Studies in Educational Evaluation xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

A longitudinal quantitative investigation into the concurrent validity of self and peer assessment applied to English-Chinese bi-directional interpretation in an undergraduate interpreting course

Chao Han[1]

College of International Studies, Southwest University, 2 Tiansheng Road, Beibei, Chongqing 400715, China

ARTICLE INFO

Keywords:
Concurrent validity
Consecutive interpreting
Peer assessment
Rating accuracy
Self-assessment
Teacher assessment

ABSTRACT

One of the productive lines of research on self-assessment (SA) and peer assessment (PA) concerns their concurrent validity with respect to a criterion measure. However, similar research has rarely been conducted for spoken-language interpreting. This article therefore reports on a longitudinal study that investigated the validity of self and peer ratings on three performance dimensions of English-Chinese consecutive interpretation (i.e., information completeness, fluency of delivery, and target language quality), taking teachers' ratings as a yardstick. Major findings include: although the students as a group were unable to replicate teachers' ratings, they were able to rank-order their performances in a fairly accurate manner and improved their SA and PA accuracy over time. Interpreting directionality seems to moderate the correlational strength of self/teacher ratings and peer/teacher ratings. These results are discussed in relation to previous literature, and pedagogical suggestions are provided to improve SA and PA for bi-directional interpretation.

1. Introduction

Self-assessment (SA) and peer assessment (PA) have been highly valued as important formative assessment tools in language learning and higher education. While SA fosters self-reflection, facilitates self-directed learning, and develops learner autonomy (e.g., Chen, 2008; Falchikov & Boud, 1989), PA encourages reflective learning through observing peers' performance, and can act as a socializing force to improve interpersonal relationships within and between learner groups (Cheng & Warren, 2005). Both SA and PA help students obtain an enhanced understanding of assessment criteria and nurture abilities and skills denied to them when only teachers are involved in assessing their work (Aryadoust, 2015; Cheng & Warren, 2005; Saito, 2008). Although these potential benefits are generally acknowledged and reported in language testing and educational assessment literature (e.g., Butler & Lee, 2010; Chen, 2008), empirical research on the validity of SA and PA has yielded inconsistent results, as can be seen in the literature review below. A number of factors have thus been identified that could affect the validity of SA and PA, and categorized into three groups: 1) the domain or skill being assessed, 2) students' individual attributes, and 3) task characteristics (for detailed review or meta-analysis, see Blanche & Merino, 1989; Boud & Falchikov, 1989; Ross, 1998; see also Butler & Lee, 2010; Chen, 2008). For example, Patri (2002) reported that providing students with peer feedback had a positive effect on the accuracy of their PA of speaking performance, but such effect was not observed for SA. These findings have sharpened our understanding of potential factors moderating the validity of SA and PA, and informed better practice of SA and PA in the language learning context.

Despite the increasing popularity of SA and PA in the context of interpreter training, similar insights have not been developed (see Iaroslavschi, 2011; Lee, 2005; Postigo Pinazo, 2008; Sawyer, 2004). In fact, little research has been conducted to examine the validity of SA and PA applied to language interpretation. Granted, it can be argued that previous findings from second/foreign language testing research could inform and improve SA and PA in interpreter education (Lee, 2017b), since assessing spoken-language interpretation and assessing second/foreign language performance, particularly speaking performance, are essentially rater-mediated assessment of language-based oral production. However, there are at least two salient differences that set them apart. Firstly, assessing speaking usually concerns monolingual performance, while interpretation assessment necessarily involves bi-directional interpretation between at least two languages, usually an interpreter's mother tongue (often known as an A language) and another working language (as a B language). Unless the rater is a

E-mail address: chao.research@gmail.com.
[1] Dr. Chao Han is associate professor in the College of International Studies at Southwest University, Chongqing, China. His research interests include interpreter performance testing and assessment, tertiary-level interpreter training and education, and research methodology. ORCID: https://orcid.org/0000-0002-6712-0555.

https://doi.org/10.1016/j.stueduc.2018.01.001
Received 8 July 2017; Received in revised form 22 November 2017; Accepted 11 January 2018

balanced bilingual, assessing into A interpretation can be different from assessing into B interpretation. Secondly, while similar assessment criteria are used in assessing speaking and interpreting (e.g., fluency, grammar, lexical diversity), evaluating interpretation entails judgement of substantive and functional equivalence between source-language input and target-language output, a dominant concern in interpretation assessment (e.g., Gile, 1999; Han, 2016; Lee, 2008; Wang, Napier, Goswell, & Carmichael, 2015). Such equivalence, also referred to as inter-textuality in translation studies, seldom figures prominently in assessing monolingual speaking performance. It is thus felt that investigating how directionality and inter-textuality play out in SA and PA of interpretation has the potential to extend and enrich the growing body of literature concerning the validity of SA and PA.

Of particular interest and at the heart of the present study is therefore the validity of SA and PA, specifically their concurrent validity with respect to teacher assessment (TA). In some literature, concurrent validity of SA and PA is synonymous to the accuracy of self and peer ratings or the competency of students, compared with their teachers, in assessing their own and peers' performance (e.g., AlFallay, 2004; Blanche & Merino, 1989; Boud & Falchikov, 1989; Lew, Alwis, & Schmidt, 2010).

