A Methodologically Improved Study on Raters’ Personality and Rating Severity in Writing Assessment

Yu Zhu1, Andy Shui-Lung Fung2, and Liuyan Yang1

Abstract
Personality is an inherent rater’s characteristic influencing rating severity, but very few studies examined their relationship and the findings were inconclusive. This study aimed to re-investigate the relationship between raters’ personality and rating severity with more control on relevant variables and more reliable analysis of rating severity. Female novice raters (n = 28) from a demographically homogeneous background were recruited to rate on two occasions essays written by 111 students in an intermediate-level Chinese as a foreign language program. Raters’ personality traits were measured using the complete version of NEO-PI-R. Many-faceted Rasch measurement model and repeated measurement were applied to yield more robust estimates of rating severity. In addition, rating order effect was carefully controlled. Extroversion was found to be positively correlated with severity, r(26) = .495, p = .010. Furthermore, Extroversion was found to be a valid predictor of severity, t(24) = 2.792, p = .010, R² = .21, Cohen’s d = .77, Hattie’s r = .37. Practical implications for developing more individualized online rater calibration for large-scale writing assessments were discussed, followed by limitations of the present study.

Keywords
personality traits, rating severity, second language writing assessment, Chinese

Introduction
It has been repeatedly found by researchers that some raters are systematically more severe or lenient than others in various subjective ratings, a phenomenon termed “rater effects” by scholars such as Myford and Wolfe (2003). Research on rater effects in second language writing has focused on raters’ characteristics such as educational background (e.g., Landy & Farr, 1980), language background (e.g., Kim, 2009), and rating experience (e.g., Bachman et al., 1995; Barkaoui, 2010; Schoonen et al., 1997; Weigle, 1994, 1998). However, relatively little attention has been paid to personality, an inherent and intrinsic characteristic of raters. Rating severity research focusing on rater personality dates back to 1980s, but as McNeal (2019, p. 19) summarized, many of the studies were “disjointed” and “fragmented” as irrelevant traits such as rater self-esteem, need for achievement, or hostility were included. This situation gradually changed when more and more researchers began to employ reliable psychological scales of personality in their measurement of rater personality in the 1990s. A milestone on this path was established by Kane and colleagues (1995), when a particular personality trait of raters (i.e., agreeableness) was suspected and proved for the first time to be an important cause of rater leniency. Strangely enough, for decades after 1995, only very few related studies have been seen in publications (i.e., Alaei et al., 2014; Bernardin et al., 2000, 2009; Carrell, 1995; Yun et al., 2005). Moreover, findings in these studies are inconsistent or even conflicting with each other. For instance, although each of the previously cited literature found personality to be correlated with severity, there are also studies indicating no relationship between these two variables (e.g., Dewberry et al., 2013; Esfandiari, 2019). Another example is that rater conscientiousness is found to be a positive predictor of severity (e.g., Bernardin et al., 2000, 2009), but it is reported as an invalid predictor in some other studies (e.g., Yun et al., 2005). Therefore, Harari et al.’s (2015) meta-analysis of 21 relevant
studies was both timely and valuable, which reported rater agreeableness and leniency having a moderately positive relationship ($\rho = .25$), extraversion and neuroticism both having a smaller positive relationship with leniency ($\rho = .12$), and conscientiousness and severity an even slightly smaller positive relationship with leniency ($\rho = .10$). However, it should be noted that the majority of those 21 studies being reviewed by Harari and colleagues (2015) are in the job performance field. It is still unclear in language testing context in general and in L2 writing assessment in particular, whether or not or how rater personality correlates with rater severity. Studies with different designs and/or in different fields may inform researchers and practitioners with more reliable and generalizable findings about the association between raters’ personality and their rating severity. With this in mind, the present study attempted to re-investigate the relationship between raters’ personality and their rating severity, with a particular focus on writing Chinese as a second language, which hopefully offers new insights into writing assessment.

**Literature Review**

In Section “Personality and Rating Severity,” a brief review of previous studies that explored the relationship between rater personality and severity is presented, followed by suggestions that might bring improvements in research method. Sections “Measuring Personality: Trait or Type” through “Rater Drift: Repeated or Single Measurement(s)” elaborate the rationale for suggested methodological improvement.

**Personality and Rating Severity**

Attempts to investigate the relationship between rater personality and rating severity started in the early 1980s. Using the Eysenck Personality Questionnaire, Branthwaite et al. (1981) were only able to identify the connection between the subscale of Lie and the students’ essay scores among all the subscales of personality. Moreover, they failed to offer a meaningful explanation for this finding due to a lack of definitive interpretation of the Lie subscale.

One of the pioneering studies to measure personality of both raters and writers was Carrell (1995), which examined how ratings might be affected by participants’ personality and genres of writing. Her study found that (a) intuitive-type raters had a tendency to assign lower scores to essays than their Sensing type counterparts; (b) feeling-type raters tended to be less severe in rating than those who fall into the thinking type; (c) the scores given by extroverts, sensing, or judging-type raters for narrative essays produced by feeling-type or introvert-type writers appeared to be most highly elevated; (d) introverts, sensing, feeling, or perceiving raters tended to give high scores to argumentative essays written by Introvert writers. These findings implied an association between raters’ personality and their rating severity.

Using NEO Five-Factor Inventory (FFI) (Costa & McCrae, 1992), NEO-FFI henceforth, Bernardin et al. (2000) investigated whether Conscientiousness and Agreeableness could predict rating leniency/severity. They reported that Conscientiousness and rating severity were positively correlated, while Agreeableness and rating severity were negatively correlated. Furthermore, raters with low Conscientiousness but high Agreeableness were found to have assigned the most lenient scores. As a step further, Bernardin et al. (2009) explored how Agreeableness and Conscientiousness may affect rating accuracy and severity when raters’ accountability level was low. Students ($n = 126$) were recruited to rate their peers after participating in group exercises during a management course. The findings were consistent with Bernardin et al. (2000).

In general, findings regarding personality domains other than Agreeableness and Conscientiousness or personality as a whole are less conclusive. For example, Alaei et al. (2014) found no statistically significant correlations between raters’ holistic ratings and their personality traits measured by a NEO-FFI inventory, but as for analytical ratings, they found that essay content scores and Agreeableness were positively correlated ($r = .787, p = .018$), which was consistent with Bernardin et al. (2000, 2009), but vocabulary score and raters’ level of Conscientiousness were also positively correlated ($r = .889, p = .003$), which was not in accord with the two previous studies.

As can be seen from the above review, the number of studies dedicated to the relationship between rater personality and rating severity is very limited. In addition, different approaches to measuring personality and different research designs render comparisons of previous findings rather difficult.

Methodologically speaking, studies reviewed in this section might benefit from a careful reconsideration of how to measure both personality and severity, attention to rater drift and order effect, and control of raters’ gender and experience. The rationales are elaborated in the following sections.

**Measuring Personality: Trait or Type**

Approaches to measuring personality characteristics basically fall into two broad categories: trait or type. The most representative inventory for the type approach is the Myers-Briggs Type Indicator (MBTI), developed by Myers and McCaulley (1985), and the most popular scale for the trait approach is the Revised NEO Personality Inventory (NEO-PI-R), proposed by Costa and McCrae (1992).

