Book Section:
Zhao, H. orcid.org/0000-0002-9502-4763 and Zhao, B. (2023) Exploring the Consistency Between Self- and Teacher Assessment: Using Co-Constructed Assessment Descriptors in EAP Writing in China. In: Chong, S. and Reinders, H., (eds.) Innovation in Learning-Oriented Language Assessment. New Language Learning and Teaching Environments . Palgrave MacMillan , Cham, Switzerland , pp. 81-103. ISBN 978-3-031-18949-4

https://doi.org/10.1007/978-3-031-18950-0_6

Exploring the consistency between self- and teacher assessment: using co-constructed assessment descriptors in EAP writing in China

Tell me and I forget, teach me and I may remember, involve me and I learn.

-Benjamin Franklin

Outline

The current study adopted multiple quantitative methods to investigate the consistency between self-and teacher assessment in a tertiary institution in China. Different from other studies that use the assessment criteria provided by instructors, this study utilised the tutor-student co-constructed criteria for self-and teacher assessment, adapted from the Common European Framework of Reference for Languages (CEFR). It employed three quantitative methods to counteract the limitations of each method and revealed the congruence and disparity of teacher and self-assessment ratings on the same assignments from different perspectives. It provided implications for using self-assessment in English for Academic Purposes (EAP) writing alongside teacher assessment.

Introduction

The role of self-assessment in language learning has been proliferated in recent years to develop learner agency. However, the development of learner agency depends on its embedded social contexts that can afford (Larsen–Freeman, 2019). In a teacher-driven learning context, learners have been heavily relying on teachers to provide feedback and doubted their own capacity of assessing their learning (Chang, 2016; Zhao, 2010). Likewise, teachers have been accustomed to teacher assessment and doubt self-assessment (Zhao, 2018a). In a teacher-driven learning context, a comparison study between self-and teacher assessment is essential to provide learners and teachers with empirical evidence of how reliable self-assessment is, where differences might lie and how to enhance self-assessment for learning.

Consistency of self- and teacher assessment

Different findings were reported across instructional settings when self and teacher assessment results were compared. Sung et al. (2005) reported that teacher assessment provided a higher average mark than self-assessment while Chang et al. (2012) observed in their study that students provided much higher ratings than teachers. Overall, the meta-analysis of studies in higher education by Falchikov and Boud (1989) shows an average coefficient of .39 between self and teacher marks and an average agreement percentage of 64% between self- and teacher assessors.

The inconsistent findings across studies were explained by Falchikov and Boud (1989) and Falchikov and Goldfinch (2000). They suggested that the congruence of teacher and student ratings could be affected by assessment focuses and assessment criteria: analytic judgement is associated with a lower agreement than holistic judgement and well-understood criteria; however, a higher level of familiarity with and ownership of criteria are associated with better agreements. Inconsistency between self- and teacher assessment could vary from the focuses of assessment. Chang et al. (2012) found that significant differences in self- and teacher grades existed in three out of seven assessment aspects. Bouzidi and Jaillet (2009) observed that the explicitness of assessment aspects and assessment instructions could decide the agreements. Training has been commonly believed to increase the congruence of teacher and student-led assessment (Rahimi, 2013; Zhao, 2014). The classroom settings (i.e., the instructional context relating to the curriculum and design of teaching and learning activities) could affect learning including assessment (Dörnyei, 2009).

Research contexts

This study was conducted during the English testing reform in China launched in 2014, aiming to develop China's Standard of English Language Ability (hereafter as CSE) that bridges teaching, learning, assessment and learner autonomy (Ministry of Education of the People's Republic of China, 2018). To fulfil this objective, this study integrated self-assessment into the existing teacher assessment with an adapted version of the European Language Portfolio (ELP), derived from CEFR and bearing resemblance with the action-oriented CSE (i.e. can-do statements) (Jiang, 2016). The project aimed to enrich the limited guidance on using action-oriented assessment descriptors in assessment and examined how the exam-driven learning culture in the Chinese would influence the implementation of self-assessment to promote coherence among teaching, learning and assessment (Zhao, 2018a).

