Benefits of self and peer assessment in teaching psychology essay

The present study is an attempt to investigate the pedagogical benefits of self- and peer-assessment in teaching writing in EFL context. The specific questions this study aims to answer are whether self-assessment and peer-assessment of writing performance over time contribute to improvement in writing performance as well as self- and peer-rating, and if so, which assessment method brings about more contribution in these regards.

Moreover, this study seeks to investigate the effect of several variables on the practice of self- and peer-assessment. The first variable is the Big Five personality traits, the effects of which are studied on learners' (a) rating accuracy in self- and peer-assessment of writing performance, and (b) degree of improvement in writing performance as well as self- and peer-rating accuracy through practicing self- and peer-assessment over time.

Finally, this research focuses on the choice of scoring method (holistic vs. analytic) in self- and peer-assessment, and examines (a) if there is any significant relationship between holistic and analytic self- or peer-assessment, (b) which scoring method is more accurate, and (c) if practicing self- and peer-assessment over time improve the holistic and analytic self- and peer-rating accuracy.

The participants of this study consisted of 198 Iranian male and female adult undergraduate students studying different English language majors, including English literature, English translation, and English language teaching, at Allameh Tabataba'i University, the South Tehran Teacher Training Branch of Islamic Azad University, and Alborz Higher Education Institute. The needed data for this study were collected from the participants

while attending the course Advanced Writing, which is a two-credit 16-week course normally offered to the students in the third term of the bachelor's program. Since intact classes were used, the classes were arbitrarily assigned to treatment and control groups by using a " semi-randomization procedure" (Mackey & Gass, 2005, p. 143). Table 3. 1 shows how the participants were assigned to the treatment and control groups.

3. 3. 3 Personality traits inventory. As mentioned in the previous chapters, several personality tests and inventories have been employed by psychologists to investigate the relationship between personality and other psychological constructs. Examples of these tests are Myers-Briggs Type Indicator, Minnesota Multiphasic Personality Inventory (MMPI), 16 Personality Factor questionnaire, Eysenck's three-factor personality questionnaire, and finally Costa and McCrae's NEO-PI-R and NEO-FFI.

According to Feist and Feist (2006), most personality psychologists have voted for McCrae and Costa's NEO-PI-R and NEO-FFI based on a five-factor theory of personality since cross-cultural support and stability over time have been observed for their theory and tests. Moreover, the five-factor theory of personality, also known as the Big Five, is inclusive in relation to other personality traits and dimensions such as the ones in Cattell's and Eysenck's (John & Srivastava, 1999; Sdorow, 1998).

The Big Five personality traits include: (1) Neuroticism (2) Extraversion, (3) Openness to Experience, (4) Agreeableness, and (5) Consciousness. There are two inventories for measuring the Big Five, one called The Revised NEO Personality Inventory (NEO-PI-R), and the other The NEO Five-Factor

Inventory (NEO-FFI). Both of these inventories are scored on a Likert scale of 1 to 5. The difference between the two lies in the fact that the NEO-PI-R consists of 240 items which allows a comprehensive assessment of adult personality on all the above-mentioned personality dimensions and their facets; however, the NEO-FFI consists of 60 items, and it is recommended to be used at time constraints, and when global information about personality is required. Both of these two inventories are of two forms: form S for self-rating and form R for observer-rating (i. e. for spouse or peer ratings) to validate the results of S form. According to the manual to the inventories, the NEO-FFI has been derived from NEO-PI-R via the validimax method (a factor analysis variant) and correlation with acceptable validity indices, taking into account that shorter scales usually tend to trade precision for time and convenience (Costa & McCrae, 1992). Comprehensive technical information and statistics about the validity of the NEO-PI-R and NEO-FFI are available in the manual to the inventories (Costa & McCrae, 1992), an account of which is out of the scope of this chapter and study.

Since general English proficiency is an important factor in determining writing performance, it was ideal to have similar treatment and control groups in terms of general English proficiency; therefore, the proficiency means of the groups were compared. Since the Kolmogorov-Smirnov and Shapiro-Wilk tests showed the data was not normally distributed ($p < .05$), the nonparametric Kruskal-Wallis Test was used to compare the proficiency means of the groups, which showed the groups were not significantly different; $H = 1.84$, $df = 2$, $p > .05$.