The MBTI inventory, an instrument of 94 items, intends to identify a respondent’s personality along the following four dimensions: extroversion–introversion (E/I), sensing–intuition (S/N), thinking–feeling (T/F), and judgment–perception (J/P). Responses to the MBTI render a respondent being labeled as one of the 16 possible combinations of personality types (e.g., ESTJ, ENFP, etc.).
The NEO-PI-R inventory offers quantitative measurements of personality traits to each respondent, based on the theory of five-factor model (FFM) of personality. It includes five domains and each domain contains six facets: Neuroticism (anxiety, hostility, depression, self-consciousness, impulsiveness, and vulnerability to stress), Extraversion (warmth, gregariousness, assertiveness, activity, excitement seeking, and positive emotion), Openness (fantasy, aesthetics, feelings, actions, ideas, and values), Agreeableness (trust, straightforwardness, altruism, compliance, modesty, and tendermindedness), and Conscientiousness (competence, order, dutifulness, achievement striving, self-discipline, and deliberation). As each facet is further measured through eight Likert-type-scale items, the inventory consists of 240 items in total. The NEO-PI-R inventory is more popular in academic research situations, while the MBTI is widely applied in counseling and business training settings (Furnham et al., 2003).

Costa (1996) detailed the satisfactory psychometric properties (i.e., alpha coefficients, content validity, convergent and discriminant validity, consensual validity, and factorial validity) of the NEO-PI-R. In addition, a number of reasons for NEO-PI-R to be recommended as a tool of personality assessment were also listed, including its known correlations with other personality instruments such as MBTI and other FFM instruments, a relatively small amount of time required for responding, a non-offensive style to respondents, and its high accessibility (both self-report and observer rating versions, as well as increasingly available translated versions). No such positive reviews were available for MBTI. Furthermore, Pittenger (1993) pointed out that there lacks sufficient evidence for one to believe that personality has exactly 16 unique types.

It is notable that many studies (e.g., Bernardin et al., 2000, 2009, Alaei et al., 2014) applying the trait approach to personality preferred the NEO-FFI (i.e., a 60-item shortened version of NEO-PI-R) based on practical considerations. While offering convenience, NEO-FFI does compromise its reliability and validity (Aluja et al., 2005). Therefore, the current study used the complete version of NEO-PI-R to ensure a more accurate measurement of personality.

**Estimates of Rater Severity: Many-Faceted Rasch Measurement Approach or Other Methods**

In the practice of large-scale writing assessments, it is fairly possible that severity estimates still vary significantly across raters even after a thorough rater calibration (e.g., Engelhard, 1992; Engelhard, 1994; Weigle, 1998). Engelhard (1992) grouped various approaches to examining rater severity into three categories, namely, analysis of variance, structural equation, and many-faceted Rasch measurement (MFRM) model (Linacre, 1989). He further argued that the first two approaches were inadequate because they were based on raw scores, which did not use individual rating as the unit of analysis, were not linear representations of rater severity, and did not allow direct comparisons among raters and between rater severity and other facets of the measurement. Engelhard regarded the MFRM model as a promising solution which makes raters’ rating of different sets of essays comparable. This echoed Linacre (1989), who believed that MFRM is the model that guarantees objectivity in examinations for it permits calibration of various facets (e.g., examinees’ ability, raters’ harshness, and task difficulty) on an interval scale that is generalizable beyond specific judging contexts. Later on, Myford and Wolfe (2003) provided an elaborated explanation of the MFRM approach with a focus on its advantage in addressing potential interactions among rater severity and other facets of the performance measurement. As a rigid measurement approach, MFRM has gained increasing popularity in assessment research and dominated studies regarding rater severity/leniency effects since the 1990s.

However, when it comes to the relationship between rater severity and personality, none of the existing studies applied the MFRM approach to obtaining estimates of raters’ harshness. In the present study, we used estimates of rater severity/leniency based on MFRM instead of raw scores.

**Order, Experience, and Gender: With Control or Not**

Order effect has been reported as “slide effect” in earlier literature in the 1960s (e.g., Bracht, 1967; Godshalk et al., 1966). It occurs when the order in which raters are rated affects the ratings they give (Hopkins, 1998). Myford and Wolfe (2003) speculated that it is an indication of “fatigue or boredom setting in” or “a shifting of the raters’ standards as the scoring session proceeds” (p. 401). Although order effect is not easily detectable (Myford and Wolfe, 2003) and less commonly examined (Irmaneerat & Yudkowski, 2007), studies on rater effects normally take different measures to counterbalance the order of essays to be rated to ward off its potential influence on essay ratings and severity estimation (e.g., Knoch et al., 2007; Schaefer, 2008). Therefore, the current study went to great lengths to randomize the scripts to control the potential order effect.

Experience effect is well studied in writing assessment. Many studies reported the differences between experienced and novice raters in their rating process and evaluation criteria (e.g., Cumming, 1990; Erdosy, 2004; Weigle, 1999). Others found differences in severity between rater groups with various levels of rating experience (e.g., Song & Caruso, 1996; Sweedler-Brown, 1985). Novice raters are more vulnerable to rater effects than experienced ones. In other words, they are more likely to be affected by their personalities while rating essays. For better control of potential experience effect, the current study used novice raters only.

Gender effect of rater is a research topic more thoroughly examined for oral performance (e.g., Aryadoust, 2016; Zhu et al., 2015). It is notable that many studies (e.g., Bernardin et al., 2000, 2009, Alaei et al., 2014) applying the trait approach to personality preferred the NEO-FFI (i.e., a 60-item shortened version of NEO-PI-R) based on practical considerations. While offering convenience, NEO-FFI does compromise its reliability and validity (Aluja et al., 2005). Therefore, the current study used the complete version of NEO-PI-R to ensure a more accurate measurement of personality.

**Order, Experience, and Gender: With Control or Not**

Order effect has been reported as “slide effect” in earlier literature in the 1960s (e.g., Bracht, 1967; Godshalk et al., 1966). It occurs when the order in which raters are rated affects the ratings they give (Hopkins, 1998). Myford and Wolfe (2003) speculated that it is an indication of “fatigue or boredom setting in” or “a shifting of the raters’ standards as the scoring session proceeds” (p. 401). Although order effect is not easily detectable (Myford and Wolfe, 2003) and less commonly examined (Irmaneerat & Yudkowski, 2007), studies on rater effects normally take different measures to counterbalance the order of essays to be rated to ward off its potential influence on essay ratings and severity estimation (e.g., Knoch et al., 2007; Schaefer, 2008). Therefore, the current study went to great lengths to randomize the scripts to control the potential order effect.

Experience effect is well studied in writing assessment. Many studies reported the differences between experienced and novice raters in their rating process and evaluation criteria (e.g., Cumming, 1990; Erdosy, 2004; Weigle, 1999). Others found differences in severity between rater groups with various levels of rating experience (e.g., Song & Caruso, 1996; Sweedler-Brown, 1985). Novice raters are more vulnerable to rater effects than experienced ones. In other words, they are more likely to be affected by their personalities while rating essays. For better control of potential experience effect, the current study used novice raters only.