Research context: teacher-driven writing instruction

Before the introduction of self-assessment, writing tutors played a dominant role in assessment. EAP writing in this research context was heavily based on a product-oriented writing approach. A typical lesson started with a teacher-led analysis of an exemplar article in terms of its structure and language use. The assessment was conducted solely by the writing tutors. Due to the large class size (over 50), little formative feedback was provided to justify the marks and explain the strengths and weaknesses of student writing. Introducing self-assessment into writing instruction was expected to develop learners' autonomy of assessing their writing and the checklist of 'I can' statements at three scales [i.e. achieved (☺), almost there (😐**)** and not there yet (☹)] was designed to use the minimum class time to maximise the value of self-assessment for improving writing quality and proficiency.

Research design

This study was carried out in three phases. A pre-assessment phase addressed learners' concerns over self-assessment and introduced the CEFR and ELP descriptors, followed by training in self-assessment with teachers' demonstration of how to use the assessment criteria. In the assessment phase, students and teachers conducted self-assessment (inside classrooms) and teacher assessment (outside of classrooms) of the same piece of writing, using the co-constructed assessment criteria by teachers and students (Zhao & Zhao, 2020). In the post-assessment phase, the participants' experiences of self-assessment were investigated. This paper focused on the data from the assessment phase and answered the following two research questions:

1. What was inter-rater agreement between self- and teacher ratings, using co-constructed assessment criteria?

2. What were the differences between self- and teacher ratings, using co-constructed assessment criteria?

The quantitative results will be discussed with potential factors in the discussion section.

Participants

Two tutors and 146 students from four classes and three subjects participated in the project for one semester voluntarily. All participants were Chinese who spoke English as a foreign language (EFL). Both tutors had been working at the institution for more than ten years and teaching the EAP module since 2016.

The student participants were second-year university students, majoring in Network Media (Class 1-2), Public Management (Class 3) and Chinese Linguistics and Literature (Class 4). Most of the students had been learning English for more than 10 years since their primary school, with an approximate English proficiency level around B1-B2, based on their entrance English exam scores, writing scores and tutors' judgement. They had limited self-assessment experience. Table 6.1 summarizes the participants' backgrounds.

Table 6.1 Student participants' backgrounds

Class ID	**Number of students**	**Gender**	**Major**	**Final writing marks (average)**
1	35 (taught by Tutor 1)	Male: 16 Female: 19	Network Media	69.69 (SD=7.66)
2	35 (taught by Tutor 1)	Male: 13 Female: 22	Network Media	67.79 (SD=7.17)
3	29 (taught by Tutor 2)	Male: 8 Female: 21	Public Management	71.31 (SD=7.79)
4	47 (taught by Tutor 2)	Male:3 Female: 44	Chinese Language and Literature	75.98 (SD=6.22)

Descriptive analysis and an ANOVA test of English writing proficiency across classes showed no significant difference in writing proficiency among Class 1-3 but students in Class 4 scored 5 points higher in their average writing score than the other three classes. It is worthy of noticing the unbalanced number of male and female students which could affect the results as existing studies have suggested the gender difference in self-efficacy and use of CEFR in self-assessment (Denies & Janssen, 2016). Due to the limited space, the impact of participants' backgrounds on the results would not be discussed in this paper.

Data collection and analysis

Data were collected from two genres: summaries of reading about *society today* and *food security* and argumentative essays on *sustainable energy and sustainable fashion*. Thirty minutes were allocated for students for self-assessment. They were asked to tick one of the three options for each descriptor in the co-constructed assessment grids. Teachers used the same criteria to assess the same essays without reading the student self-ratings to avoid their possible influence on teacher ratings.

Comparative analysis was conducted between self- and teacher ratings for each task. Kappa agreement for nominal data (K) was used to assess the agreement between teacher and self-assessment on a scale of slight (0.0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.00) (Frey, 2018). Difference analysis was carried out to reveal whether students over- or underrated themselves, compared to their teachers via the Wilcoxon signed-rank test (WST) which suggested the specific number of the same self- and teacher ratings and different ratings with higher self-ratings and lower self-ratings, respectively.

Findings

The results were reported in the order of assessing constructing summaries, language use in summaries, constructing argumentative essays and language use in argumentative essays.

Self- and teacher ratings on constructing summaries

Table 6.2 showed that teachers and students achieved significant agreements in six out of the nine descriptors (p<.05). A mean of 0.22 Kappa value showed a fair agreement in the six descriptors among the 134 sets of self- and teacher assessment data. The range of Kappa values (.146 - .389) suggested a slight to fair yet different agreement across the descriptors, indicating the need of examining the difference between self- and teacher ratings.