3. 4. 2 Rater training. The researcher of this study as well as two EFL instructors, who are experienced English language teachers at institute and university level holding Master's and Bachelor's in TEFL, rated the writing performances of the participants. The rater training was conducted based on the procedures of Educational Testing Service elaborated by Weigle (2002) and the guidelines outlined by Jacobs et al. (1981). To check the holistic and analytic interrater reliability of the raters, 30 paragraphs by the self-assessment group on the pretest were rated by the raters, and the interrater reliabilities for holistic and analytic scorings were calculated via intraclass correlation (ICC), which turned out to be . 94 and . 92 respectively. It should be noted that the raters scored each paragraph first holistically and then analytically. The holistic scores were given based on percentage. This was done based on the suggestion of Falchikov and Goldfinch (2000) in order to have both analytic and holistic scoring based on a similar range, and to provide the raters, and then the self- and peer-raters, with a familiar range.

3. 4. 3 Self-/peer-assessment training and practice. After the administration of the pretest, the writing course actually started with a two-hour session on the basics of paragraph writing such as topic, topic sentence, supporting sentences, coherence, and cohesion. Most of the instructions were based on Arnaudet and Barrett's Paragraph Development (1990). The second session, the ESL composition profile accompanied by the related pamphlet containing the full descriptors, illustrations, and anchor scripts was introduced to the students. The third session was also spent on the scale elaboration, and then sample paragraphs including the ones written on the pretest were given to the students to be rated first holistically and then analytically based on the

scale and the anchor scripts. The students' ratings were then compared with those of the raters, and the rating ambiguities were discussed and resolved by the instructors during the session.

After the sessions spent on the introduction of the scale by the instructors and rating practice by the students, one method of paragraph development was introduced to the students every session. Having done the book exercises, the students were given a choice of two topics for paragraph writing. In the peer-assessment group, the participants exchanged their paragraphs with those of their peers for peer-assessment; however, the participants of the self-assessment group rated their own paragraphs. This was done for nine sessions afterwards since there were as a whole nine paragraph development methods introduced to the students. The students were told that self- and peer-rating data were to be used in partial determination of the class participation grade for each student.

After the ninth session, a posttest was also administered to check the improvement of the participants in writing performance and rating accuracy. Every session, the participants' paragraphs from the previous session were rated by the raters both holistically and analytically, and the necessary feedback was given to the students. At times, some sample paragraphs were also read aloud by the students to be rated by both the teachers and students together in the class. Moreover, the participants in the peer-assessment group compared their own ratings with those of the raters every session. It should be noted that the control and treatment groups were similar in terms of the amount of instructions they received on paragraph

writing; however, self-/peer-assessment practice was only carried out in the treatment groups.

3. 5. 2 Research design. With regard to the variables and the research questions of the present study, the following research methods and designs were utilized. The first research method employed in this study was quasi-experimental with the pretest-posttest nonequivalent-groups design (Best & Kahn, 2006). There were two treatment groups and one control group in this study. The treatment groups received writing instructions accompanied by peer-assessment or self-assessment training and practice, and the control group receives paragraph writing instructions without any peer-assessment and self-assessment training and practice.

To analyze the collected data and test the hypotheses of this study, different statistical measures were employed. First, the Kolmogorov-Smirnov and Shapiro-Wilk tests were run to check the normality of the data; then parametric and non-parametric statistics were selected accordingly. Next, non-parametric Wilcoxon Signed Rank and Kruskal-Wallis tests, and parametric t test were run to compare the means of the groups. Whenever the effect of a covariate was existent, the analysis of covariance (ANCOVA) was employed to compare the means controlling for the effect of the covariate. When the means of the groups were found to be different, Games Howell test was run to do the post hoc multiple comparisons. Finally, the Pearson correlation coefficient was calculated to investigate relationships between the variables of this study.