Gender effect of rater is a research topic more thoroughly examined for oral performance (e.g., Aryadoust, 2016; Zhu et al., 2015). It is notable that many studies (e.g., Bernardin et al., 2000, 2009, Alaei et al., 2014) applying the trait approach to personality preferred the NEO-FFI (i.e., a 60-item shortened version of NEO-PI-R) based on practical considerations. While offering convenience, NEO-FFI does compromise its reliability and validity (Aluja et al., 2005). Therefore, the current study used the complete version of NEO-PI-R to ensure a more accurate measurement of personality.
Rater Drift: Repeated or Single Measurement(s)

In writing assessment, rater drift effect was first reported by Lunz and Stahl (1990), where inconsistencies in rater severity were found across various time periods for all the three contexts of rating. Similarly, Wilson and Case (2000) also noticed a general pattern which indicated that table leaders’ severity estimates drifted more dramatically than those of the fellow raters of each table. In a more thorough investigation on rater drift effect, Congdon and McQueen (2000) revealed explicitly how serious the effect could be even for trained raters: over 60% of their severity varied significantly on the seventh day from the first day. Moreover, individual rater’s severity on each day was found to differ dramatically from their single, on-average estimate for the whole rating period. Wolfe et al. (2001) also reported on this effect. All these studies confirmed a fact that rater severity is not as stable as commonly believed. With this in mind, the current study applied a repeated measurement design to gauge rater severity.

Research Questions

Based on the above literature review, the present study intends to investigate the following research questions:

1. Does essay raters’ personality score correlate with their severity level, when rater severity is repeatedly measured with MFRM, rater personality measured with NEO-PI-R, and with control of gender, experience, and order effect?
2. Can essay raters’ personality score predict their severity level?

Null hypotheses for each of these two research questions were set and two-tailed statistical tests were performed in the analysis following data collection.

Method

Participants

Raters of the present study were 28 master students in the program of Chinese as a Foreign Language (CFL) Education at a top-ranked research university in China. As Dai and colleagues (2004) cautioned that the scores of NEO-PI-R might be significantly affected by demographic factors such as age, education, and gender, the study recruited volunteering raters with careful control of the above-mentioned variables. More specifically, they were all females, aged 21 to 23, in the third semester of the program, with no experience of essay rating for large scale writing assessment.

The essays to be rated were written by 111 volunteers who were studying intermediate Chinese at the same university as the raters. Over two-thirds of them were females and about half were from Thailand or Indonesia, 23% from South Korea, Japan, or Russia, and the rest of them were from over 20 countries around the world. All raters and student writers were informed of the nature of this research and gave their verbal consent for their involvement.

Materials

Essays and benchmark samples. The writing task of the present study is a simulation of the writing part of the New HSK Level 5. Since it was launched by the Office of Chinese Language Council International in 2009, the six-level New HSK has become the most authoritative proficiency test widely taken by CFL learners all over the world. The New HSK Level 5 writing test contains two tasks: to write a short essay with the five keywords provided and a short essay based on the picture given. The length of each essay is no less than 80 characters. The time limit for a writing test is 40 min. The present study administered two tests (see Appendix A), with a short break of 5 min in between. Each test, following the New HSK Level 5, contains one essay with key words as prompt (hereinafter referred to as “KW task”) and one with a picture as prompt (hereinafter referred to as “PIC task”).

All 111 student writers took part in the writing tests, who were all registered in the HSK-5 preparation class. There was a short break of 5 min between the two tests. During essay writing, students had no access to the internet or dictionaries in any form. As the students were required to take the formal HSK Test scheduled 2 weeks after the simulated tests, they were motivated to take the simulation seriously.

To control rating order effect, each student’s writing competence needs to be pre-determined. Four expert raters were invited to pre-rate all the scripts of the two writing tests. Two of the raters had experience of teaching CFL writing courses for over 5 years, and a doctoral degree in Chinese linguistics or foreign language education. The other two had been teaching CFL writing for more than 10 years and had a master’s degree in Chinese language and literature. Furthermore, they all had participated in rating large scale CFL assessments, including the HSK Test. Details of the rating scale are provided in Section “Rating scale.”

A script, which includes the two essays in a writing test, would be considered as a benchmark if both of the two essays had received exactly the same rating from at least three of the four experts. The distribution of benchmark samples is listed in Table 1.
Table 1. Benchmark Samples for Each Writing Test.

| Score | No. of scripts |
|-------|---------------|
| 0     | 1             |
| 1     | 2             |
| 2     | 3             |
| 3     | 4             |
| 4     | 3             |
| 5     | 2             |
| Total | 15            |

Rating scale. The rating scale adopted by the present study was the same as the official rubric for the New HSK Level 5 writing test (see Appendices B and C). The rubric provides overall descriptions for the six bands (0–5) and break-down descriptions for grammar and vocabulary performance within each band. For each essay, a rater needs to assign an appropriate band, and then make specific judgment about performance on grammar and vocabulary respectively. If grammar or vocabulary performance generally matches the corresponding description, a score of 2 should be given; 3 points will be given to slightly better performance than described, or 1 point in case of slightly inferior performance. Note that the highest possible score for grammar or vocabulary in band 5 is 2, and the lowest possible score for grammar or vocabulary in band 1 is 2. After conversion, a test-taker’s writing scores could range from 0 to 21.

NEO-PI-R. The present study applied a Chinese version of the NEO-PI-R (Yang et al., 1999), which possessed satisfactory construct validity and reasonable reliabilities as reported in Dai et al. (2004). Based on data collected for this study, the Cronbach’s alpha for each domain of the scale is as follows: Neuroticism (α = .811), Extroversion (α = .842), Openness (α = .742), Agreeableness (α = .706), and Conscientiousness (α = .853).

Procedure

Randomization of the scripts. After the 15 benchmarks were chosen, the remaining 96 scripts for each test were randomized and assigned to the raters. As two tests were administered, scripts from the second test were subject to the same randomization procedure. The steps are as follows:

1. 96 scripts were evenly divided into four sets (i.e., high, upper middle, lower middle, low) based on the score given by four expert raters, each containing 24 scripts;
2. 8 scripts are randomly drawn from each set to form one of three groups of 32 scripts;
3. Every two of the three script groups are put together to form a batch for rating. For instance, script Groups 1 and 2 would form Batch A, script Groups 1 and 3 would form Batch B, and Groups 2 and 3 would form Batch C. Each batch contained 64 scripts.

Administration of NEO-PI-R. Prior to rater calibration, the raters were asked to respond in a non-anonymous manner to the Chinese version of NEO-PI-R, on condition that their responses would remain confidential and only be used for research purposes.

Rater calibration and formal rating. As there were two writing tests, the ratings were done separately, with a 1-month interval. At the beginning of each rating occasion, a calibration session was carried out, lasting about 2 hr according to the following procedure. The trainer first introduced the rating scale and answered raters’ questions about the scale. Then the raters were asked to rate six benchmark samples, with one sample representing each of the six bands. Next, five benchmark samples were assigned, representing Bands 1 to 5. The last set of benchmark samples included two from Band 3, one from Band 2, and one from Band 4. The raters were required to complete the rating independently and were then given feedback and allowed to discuss among themselves and/or with the trainer until reaching a unanimous decision about the score of each benchmark sample.