Table 6.2 Kappa inter-rater agreement between teacher and self-ratings for constructing summaries

Descriptors	Kappa Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
D1: give a simple summary	-.008	.063	-.148	.883
D2: paraphrase	**.146**	.076	2.069	**.039***
D3: reproduce language from the reading text	**.241**	.072	3.564	**.000***
D4: select the information for summaries	**.208**	.068	3.710	**.000***
D5: write summaries independently	**.389**	.073	5.432	**.000***
D6: summarise the plots	.039	.063	.650	.516
D7: summarise the main themes	**.166**	.062	3.343	**.001***
D8: make notes	**.159**	.054	3.875	**.000***
D9: summarise the background	-.001	.051	-.027	.978
N of Valid Cases	134			
a. Not assuming the null hypothesis.				
b. Using the asymptotic standard error assuming the null hypothesis.				

The Wilcoxon signed-rank test (WST) showed significant differences between self- and teacher ratings for seven out of the nine descriptors (p<.00).

Table 6.3 WST: differences between teacher and self-ratings on constructing summaries

	TAS1 D1 – SAS1 D1	TAS1 D2 – SAS1 D02	TAS1 D3 – SAS1 D3	TAS1 D4 – SAS1 D4	TAS1 D5 - SAS1 D5	TAS1 D6 - SAS1 D6	TAS1 D7 - SAS1 D7	TAS1 D8 - SAS1 D8	TAS1 D9 - SAS1 D9
Z	-5.253	-.692	-1.838	-3.095	-3.130	-4.866	-5.416	-5.336	-3.252
Sig. (2-tailed)	.000*	.489	.066	.002*	.002*	.000*	.000*	.000*	.001*

*Statistically significant differences

Note: TAS1 = teacher assessment in constructing summaries; SAS1 = self-assessment in constructing summaries.

Table 6.4 showed more assignments received higher ratings from teachers than the students themselves for all but Descriptor 6. However, on average, the number of ties suggested that 51% of the assignments received the same ratings from students themselves and their tutors. This applies to six out of the nine assessment descriptors, suggesting half of the students shared the same understanding of their proficiency in these assessment aspects with their tutors.

Table 6.4 WST: congruence between teacher and self-ratings for constructing summaries

		N	Mean Rank	Sum of Ranks
TAS1D1 – SAS1D1	Negative Ranks[a]	17	40.00	680.00
	Positive Ranks[b]	64	41.27	2641.00
	Ties[c]	53		
	Total	134		
TAS1D2 – SAS1D2	Negative Ranks	29	33.74	978.50
	Positive Ranks	36	32.40	1166.50
	Ties	**69**		
	Total	134		
TAS1D3 – SAS1D3	Negative Ranks	22	29.50	649.00
	Positive Ranks	36	29.50	1062.00
	Ties	**76**		
	Total	134		
TAS1D4 – SAS1D4	Negative Ranks	16	28.75	460.00
	Positive Ranks	40	28.40	1136.00
	Ties	**78**		
	Total	134		
TAS1D5 – SAS1D5	Negative Ranks	11	21.00	231.00
	Positive Ranks	31	21.68	672.00
	Ties	**92**		
	Total	134		
TAS1D6 – SAS1D6	Negative Ranks	61	40.83	2490.50
	Positive Ranks	18	37.19	669.50
	Ties	55		
	Total	134		
TAS1D7 – SAS1D7	Negative Ranks	8	28.00	224.00
	Positive Ranks	49	29.16	1429.00
	Ties	**75**		
	Total	132		
TAS1D8 – SAS1D8	Negative Ranks	11	32.50	357.50
	Positive Ranks	54	33.10	1787.50
	Ties	**69**		
	Total	134		
TAS1D9 – SAS1D9	Negative Ranks	32	42.88	1372.00
	Positive Ranks	60	48.43	2906.00
	Ties	42		
	Total	134		
a. negative ranks: teacher assessment ratings are lower than self-assessment ratings				
b. positive ranks: teacher assessment ratings are higher than self-assessment ratings.				
c. ties: teacher assessment ratings equalled to self-assessment ratings.				