Investigating the validity or accuracy of SA and PA is of practical importance.[2] Firstly, the accuracy of SA and PA is believed to be a condition of learner autonomy (Blanche & Merino, 1989), and effective learners are realistic judges of their performance (Boud & Falchikov, 1989). As such, there is a practical need for students to develop abilities to assess themselves and their peers accurately. Secondly, given that many teachers doubt the validity of SA and PA, they are reluctant to incorporate SA or PA results into final grades, and even refuse to implement SA and PA in classroom (Saito & Fujita, 2004). More investigation could help elucidate the strengths and weaknesses of SA and PA, leading to informed decision making. Thirdly, a more practical reason for such an investigation is that little knowledge has been obtained regarding to what extent SA and PA practiced in interpreter training are accurate and trustworthy. The validity of SA and PA is under-researched and poorly understood in the field of interpretation testing and evaluation. Against this background, the study attempts to examine the validity of self and peer ratings on English-Chinese bi-directional consecutive interpretation.

2. Literature review

This section first examines some commonly used methods to investigate concurrent validity of SA and PA. It then synthesizes pertinent literature in second/foreign language testing and educational assessment, particularly SA and PA of speaking performance. Finally, the section reviews SA and PA applied to language interpretation and problematizes the practice of and research on SA and PA in interpreter education.

2.1. Concurrent validity of SA and PA

To examine concurrent validity, SA and PA results are usually correlated with more objective and trustworthy criterion measures such as objective tests, final grades and teachers' ratings (Butler & Lee, 2010; Chen, 2008; Ross, 1998). By far, the most commonly used metric is Pearson product-moment correlation coefficients (Blanche & Merino, 1989; Ross, 1998). In many cases, TA results or teachers' ratings are chosen as the criterion measure, as they are generally easier and quicker to obtain than other measures (see AlFallay, 2004; Boud & Falchikov, 1989; Chen, 2008). However, using TA results as an objective yardstick is controversial, due to variability of teacher ratings which poses a potential threat to construct validity (Aryadoust, 2015; Ward, Gruppen, & Regehr, 2002). Furthermore, based on the model of concurrent validity, TA measures need to be demonstrated as a valid criterion, which may imply a tricky loop of endless validation (Bachman, 1990). To justify TA as a trustworthy and valid benchmark, some sound practices have been suggested in the literature: 1) ensuring that scoring schemes (e.g., assessment criteria, scalar descriptors, scoring rubrics) provide an operational definition of the constructs we want to measure (Knoch, 2011); 2) selecting raters who have extensive rating experience and receive rater training so that ratings could properly reflect the degree and extent of the constructs being measured (Cheng & Warren, 2008); 3) ascertaining reliability of TA results before further statistical analysis (Ward et al., 2002), and 4) using pooled ratings from multiple raters as fairer measures (Sullivan & Hall, 1997). Another commonly used method is to employ significance testing. Depending on the characteristics of collected data, inferential statistical analyses such as *t* test, analysis of variance and Wilcoxon signed-ranks test have been used (e.g., Babaii, Taghaddomi, & Pashmforoosh, 2016; Chen, 2008; Sullivan & Hall, 1997). Such tests are able to reveal whether SA or PA results are, on average, different from those of TA in a statistically significant manner.

2.2. SA and PA of speaking

Although much of the literature on SA and/or PA of second/foreign language performance seems to concern writing, a number of empirical studies are dedicated to SA and/or PA of speaking and there are also a few excellent reviews (Ross, 1998; Saito, 2008). Regarding SA of speaking, Ross (1998) in a meta-analytic study reported considerable variation in the students' ability to accurately estimate their own speaking performance, as correlation coefficients between SA and criterion measures ranged from 0.09 to 0.78. It was also found that SA of speaking was less accurate than SA of other skills such as reading, listening, and writing. This finding was partially explained by the fact that in many foreign language programs exposure to the written word precedes extensive practice of listening and speaking, which may thus affect the relative accuracy of SA. In a longitudinal investigation into SA of oral presentations, Chen (2008) found that the Spearman's correlation coefficient between students and teachers' ratings improved over ten weeks from 0.55 ($\rho < 0.05$) in the first assessment cycle to 0.79 ($\rho < 0.05$) in the second cycle. In addition, using Wilcoxon sign-ranks test, a statistically significant difference ($\rho < 0.05$) was detected between students' and teachers' ratings in the first cycle, but did not occur in the second cycle. Chen (2008) thus concluded that students' knowledge and skills obtained from the first cycle of SA might have contributed to improvement in SA accuracy. In another longitudinal study, Babaii et al. (2016) also found that Pearson correlation between students' and teachers' ratings increased from 0.73 ($\rho < 0.01$) to 0.90 ($\rho < 0.01$), after the students were provided with detailed assessment criteria and received rater training. Both these two longitudinal studies suggest that with practice and training SA accuracy can improve over time.

When it comes to PA of speaking, Magin and Helmore (2001) analyzed a multi-year data and found fairly strong correlation (Pearson's *r*) between students' and teacher' ratings on oral presentations, ranging from 0.48 to 0.69. Cheng and Warren (2005) also reported agreement between student PA and teacher's marks on three scoring dimensions of oral presentation (i.e., preparation & content, delivery, language) for three different classes. Using *t*-test, they also found that no statistically significant differences were detected between students' and teacher's marks, with only one exception. It is worth noting, however, that Cheng and Warren (2005) defined agreement as that student's mean mark lied within one standard deviation of the teacher's mean mark. In addition, Saito (2008) reported that rater training did not lead to substantial

[2] In the present study, the terms "concurrent validity" and "accuracy" were used interchangeably. When correlational methods are used to examine concurrent validity of student self or peer ratings, "accuracy" could be interpreted as the accurate rank-ordering of performances with respect to a criterion measure (see otherwise, Wolfe, Jiao, & Song, 2015).