After calibration, the 28 raters were appointed randomly into Group 1 (n = 10), Group 2 (n = 9), and Group 3 (n = 9). Each group was responsible for rating one of the three script batches. Rating the scripts for one writing test generally lasted about 2.5 hr.

To find out whether or not rater severity varied noticeably over time, each rating occasion was artificially divided into three sessions (for the purpose of statistical analysis only), corresponding to each rater’s rating of the first and second 21 scripts, and the remaining 22 scripts respectively.

Analyses and Results

The MFRM Analysis

The model applied for FACETS analysis. Linacre (1989) extended Rasch measurement models to MFRM, which has been used widely to examine rater severity (e.g., Engelhard, 1992, 1994); The five-facet model (writer competence, rater severity, session rigorousness, essay difficulty, and item difficulty) applied in the present study can be expressed as follows:

\[
\ln \left( \frac{P_{nkilm}}{P_{nkilm-1}} \right) = \beta_n - \delta_i - \zeta_j - \eta_k - \lambda_l - \tau_m
\]

where \(P_{nkilm}\) is the probability of writer \(n\) being rated \(m\) during session \(j\) for essay \(k\) on item \(l\) by rater \(i\); \(P_{nkilm-1}\) is the probability of writer \(n\) being rated \(m - 1\) during session \(j\) for essay \(k\) on item \(l\) by rater \(i\); \(\beta_n\) is the writing competence of writer \(n\), and \(n = 1, 2, 3, \ldots 77\) for the first test, \(n = 1, 2, 3, \ldots 79\) for the second test; \(\delta_i\) is the severity of rater \(i\), and \(i = 1, 2, 3, \ldots 29\) for the first test, \(i = 1, 2, 3, \ldots 28\) for the second test; \(\zeta_j\) is the rigorousness of rating session \(j\), and
The FACETS measures of rater severity and related facets. The above-mentioned MFRM detected strong evidence of variance in rating severity across raters. According to the standards introduced by Myford and Wolfe (2004), the raters in the present study were found to have very different severity estimates in the first rating occasion, with $\chi^2 = 285$, $df = 27$, $p < .01$, Separation = 3.04, Strata = 4.38, and Reliability of Separation = 0.90. Similar results were found for the second rating occasion, where $\chi^2 = 231.3$, $df = 26$, $p < .01$, Separation = 2.75, Strata = 4.00, and Reliability of Separation = 0.88. Findings remained basically the same when rating data for the two tests were combined, which suggested that severity estimates varied notably across raters ($\chi^2 = 313.1$, $df = 27$, $p < .01$, Separation = 3.40, Strata = 4.86, Reliability of Separation = 0.92). The raters’ severity estimates were between 0.46 and 1.25.

Table 2 offered individual rater’s severity estimates together with relevant statistics for each rating occasion and combined data. The infit and outfit mean squares for the rater severity estimates for each rating occasion and for the combined data were almost all found to be within the range of 0.5 to 1.5, indicating the quality of severity estimation was good enough (Linacre, 2012, p. 269). The combined data resulted in better severity estimates. Moreover, the analysis of the rater facet indicated that the raters were all stricter than expected because even the most “lenient” rater (i.e., rater CDD in Table 2) had a severity estimate substantially higher than zero.

Valid observations for FACETS estimate of writer competence for Tests 1 and 2 were 77 and 79, respectively, out of the 96 student writers. Analysis of the writer competence facet showed that students’ writing ability varied greatly (see Table 3). The quality of writer competence estimates was not as satisfying as that of the rater severity; however, it can be seen that the majority of the infit or outfit mean squares still fell into the range between 0.5 and 1.5. We consider a FACETS estimate as distorted only when its infit and/or outfit mean square was bigger than 2, based on Linacre (2012). Although the writers’ Chinese proficiency as a whole was

### Table 2. Raters’ Severity and Infit/Outfit Mean Square.

| Rater | Severity in Test 1 | Infit, outfit | Severity in Test 2 | Infit, outfit | Severity from both tests | Infit, outfit |
|-------|--------------------|---------------|--------------------|---------------|--------------------------|---------------|
| CDD   | 0.62               | 1.25, 1.28    | 0.41               | 1.19, 1.20    | 0.46                     | 1.25, 1.25    |
| XT    | 0.72               | 1.36, 1.41    | 0.61               | 0.84, 0.85    | 0.60                     | 1.34, 1.36    |
| LJJ   | 0.61               | 0.91, 0.91    | 0.65               | 1.07, 1.07    | 0.64                     | 1.22, 1.22    |
| ZQ    | 0.73               | 1.21, 1.24    | 0.54               | 1.24, 1.25    | 0.65                     | 1.16, 1.17    |
| ZSH   | 0.91               | 1.39, 1.43    | 0.58               | 0.85, 0.87    | 0.65                     | 0.85, 0.85    |
| ZJL   | 0.86               | 1.04, 1.06    | 0.55               | 0.98, 0.98    | 0.69                     | 0.92, 0.93    |
| GQ    | 0.99               | 1.02, 1.00    | 0.54               | 0.87, 0.89    | 0.73                     | 0.84, 0.86    |
| CDJ   | 1.09               | 1.04, 1.12    | 0.59               | 1.06, 1.06    | 0.74                     | 1.12, 1.12    |
| WHJ   | 1.03               | 0.98, 0.99    | 0.67               | 1.12, 1.12    | 0.75                     | 1.09, 1.09    |
| MQT   | 1.04               | 0.92, 1.01    | 0.67               | 1.25, 1.26    | 0.75                     | 1.10, 1.10    |
| CMH   | 1.03               | 1.32, 1.40    | 0.67               | 1.25, 1.26    | 0.75                     | 1.10, 1.10    |
| YLD   | 0.95               | 0.94, 1.00    | 0.74               | 0.93, 0.94    | 0.76                     | 0.85, 0.86    |
| JY    | 0.74               | 0.81, 0.81    | 0.98               | 0.93, 0.97    | 0.78                     | 0.87, 0.87    |
| YYY   | 1.06               | 0.98, 0.97    | 0.70               | 0.99, 1.01    | 0.79                     | 0.88, 0.88    |
| TYR   | 1.19               | 1.02, 1.04    | 0.59               | 1.02, 1.05    | 0.80                     | 1.08, 1.10    |
| XYR   | 1.07               | 1.06, 1.08    | 0.73               | 0.98, 0.98    | 0.82                     | 1.05, 1.06    |
| ZLP   | 1.00               | 0.85, 0.84    | 0.78               | 0.94, 0.96    | 0.82                     | 0.99, 1.01    |
| FGY   | 1.24               | 0.71, 0.70    | 0.62               | 0.79, 0.79    | 0.83                     | 0.86, 0.87    |
| LB    | 1.16               | 0.88, 0.89    | 0.85               | 1.06, 1.07    | 0.88                     | 1.05, 1.06    |
| TLY   | 1.22               | 1.28, 1.41    | 0.82               | 1.32, 1.33    | 0.89                     | 1.30, 1.30    |
| LS    | 1.31               | 0.78, 0.80    | 0.68               | 0.78, 0.79    | 0.89                     | 0.86, 0.87    |
| WAQ   | 1.26               | 0.84, 0.86    | 0.79               | 0.74, 0.76    | 0.92                     | 0.67, 0.68    |
| LDX   | 1.27               | 0.86, 0.96    | 0.85               | 0.87, 0.91    | 0.94                     | 0.74, 0.77    |
| XTT   | 0.94               | 1.02, 1.00    | 1.15               | 1.05, 1.10    | 0.94                     | 1.18, 1.19    |
| KZL   | 1.20               | 0.87, 0.90    | 0.88               | 1.23, 1.25    | 0.94                     | 1.05, 1.06    |
| GQX   | 1.09               | 0.88, 0.94    | 1.19               | 0.73, 0.78    | 1.02                     | 0.86, 0.88    |
| ZTT   | 1.38               | 0.75, 0.84    | 1.14               | 0.86, 0.90    | 1.11                     | 0.88, 0.91    |
| LXX   | 1.56               | 0.87, 0.95    | 1.25               | 0.83, 0.83    |                          |               |