Self- and teacher ratings on the language use of summaries

Table 6.5 showed that significant agreements existed in eight out of the twelve descriptors (p<.05). However, a mean of 0.19 Kappa values suggested slight to fair agreements across descriptors, slightly lower than ratings in constructing summaries. The different k values with a range of .142 and .286 indicated the variance across descriptors.

Table 6.5 Kappa inter-rater agreement between teacher and self-ratings for the language use of summaries

Descriptors	Number of valid case	Kappa Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
D1: control of vocabulary use	131	.099	.075	1.321	.186
D2: grammatical accuracy	130	.255	.067	4.498	.000*
D3: punctuation accuracy	131	.161	.056	2.759	.006*
D4: spelling accuracy	129	.052	.081	.652	.514
D5: sentence structure	131	.098	.048	2.108	.044*
D6: tenses	130	.004	.069	.065	.948
D7: use of linking words	131	.130	.063	2.282	.022*
D8: linging discrete points	123	.173	.067	3.197	.001*
D9: use of connectors	132	-.028	.065	-.437	.662
D10: clarity	130	.142	.056	2.58	.010*
D11: qualify opinions	131	.248	.066	4.062	.000*
D12: convey information	131	.286	.068	4.389	.000*
a. Not assuming the null hypothesis.					
b. Using the asymptotic standard error assuming the null hypothesis.					
c. * means significant Kappa Value					

WST revealed the significant differences existing in ten out of the twelve assessment descriptors (p<.05) (Table 6.6).

Table 6.6 WST: differences between teacher- and self-ratings on the language use of summaries

	TAS2D1 – SAS2D1	TAS2D2 – SAS2D02	TAS2D3 – SAS2D3	TAS2D4 – SAS2D4	TAS2D5 – SAS2D5	TAS2D6 – SAS2D6	TAS2D7 – SAS2D7	TAS2D8 – SAS2D8	TAS2D9 – SAS2D99	TAS2D10 – SAS2D10	TAS2D11 - SAS2D11	TAS2D12 – SAS2D12
Z	-1.156^b	-2.248^c	-5.724^b	-.938^c	-6.575^b	-2.281^b	-2.609^b	-1.983^b	-5.063^b	-4.400^b	-5.632^b	-3.459^b
Sig. (2-tailed)	.248	**.025***	**.000***	.348	**.000***	**.023***	**.009***	**.047***	**.000***	**.000***	**.000***	**.001***

*Statistically significant differences

Note: TAS2 = teacher assessment of language use in summaries; SAS2 = self-assessment of language use in summaries.

Descriptive statistics in Table 6.7 further revealed that more assignments received higher teacher than self-ratings on Descriptors 1, 3, 5-12 but lower teacher ratings on Descriptors 2 and 4. In addition, 53% of the assignments received the same rating (i.e. ties) from teachers and students. Nine out of the 12 descriptors in over half of the assignments received the same ratings from students themselves and the tutors.

Table 6.7 WST: differences between teacher- and self-ratings on the language use of summaries

		N	Mean Rank	Sum of Ranks
TAS2D1 – SAS2D1	Negative Ranks[a]	26	31.88	829.00
	Positive Ranks[b]	36	31.22	1124.00
	Ties[c]	69		
	Total	131		
TAS2D2 – SAS2D02	Negative Ranks	34	26.79	911.00
	Positive Ranks	18	25.94	467.00
	Ties	78		
	Total	130		
TAS2D3 – SAS2D3	Negative Ranks	7	33.86	237.00
	Positive Ranks	54	30.63	1654.00
	Ties	70		
	Total	131		
TAS2D4 – SAS2D4	Negative Ranks	31	29.11	902.50
	Positive Ranks	25	27.74	693.50
	Ties	73		
	Total	129		
TAS2D5 – SAS2D5	Negative Ranks	10	35.50	355.00
	Positive Ranks	69	40.65	2805.00
	Ties	52		
	Total	131		
TAS2D6 – SAS2D6	Negative Ranks	26	34.35	893.00
	Positive Ranks	44	36.18	1592.00
	Ties	60		
	Total	130		
TAS2D7 – SAS2D7	Negative Ranks	17	28.74	488.50
	Positive Ranks	38	27.67	1051.50
	Ties	76		
	Total	131		
TAS2D8 – SAS2D8	Negative Ranks	17	24.41	415.00
	Positive Ranks	31	24.55	761.00
	Ties	75		
	Total	123		
TAS2D9 – SAS2D9	Negative Ranks	16	38.00	608.00
	Positive Ranks	60	38.63	2318.00
	Ties	56		
	Total	132		
TAS2D10 – SAS2D10	Negative Ranks	15	28.00	420.00
	Positive Ranks	48	33.25	1596.00
	Ties	67		
	Total	130		
TAS2D11 - SAS2D11	Negative Ranks	6	24.50	147.00
	Positive Ranks	48	27.88	1338.00
	Ties	77		
	Total	131		
TAS2D12 – SAS2D12	Negative Ranks	14	25.00	350.00
	Positive Ranks	38	27.05	1028.00
	Ties	79		
	Total	131		
a. negative ranks: teacher assessment ratings are lower than self-assessment ratings				
b. positive ranks: teacher assessment ratings are higher than self-assessment ratings.				
c. ties: teacher assessment ratings equalled to self-assessment ratings.				

