also been shown to contribute to variability in assessment scores for performance-based assessments (Floreck & De Champlain, 2001).

Rater variability in assessment of student performance during objective structured clinical assessments or examinations [OSCAs or OSCEs] (Liao, Hunt, & Chen, 2010; Yanhua & Watson, 2011) and in the clinical setting (McCarthy & Murphy, 2008) have also been reported, often despite raters having undertaken training prior to assessing students. If significant variability exists between and within raters, then assessment of student performance becomes a function of the rater, rather than a reflection of student learning. Many assessment tasks undertaken by nursing students are 'high-stakes' assessments, particularly practical assessments such as OSCAs which are graded on a satisfactory/unsatisfactory basis, with satisfactory performance required for course progression or completion and registration. Apart from the issue of fairness to students, there can be negative consequences for students as a result of being graded by a harsh or inconsistent rater. Further, there are significant costs associated with marking both theoretical and practical assessments in nursing programs, and the validity of these ratings is paramount for quality teaching and learning practice. Rather than presuming that these effects will ultimately be absorbed or work themselves out (Iramaneerat & Yudkowsky, 2007) it is judicious to take a proactive approach to detect and rectify these issues. Hence, the aim of this study was to determine the extent of assessor variability in clinical skills assessments in an undergraduate nursing program.

2. Method

This study was part of a larger multi-phase mixed methods study that sought to explore assessor variability in clinical skills assessments in an undergraduate nursing program. The current paper reports the results of Phase One which examined assessor stringency and leniency in three clinical skills units (subjects) across each year of the nursing program using a prospective follow-up design.

2.1. Study setting

The study was undertaken at a large metropolitan university which offers a three-year Bachelor of Nursing program on three campuses. Approximately 3500 undergraduate students are enrolled across the program. The university has a significant multicultural student population, with almost one quarter (24.5%) coming from low socioeconomic backgrounds (University of Western Sydney, 2014).

2.2. Data collection

The results of students' clinical skills assessments in three units (semesters one, three and five) were extracted from an administrative database. In the current study setting, clinical skills were graded on a pass/fail (satisfactory/unsatisfactory) basis. Results were matched to student demographic data (age, gender, language spoken at home, country of birth) and the assessor. Assessors were categorised as contract or continuing (tenured) staff or casually employed (sessional) staff. Only those who had graded 20 or more students, representing at least two days of assessment, were included in the analysis. This was to ensure that each assessor had undertaken at least two days of assessments during this period, and to ameliorate the likelihood of lower pass rates than usual as the assessors may be 'having a bad day' (McGinn et al., 2004).

As the researchers were employed within the School, a research assistant extracted all data and replaced all identifiers (student number and staff name) with a code to ensure the anonymity of students and assessors. The study was approved by the university Human Research Ethics Committee (H11075).

2.3. Data analysis

Data analysis was performed using SPSS Version 22.0 (2013). We computed the means and standard deviation for continuous variables, and percentages for categorical variables, to summarise sample characteristics. Chi-square test was used to investigate relationships between pass rates and student characteristics. Tertiles were used to split the assessors into three groups (low, mid and high-range) based on the percentage of pass grades allocated. Logistic regression was used to identify predictors of pass grades in clinical skills assessments. Results are reported as adjusted odds ratios with 95% confidence intervals (CI). A p value of <0.05 was considered statistically significant.

3. Results

A total of 2339 graded clinical skills assessments of students in the undergraduate nursing program were available for analysis, representing 75% (2339 of 3131) of students enrolled in three nursing skills or practical units. Overall these results indicated that 80% of students passed their clinical skills assessment, these were assessed by 85 academic staff, almost 80% of whom were casual (Table 1).

3.1. Variability in pass grades across units and campus

The variability in pass grades across the three units ranged from 74% to 85% (χ^2: 32.32, df: 2, $p < 0.001$). Although small, statistically significant differences were also noted in pass rates across campuses, with campus 2 having lower pass grades compared to the other two campuses (χ^2: 17.81, df: 2, $p < 0.001$).

Variability in pass grades by students' demographic factors

Eighty five percent of Australian born students passed their clinical skills assessments, however, only 78% of students born overseas passed their assessment (Fig. 1). This difference was statistically significant ($p < 0.001$). No other differences were noted in pass grades based on age, gender, or language spoken at home.

3.2. Variability in pass grades by assessors

There was a wide range of variability in the percentage of pass grades by assessors, varying from 36% to 100% (Fig. 2). When the

Please cite this article in press as: Daly, M., et al. Hawks and doves: The influence of nurse assessor stringency and leniency on pass grades in clinical skills assessments. *Collegian* (2016), http://dx.doi.org/10.1016/j.colegn.2016.09.009