Mean severity, Mean SE; [infit range] [outfit range]  
1.05, 0.07; [0.69, 1.44] 0.74, 0.07; [0.74, 1.38] 0.81, 0.05; [0.67, 1.34] 0.68, 1.36}
quite homogeneous, their writing abilities were quite diverse and 10% of them were found to mismatch the model. But when data for the two tests were combined, misfitting writers dropped sharply to less than 3%. It is also notable that variance of writer competence reduced greatly from Test 1 to Test 2, which could be due to the regression to the mean effect common in test–retest situations (Dallal, 2012).

For the facet of rating sessions, when the data of the two tests were combined, each of the six rating sessions had an estimate of difficulty of 0.00, with infit and outfit mean squares slightly varied from 0.89 to 1.24. It can be concluded that there is little variance across the six rating sessions.

Statistics for the essay facet were listed in Table 4. When the data of the two tests were combined, the first writing task of Test 1 (KW1 henceforth) was found to be significantly harder than all the other three tasks, $\chi^2(3) = 283.80, p < .01$, Separation = 8.37, Strata = 11.49, Reliability of Separation = 0.99.

Apart from the five main facets, various interaction effects involving rater facet were also explored (i.e., interactions between rater and either essay, session, or item facet). We tried to test these three interactions by adding them one-by-one into the five facets model as specified in Section “The model applied for FACETS analysis” and found that the model with an interaction between raters and essays was the only one that matched our data. This meant raters might behave differently in terms of their severity for some essays. For example, while rater 104 was estimated as more lenient than rater 301 in general, it was found in the bias/interaction report of FACETS that for rating KW1, rater 104 actually was severer than rater 301.

Figure 1 shows the relative position of the estimates of the five facets (i.e., rater severity, writer competence, sessions, tasks, and items) according to the table of All Facet Vertical Rulers of FACETS output.

### Table 3. The Result of the Analysis of Writer Competence Facet.

| Writer competence estimate | $\chi^2(df)$, $p$ | Separation, reliability | Min. | Max. | Infit mean square | Outfit mean square |
|----------------------------|------------------|------------------------|------|------|------------------|-------------------|
| Test 1                     | 4,477(76), <.01  | 7.06, 0.98             | -1.93| 1.45 | [0.45, 2.34],    | [0.35, 2.62],     |
|                            |                  |                        |      |      | 10% ∈ (2.0, +∞) | 12% ∈ (2.0, +∞)  |
|                            |                  |                        |      |      | 13% ∈ (1.5, 2.0) | 10% ∈ (1.5, 2.0)  |
|                            |                  |                        |      |      | 69% ∈ [0.5, 1.5] | 70% ∈ [0.5, 1.5]  |
|                            |                  |                        |      |      | 8% ∈ [0, 0.5)    | 8% ∈ [0, 0.5)     |
| Test 2                     | 1,814.1(78), <.01 | 4.83, 0.96             | -1.85| 0.72 | [0.37, 2.46],    | [0.38, 2.52],     |
|                            |                  |                        |      |      | 5% ∈ (2.0, +∞)  | 5% ∈ (2.0, +∞)    |
|                            |                  |                        |      |      | 22% ∈ (1.5, 2.0) | 22% ∈ (1.5, 2.0)  |
|                            |                  |                        |      |      | 57% ∈ [0.5, 1.5] | 57% ∈ [0.5, 1.5]  |
|                            |                  |                        |      |      | 16% ∈ [0, 0.5)   | 16% ∈ [0, 0.5)    |
| Combination of the two tests | 3,576.4(78), <.01 | 6.72, 0.98             | -1.38| 0.80 | [0.33, 1.99],    | [0.33, 2.02],     |
|                            |                  |                        |      |      | 0% ∈ (2.0, +∞)  | 3% ∈ (2.0, +∞)    |
|                            |                  |                        |      |      | 13% ∈ (1.5, 2.0) | 13% ∈ (1.5, 2.0)  |
|                            |                  |                        |      |      | 73% ∈ [0.5, 1.5] | 70% ∈ [0.5, 1.5]  |
|                            |                  |                        |      |      | 14% ∈ [0, 0.5)   | 14% ∈ [0, 0.5)    |

### Table 4. Estimates of Writing Task Difficulties Based on Combined Data of Tests 1 and 2.

| Writing task estimate | Measure in logit | SE | Infit mean square | Outfit mean square |
|-----------------------|------------------|----|------------------|-------------------|
| KW 2                  | -.09             | .02| 0.99             | 1.00              |
| PIC 1                 | -.09             | .02| 0.93             | 0.94              |
| PIC 2                 | -.07             | .02| 1.12             | 1.13              |
| KW 1                  | .25              | .02| 0.98             | 0.98              |

Raters’ Personality Traits as Measured by NEO-PI-R

Instead of domain raw scores, varimax rotated principal component scores (VRPCS) were calculated for each of the five domains of the NEO-PI-R scale. VRPCS were believed to be more reliable, accurate, and valid measurement than domain raw scores because they addressed the issue of non-orthogonality of the domain scores (Costa & McCrae, 1992).
Table 5. Estimates of Item Facet Based on the Combined Data of Tests 1 and 2.

| Items/rating domains | Measure in logit | SE  | Infit mean square | Outfit mean square |
|----------------------|-----------------|-----|------------------|-------------------|
| General              | -.72            | .01 | 1.44             | 1.45              |
| Grammar              | .33             | .01 | 0.82             | 0.83              |
| Vocabulary           | .39             | .01 | 0.75             | 0.76              |

and these advantages were further evidenced in both the lab and field study (DeCostanza & colleagues, 2017). Table 6 is a summary of the descriptive statistics of each personality domain of the 28 raters.

Figure 2 is a box-plot of the personality traits scores of the raters. It shows that the raters in general had a relatively higher level of Agreeableness and a relatively lower level of Extroversion.

