Thinking Skills and Creativity 15 (2015) 13–25

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: http://www.elsevier.com/locate/tsc

Rater effects in creativity assessment: A mixed methods investigation

Haiying Long[a,1], Weiguo Pang[b,*]

[a] Department of Leadership and Professional Studies, Florida International University, 11200 SW 8th Street, Miami, FL 33199, USA
[b] School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 17 October 2014
Accepted 23 October 2014
Available online 18 November 2014

Keywords:
Creativity assessment
Rater effects
Generalizability theory
Rater cognition
Mixed methods research

ABSTRACT

Rater effects in assessment are defined as the idiosyncrasies that exist in rater behaviors and cognitive process. They are composed of two aspects: the analysis of raw rating and rater cognition. This study employed mixed methods research to examine the two aspects of rater effects in creativity assessment that relies on raters' personal judgment. Quantitative data were collected from 2160 raw ratings made by 45 raters in three group and were analyzed by generalizability theory. Qualitative data were collected from raters' explanation of rationales for rating and their answers for questions about rating process as well as from 12 in-depth interviews and both were analyzed by framing analysis. The results indicated that the dependability coefficients were low for all the three rater groups, which were further explained by the variations and inconsistencies in raters' rating procedure, use of rating scales, and their beliefs about creativity.

1. Introduction

Using human judges to score individual works or behaviors is not an uncommon measurement process in social sciences. Requiring teachers to score responses of constructed items in standardized tests is such an instance (Crisp, 2012). Other examples include counseling psychologists measuring high school students' degree of pathology and intensity of violence; graduate students in social work program assigning scores to evaluate children's behaviors at their homes; principals observing classroom teaching and evaluating teachers' performance. In creativity studies, researchers also rely heavily on raters' judgment of the products generated from participants, including the ideas produced in divergent thinking tests, creative solutions to real world problems, and artifacts of creative writing and art (Author, 2014b).

Research on creativity raters in recent years (e.g., Kaufman & Baer, 2012; Kaufman, Gentile, & Baer, 2005; Kaufman, Baer, & Cole, 2009; Kaufman, Baer, & Gentile, 2004; Kaufman, Baer, Cole, & Sexton, 2008) mostly focused on the influence of raters with different expertise on the results of assessment. However, this line of research does not shed light on the issue of rater effects (Hung, Chen, & Chen, 2012). The present research aims to fill this gap by employing mixed methods methodology to examine the rater effects in assessing the creativity of two science tasks. The examination of this issue is crucial because raters and their judgment are an indispensable part of the assessment. In addition, examining rater effects

* Corresponding author. Tel.: +86 21 6223 2910.
E-mail addresses: haiying.long@fiu.edu (H. Long), wgpang@psy.ecnu.edu.cn (W. Pang).
[1] Tel.: +1 305 348 3228.

http://dx.doi.org/10.1016/j.tsc.2014.10.004
1871-1871/

reveals the behaviors and cognitive process of raters during the assessment, which would further facilitate possible trainings for raters in the future, hence, help improve the assessment procedure.

2. Literature review

2.1. Rater effects

When human judgment is involved in the measurement, the measurement process becomes more complex than it appears. After the products are made, they are presented before the raters who assign ratings based on the criteria provided. The final decision made about individuals' traits, then, is not only determined by how individuals perform in the tasks but also by how raters perform in the assessment process. Traditionally, only the consistency or agreement among raters is the demonstration of rater performance and most rater trainings focus on how to achieve a high rater agreement. However, since the 1970s, researchers began to realize that no matter how many trainings and monitorings that raters go through before and during the assessment, their performance is still greatly affected by the idiosyncrasies that exist in their behaviors and cognitive process (Charney, 1984; Hamp-Lyons, 2007; Noyes, 1963). These idiosyncrasies are defined as rater effects (Wolfe, 2004).

According to Wolfe and McVay (2012), there are two major aspects of rater effects. One is the manifest level of the effects, which is reflected by the raw ratings assigned by raters. The other is the underlying level, which is shown by raters' thinking process or rater cognition. These two aspects are closely associated with measurement reliability and validity. On one hand, the raw ratings are a potential source of measurement errors in estimating reliability among raters (Campbell & Fiske, 1959; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991). This is as Guilford (1936) stated, "... raters are...subject to all the errors to which humankind must plead guilty" (p. 272). On the other hand, raters' idiosyncrasies interfere with the construct measured (Cumming, Kantor, & Powers, 2001), thus become a construct-irrelevant variance, which is one of the major threats to construct validity (Messick, 1995). Rater cognition is also substantive aspect of validity that focuses on how judges evaluate works as well as whether judges' processes are consistent with their interpretation of the construct (Messick, 1995). Its significance in validation process is highlighted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999),

> If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided. (p. 19)

In addition, under Kane's (2006) framework of argument-based validation, rater cognition-related factors, such as, whether raters follow rating criteria and whether they use the categories in the intended manner, help establish an interpretive argument.

Furthermore, the two aspects of rater effects are closely related to each other in a way that the understanding of raw ratings, or even the consistency among raters, "depends on an intuitive, if not explicit, understanding of rater cognition" (Bejar, 2012, p. 3). However, as two aspects of rater effects, raw ratings and rater cognition are examined by different research methodologies. The ratings are typically analyzed by quantitative methodologies such as, generalizability theory and latent trait measurement models. Rater cognition is investigated by qualitative methodologies, such as, think aloud and verbal protocol analysis (Wolfe & McVay, 2012).

2.2. Use of generalizability theory in analysis of raw scores

Under the framework of classical test theory (CCT), the evaluation decision of raters is often expressed as a raw score and the consistency among raters is estimated by interrater reliability. In general, there are three categories of interrater reliability: consensus, consistency, and measurement estimates. When two raters do not share common meanings of the rating scales but are able to be "consistent in classifying the phenomenon according to his or her own definition of the scale" (Stemler, Consistency Estimates section, para. 1), a situation that resembles creativity assessment, particularly Consensual Assessment Technique, it is best to use consistency estimate as indicated by Cronbach's alpha.

However, according to Stemler (2004), Cronbach's alpha has a few weaknesses. For example, raters may have different interpretations in rating scores and rating categories, and it is highly sensitive to the distribution of the data. In addition, even a high alpha does not necessarily suggest a high consensus among raters because a high alpha may result from a large number of raters. What's more, because CCT attributes variation of observed scores only to a true score and a random error, the raw score under this framework cannot reflect variations of raters, such as, rater severity, interactions between raters and other aspects in the evaluation, and other random errors. For these reasons, Cronbach (2004, p. 394) himself claimed, "Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components" and he and his colleagues further developed generalizability (G) theory (Cronbach et al., 1963; Shavelson & Webb, 1991).

In G theory, an observed score is assumed to be a sample drawn from a universe of possible observations and each aspect of the measurement is defined as a facet. Each facet involved and the interactions among them are variations of the universe scores and the theory aims to more accurately disentangle the contribution of each error to the total variation (Shavelson