**Correlations Between Raters’ Personality Traits and Rating Severity**

Cook’s $D$ was computed to identify any potential outliers in the observed data. A data point would be considered highly influential on fitted values if the percentile value of its Cook’s $D$ is near or above 50 in the $F(p, n-p)$ distribution, while any percentile values of Cook’s $D$ between 10 and 50 suggest their corresponding data points as candidates of outliers. Accordingly, observation number 28, with a percentile value of Cook’s $D$ being 41.3, and observation number 14 with a percentile value of Cook’s $D$ as 38.3 were identified as suspects of outliers. Therefore, these two observations were deleted for the statistical analyses that followed.

Table 7 showed that Extroversion correlated positively at a notable level with rating severity.

**Regressing Raters’ Personality Traits on Their Rating Severity**

Although correlation analysis revealed that rating severity was correlated positively with raters’ level of extroversion, identifying valid predictors of rating severity requires a regression analysis. Using stepwise selection method in SPSS, linear regression models were fitted and the result was summarized in Table 8.

Table 8 shows that Extroversion is the only valid predictor of severity, with a large effect size as defined by Cohen (1988) or within the zone of desired effects according to Hattie (2009). Costa and McCrae (1992) described extroverts as sociable, assertive, active, and talkative. In addition, extroverts are believed to like excitement and stimulation. In the context of L2 writing assessment, it would not be surprising if highly extroverted raters find the writings of foreign learners in general to be somewhat less stimulating than those of native writers and consequently assign lower scores without reasonable and prudent rationale.

**Discussions and Implications**

The fact that rating severity yielded from the second test was very different from that of the first test strongly suggested that in some cases, especially for novice raters who in general not yet developed consistent levels of severity, raters’ severity may not be constant as they are supposed to be. Instead, they may fluctuate noticeably across rating occasions when a rater is judging the same writer’s parallel writings, using the same rating rubrics. This rater drift effect (Wilson & Case, 2000) was also evident by the significant interaction between raters and essay as revealed by the FACETS analysis (see Section “The FACETS measures of rater severity and related facets” for details). Thus, to obtain a more robust conclusion about the relationship between rating severity and personality traits, researchers need to adopt a repeated measurement design, as this study did, rather than relying on findings based on a single rating occasion of single or multiple writing task(s).

One more improvement this study made in the research design is the procedures to eliminate the potential order effect of essay rating, as is evidenced in the non-significant interaction effect between rater and rating session. Figure 2 also demonstrated graphically that the six rating sessions were at about the same level. With assurance of no rating order effects present, one can have more confidence in the findings of the study.

Applying FACETS estimates of raters’ severity and measuring personality with a complete version of NEO-PI-R allow a more accurate and reliable analysis of the relationship between raters’ rating severity and their personality traits.

In this study, severity was found to be positively correlated with raters’ Extroversion in the context of assessing writing of Chinese as a second language. A statistically significant positive correlation between Extroversion and severity has not been reported in any previous studies. But, clues of such a correlation have been reported as preliminary findings of some studies. For example, based on examinations of rating scores, think-aloud protocol and retrospective interview of four raters in an English L2 argumentative writings context in Korea, Choi and Lee (2019) provided a brief explanation of how extroverted raters differed in their rating behavior from those with an introverted personality type. One of the raters with high Extroversion trait reported during the interview:

When I read the prompt for the first time, I drew an outline in my head like I was writing the essay too and when the students’
Figure 1. FACETS graphic output for each Facet, based on a combined data of Tests 1 and 2.
essays seemed different than what I had in mind, I looked at the ideas more carefully. I have this personality that makes me want to be convinced by different or new ideas and I would be very critical before I am convinced.

It seems that native speaker raters with Extroversion trait tended to be harsher when the writings did not meet their expectations, which is quite common in L2 writing situations. However, with only four raters being included in Choi and Lee (2019), such findings should be referred to with caution, and further studies are needed, both qualitative and quantitative, to draw a convincing conclusion.

Other than examining the relationship between raters’ personality and severity, the study also attempted to regress raters’ personality measurements on their rating severity. The model was fitted without influential data points to acquire a higher level of model accuracy and had a higher portion of variation explained by the model. It was found that Extroversion positively predicted raters’ severity, \( b = .007, t(25) = 2.792, p = .010 \), and it explained a noticeable proportion of variance in rating severity, \( R^2 = .21, F(1, 24) = 7.794, p = .010 \).

Exploring whether or not raters’ personality traits can predict their severity of rating has practical implications for rater calibration of large-scale writing assessment. In recent years, online rater training is gaining increasing popularity due to its convenience and efficiency. By knowing a rater’s tendency of being too severe or lenient in rating in advance, trainers can develop more individualized training materials and deliver them to each target raters via various online systems. However, in most cases, a rater’s severity is unknown prior to his or her actual rating. This makes searching predictors of rating severity meaningful, especially when such variables are relatively stable and easily measurable.

Since this study revealed that extroversion is a valid predictor of rater severity, rater trainers of large-scale writing tests can simply ask each rater to respond to the 48 items of the Extroversion domain of the NEO-PI-R inventory prior to rater calibration. It will take only 6 min for a rater to finish the subscale of Extroversion. Trainers can then identify those “at-risk” raters and send specially designed training materials via the internet to them. Such an approach can be a useful complement to more generic and centralized rater calibration/training. Of course, more convincing evidence based on more sophisticated design and analysis is needed to justify such a training approach. The relationship between raters’ personality or cognitive strategy and accuracy of their rating are other directions that are worth further investigation in the future as suggested by some recent studies (e.g., Wang et al., 2017; Zhang, 2016). With regards to innovative methods for relevant studies, one may refer to Wang and Engelhard (2017) and Engelhard et al. (2018), which suggested to combine multifocal lens model and Rasch measurement theory to evaluate the quality of human rating in writing assessments.

The sample size of the raters of the present study seems not sufficiently large enough as compared to relevant studies such as Bernardin et al. (2000, 2009), both of which used a sample size of over 100. However, it should be noted that those two studies recruited peer-reviewers of undergraduate students instead of using native or expert raters, which is a more conventional practice in L2 writing assessment.

### Table 6. Descriptive Statistics of the Raters’ Personality as Measured by NEO-PI-R.

| Personality traits  | M   | SD  |
|---------------------|-----|-----|
| Neuroticism (N)     | 43.41 | 7.03 |
| Extraversion (E)    | 39.10 | 8.79 |
| Openness (O)        | 54.88 | 8.48 |
| Agreeableness (A)   | 58.53 | 8.38 |
| Conscientiousness (C) | 48.08 | 8.99 |

### Table 7. Pearson’s Correlations between Severities and Personality.

| Personality traits | \( r \) | \( p \) | 95% CI               |
|--------------------|--------|--------|---------------------|
| Neuroticism        | \( -.040 \) | .85    | \([- .43, .32]\)   |
| Extraversion       | \( .495^* \) | .01    | \([.36, .72]\)    |
| Openness           | \( -.230 \) | .26    | \([- .60, .20]\)  |
| Agreeableness      | \( -.330 \) | .10    | \([- .56, -.07]\) |
| Conscientiousness  | \( -.100 \) | .51    | \([- .42, -.23]\) |

Note. Severity was estimated by FACETS using rating data of both Tests 1 and test 2. Here \( n = 26 \) because we deleted two outliers (see Section “Correlations between raters’ personality traits and rating severity” for details).

*indicates statistically significant at \( a = 0.05 \) level.

---

*Choi & Lee (2019)*.
contexts. Moreover, previous studies have used a sample size similar to the present study, such as Carrell (1995) \(n = 20\) and Alaei et al. (2014) \(n = 31\).

Another piece of evidence that enhanced our confidence in the findings is the quite satisfactory effect sizes achieved throughout the statistical analyses of the present study. This effect size was for a large part due to our strict rater selection procedure, which resulted in a relatively smaller but more homogeneous sample of the rater populations in terms of their gender, age, education, and rating experience. This appeared to be a clear contrast to Esfandiari (2019), where raters varied significantly in age, gender, education, which may have contributed to the result of no significant findings between personality traits and rating severity.

That said, the sample size of this study could still be considered a limitation that deserves attention in future investigations and more rigorous control of raters (especially for demographic variables unique in the Chinese context, e.g., raters having siblings or not) is always desirable.

Another potential limitation concerns the generalizability of the findings of the present study. Despite the finding that extroversion is the only valid and significant predictor of rater severity in Chinese L2 writing assessment, it is still unclear whether such a finding could be safely generalized to other rater populations and/or in other L2 writing assessment contexts.

### Appendix A

**Writing Tasks of the Present Study**

Task 1 of Test 1 (KW1)

**Topic:** Endeavor and success

Directions: Please write an essay around 80 characters, and make sure to include all the five keywords given below.

人生, 理想, 坚强, 赞成, 成功 (Translation: life; ideality; perseverance; laudation; success)

Task 2 of Test 1 (PIC1)

**Topic:** Urban traffic

Directions: Please write an essay around 80 characters based on the picture given below.

---

Table 8. Severity Regressed on Domain Level Personality, Excluding the Outliers.

| Model | \(B\) | SE | \(\beta\) | \(t\) | Sig. | Effect size\(^b\) \((d, r)\) |
|-------|------|----|--------|-----|-----|----------------|
| (Constant) | 0.487 | .108 | | 4.527 | .000 | |
| Neuroticism | 0.036 | | | 0.197 | .845 | |
| Extroversion | 0.007 | .003 | .495 | 2.792 | .010 | (.77, .367) |
| Openness | −2.11 | | | −1.199 | .243 | |
| Agreeableness | −0.066 | | | −0.296 | .770 | |
| Conscientiousness | −0.165 | | | −0.928 | .363 | |

\(^a_n = 26\) (excluding observation #14 and #28), adjusted \(R^2 = .214\), standard error of estimate = .116.  
\(^b\)Effect sizes in this table were estimated through Ellis (2009).
### Table B1. The Rubric for Essay Writing Test of HSK-5.

| Band | General impression | Grammar | Vocabulary | Note |
|------|--------------------|---------|------------|------|
| 5    | Demonstrates clear competence in writing though it may have occasional errors. An essay in this category a. has substantial content, complete structure, logical development, clear organization, smooth coherence, and fits appropriate register b. uses native expressions and has high readability c. accomplishes the requirements of the writing task, using all the keywords provided or make content matching the given picture d. has no less than 60 Chinese characters | a. fluent sentences, abundant forms of expression, proper uses of sophisticated sentence structures b. accurate grammar, no obvious mistakes of grammar, especially for word order and/or functional categories | a. demonstrates syntactic variety. Accurate, and appropriate uses of vocabulary. May contain minor mistakes that have no influence on smooth communication b. may contain no more than two mis-spelled characters | Allow misuses of one out of the five keywords provided in the writing task |
| 4    | Demonstrates competence in writing. An essay in this category a. has roughly substantial content, adequately complete structure, a sense of logic, clear organization, generally smooth coherence, minor problems in sentence connection b. is easy to read, and may contain some errors that occasionally obscure meaning c. accomplishes the requirements of the writing task, using all the keywords provided or make content matching the given picture d. has no less than 60 Chinese characters | a. fluent sentences, demonstrate form variety, may contain errors in sophisticated structures, but do not obscure meaning b. may contain sparse grammar errors, mainly on word order or functional categories, and do not affect accuracy of expression | a. adequate vocabulary, generally accurate uses of vocabulary. Generally appropriate word choices and proper register. May contain minor mistakes that do not affect effective communication b. may contain several mis-spelled characters | |
| 3    | Demonstrates minimal competence in writing. An essay in this category a. has generally complete content and structure b. is adequately developed, organized, and coherent c. uses effective expression in general. May contain errors that obscure meaning, but still readable in general d. accomplishes the requirements of the writing task, using all the keywords provided or make content matching the given picture e. has no less than 60 Chinese characters | a. basic fluence, simple expressions b. may contain certain grammar mistakes, mainly on word order and/or functional categories | a. demonstrates a mastery of certain amount of vocabulary, can use simple, high frequency words to satisfy task requirements. May contain inappropriate uses of words b. may contain a small amount of mis-spelled characters | Allow misuses of no more than two out of the five keywords provided in the writing task |
### Table B1. (continued)

| Band | General Impression | Grammar | Vocabulary | Note |
|------|--------------------|---------|------------|------|
| 2    | Demonstrates some developing competence in writing. An essay in this category  
|      | a. addresses the topic inadequately  
|      | b. has inadequate organization and weak coherent. Main ideas are basically understandable, but from time to time readers need to comprehend by guessing  
|      | c. uses only three or four out of the five keywords provided or partly mismatch the given picture  
|      | d. has 40 to 60 characters | a. simple sentences, lack fluency  
|      | b. many grammar mistakes which obscure meaning and understanding. Inappropriate or misuses of simple conjunctions | a. inappropriate uses of some of the five keywords provided in the writing task  
|      | b. uses limited vocabulary. Misuses of words causing obscured meaning | |
| 1    | Demonstrates incompetence in writing. An essay in this category reveals one or more of the following weakness  
|      | a. serious disorganization or underdevelopment, and lack of necessary coherence  
|      | b. hard to understand  
|      | c. Uses less than three out of the five keywords provided or basically mismatch the given picture  
|      | d. Has less than 40 characters | serious grammar mistakes all over the text, causing failure in communication  
|      | a. noticeable amount of inappropriate word choices, misuses of words | a. many mis-spelled characters |
| 0    | Demonstrates severe incompetence in writing. An essay in this category may exhibit one or more of the following problems  
|      | a. using none of the five keywords provided or is irrelevant to the given picture  
|      | b. blank script  
|      | c. having no more than one complete sentence or no message conveyed, so that there is no way to judge writing competence  
|      | d. copying the writing tasks only  
|      | e. duplicating essays of others | |

| Score | Grammar or vocabulary |
|-------|----------------------|
| 3     | Better than the grammar/vocabulary description of a particular level, but does not reach an upper level. |
| 2     | Fits exactly the description of a particular level. |
| 1     | Inferior to the description of a particular level, but still better than the description of a lower level. |
## Table C1. Conversion Table for HSK-5 Essay Writing Score.

| Band score | Grammar score | Vocabulary score | Final score |
|------------|---------------|------------------|-------------|
| 5          | 2             | 2                | 21          |
| 2          | 1             | 2                | 20          |
| 1          | 2             | 1                | 19          |
| 4          | 3             | 3                | 18          |
| 3          | 2             | 2                | 17          |
| 2          | 3             | 1                | 16          |
| 1          | 3             | 3                | 15          |
| 2          | 2             | 1                | 14          |
| 3          | 3             | 3                | 13          |
| 2          | 3             | 2                | 12          |
| 1          | 3             | 1                | 11          |
| 2          | 2             | 2                | 10          |
| 1          | 2             | 1                | 9           |
| 2          | 3             | 3                | 8           |
| 1          | 3             | 2                | 7           |
| 2          | 3             | 1                | 6           |
| 1          | 3             | 3                | 5           |
| 2          | 2             | 2                | 4           |
| 1          | 2             | 1                | 3           |
| 2          | 2             | 3                | 2           |
| 1          | 2             | 2                | 1           |
| 0          | 0             | 0                | 0           |

## Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Fundamental Research Funds for the Central Universities in China (Grant No. 20720181002).

## Research Ethics
All participants were informed of the nature of this research and gave their verbal consent to be involved.

## References

Alaei, M. M., Ahmadi, M., & Zadeh, N. S. (2014). The impact of rater’s personality traits on holistic and analytic scores: Does genre make any difference too? *Procedia - Social and Behavioral Sciences*, 98, 1240–1248.

Aluja, A., Garcia, O., Rossier, J., & Garcia, L. F. (2005). Comparison of the NEO-FFI, the NEO-FFI-R and an alternative short version of the NEO-PI-R (NEO-60) in Swiss and Spanish samples. *Personality and Individual Differences*, 38(3), 591–604.

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly*, 13, 1–24.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12, 238–257.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44, 31–57.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *The Journal of Applied Psychology*, 85(2), 232–236.

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment*, 17(3), 300–310.

Bracht, G. H. (1967). The comparative values of objective and essay testing in undergraduate education: Implications for valid assessment of instruction [Unpublished doctoral dissertation]. University of Colorado.

Carrell, P. L. (1995). The effect of writers’ personalities and raters’ personalities on the holistic evaluation of writing. *Assessing Writing*, 2(2), 153–190.

Chapman, B. P., Duberstein, P. R., Sörensen, S., & Lyness, J. M. (2007). Gender differences in Five Factor Model personality traits in an elderly cohort. *Personality and Individual Differences*, 43(6), 1594–1603.

Choi, Y. H., & Lee, S. (2019). The role of raters’ personality traits in the rating process of L2 argumentative writing: Interactions with argument positions. *Journal of Research in Curriculum & Instruction*, 23(2), 69–82.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Laurence Erlbaum.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163–178.

Costa, P. T. (1996). Work and personality: Use of the NEO-PI-R in industrial/organizational psychology. *Applied Psychology: An International Review*, 45(3), 225–241.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7, 31–51.

## ORCID iDs

Yu Zhu [https://orcid.org/0000-0001-7539-2465](https://orcid.org/0000-0001-7539-2465)

Liuyan Yang [https://orcid.org/0000-0001-9666-8790](https://orcid.org/0000-0001-9666-8790)
Dai, X., Yao, S., Cai, T., & Yang, J. (2004). Reliability and validity of the NEO-PI R in mainland China. *Chinese Mental Health Journal, 18*(2), 171–175.

Dallal, G. (2012). *The regression effect*. http://www.jerrydallal.com/LHSP/regeff.htm

DeCostanza, A. H., & colleagues. (2017). Using principal component scores to enhance the validity and reliability of big five personality measures. *Journal of Individual Differences, 38*(2), 83–93.

Dewberry, C., Davies-Muir, A., & Newell, S. (2013). Impact and causes of rater severity/leniency in appraisals without postevaluation communication between raters and rates. *International Journal of Selection and Assessment, 21*(3), 286–293.

Ellis, P. D. (2009). *Effect size calculators*. http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

Engelhard, G. (1992). The measurement of writing ability with a Many-Faceted Rasch Model. *Applied Measurement in Education, 5*(3), 171–191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement, 31*(2), 93–112.

Engelhard, G., Jr., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling, 60*(1), 33–52.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (TOEFL Research Report RR-03-17). Educational Testing Service.

Esfandiari, R. (2019). How predictable ratings are: The role of personality traits. *Journal of Modern Research in English Language Studies, 6*(3), 33–55.

Furnham, A., Moutafi, J., & Crump, J. (2003). The relationship between the Revised Neo-Personality Inventory and the Myers-Briggs Type Indicator. *Social Behavior and Personality, 31*(6), 577–584.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. College Entrance Examination Board.

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*(3), 225–246.

Harari, M. B., Rudolph, C. W., & Laginess, A. J. (2015). Does rater personality matter? A meta-analysis of Big Five-performance rating relationships. *Journal of Occupational and Organizational Psychology, 88*(2), 387–414.

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses in education*. Routledge.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Allyn and Bacon.

Iramaneerat, C., & Yudkowski, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation & the Health Professions, 30*(3), 266–283.

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38*, 1039–1051.

Kim, Y.-H. (2009). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics, 30*(3), 435–440.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26–43.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

Linacre, J. M. (1989). *Objectivity for judge-intermediated certification examinations* [Paper presentation]. Annual Meeting of the American Educational Research Association, San Francisco, CA.

Linacre, J. M. (2012). *A user’s guide to FACETS, Rasch-model computer programs*. www.winsteps.com

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*(4), 425–444.

McBride, C., Bacchiochi, J. R., & Bagby, R. M. (2005). Gender differences in the manifestation of sociotropy and autonomy personality traits. *Personality and Individual Differences, 38*, 129–136.

McNeal, K. (2019). *The effects of rater performance and perspective on rating leniency* [Doctoral dissertation]. University of Minnesota.

Myers, L. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

O’Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169–192.

O’Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*, 373–386.

Pittenger, D. J. (1993). The utility of the Myers-Briggs Type Indicator. *Review of Educational Research, 63*(4), 467–488.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493.

Schoonen, R., Verger, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*, 157–184.

Song, C. B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163–182.

Sunderland, J. (1995). *Gender and language testing*. *Language Testing, 17*, 24–35.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal, 74*, 49–55.

Wang, J., & Engelhard, G., Jr. (2017). Using a multifocal lens model and Rasch measurement theory to evaluate rating quality in writing assessments. *Pensamiento Educativo. Latin-American Educational Research Journal, 2*, 1–16.

Wang, J., Engelhard, G., Jr., Racynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36–47.

Weigle, S. C. (1994). *Effects of training on raters of ESL compositions: Quantitative and qualitative approaches* [Unpublished doctoral dissertation]. University of California.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*, 145–178.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. *Objective Measurement: Theory into Practice, 5*, 113–133.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*, 256–280.

Yang, J., McCrae, R. R., & Costa, P. T., Jr. (1999). Cross-cultural personality assessment in psychiatric populations: The NEO-PI-R in the People’s Republic of China. *Psychological Assessment, 11*(3), 359–368.

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97–107.

Zhang, J. (2016). Same text different processing? Exploring how raters’ cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37–53.