Self- and teacher ratings on constructing argumentative essays

Table 6.8 shows that students and tutors obtained significant agreements on five out of the seven descriptors (p< .05). A mean of 0.23 Kappa values for the five descriptors suggested a slight to fair agreement across descriptors in the 142 assignments.

Table 6.8 Kappa inter-rater agreement between teacher and self-ratings for constructing argumentative essays

Descriptors	Measure of Agreement Kappa value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
D1: using information from the text to support arguments	.366	.073	4.657	.000*
D2: form a line of arguments	.143	.075	1.885	.059
D3: support arguments with supporting details	.050	.060	.849	.396
D4: develop lines of arguments	.212	.068	3.664	.000*
D5: structure different lines of arguments	.246	.075	4.759	.000*
D6: synthesising information	.162	.066	2.953	.003*
D7: reconstruct arguments coherently	.162	.067	3.327	.001*
N of Valid Cases	142			
a. Not assuming the null hypothesis.				
b. Using the asymptotic standard error assuming the null hypothesis.				

Table 6.9 shows significant differences in five out of the seven descriptors (p<.05). Table 6.10 further showed that more assignments received higher teacher than self-ratings for all descriptors except Descriptor 2. In addition, 59% received the same ratings (i.e. ties) from teachers and students for all the seven descriptors.

Table 6.9 WST: Differences between teacher and self-ratings for constructing argumentative essays

	TAA1D1 – SAA1D1	TAA1D2 – SAA1D2	TAA1D3 – SAA1D3	TAA1D4 – SAA1D4	TAA1D5 – SAA1D5	TAA1D6 – SAA1D6	TAA1D7 – SAA1D7
Z	-2.654	-1.054	-4.336	-4.185	-4.523	-5.333	-3.052
Asymp. Sig. (2-tailed)	.008*	.292	.000*	.348	.000*	.000*	.002*

*Statistically significant differences

Note: TAA1 = teacher assessment of constructing argumentative essays; SAA1 = self-assessment of constructing argumentative essays

Table 6.10 WST: Differences between teacher and self-ratings for constructing argumentative essays

		N	Mean Rank	Sum of Ranks
TAA2D1 – SAA2D1	Negative Ranks[a]	14	23.50	329.00
	Positive Ranks[b]	32	23.50	752.00
	Ties[c]	96		
	Total	142		
TAA2D2 – SAA2D02	Negative Ranks	36	32.78	1180.00
	Positive Ranks	28	32.14	900.00
	Ties	78		
	Total	142		
TAA2D3 – SAA2D3	Negative Ranks	24	36.75	882.00
	Positive Ranks	60	44.80	2688.00
	Ties	58		
	Total	142		
TAA2D4 – SAA2D4	Negative Ranks	15	29.50	442.50
	Positive Ranks	47	32.14	1510.50
	Ties	80		
	Total	142		
TAA2D5 – SAA2D5	Negative Ranks	7	22.50	157.50
	Positive Ranks	37	22.50	832.50
	Ties	98		
	Total	142		
TAA2D6 – SAA2D6	Negative Ranks	9	28.00	252.00
	Positive Ranks	50	30.36	1518.00
	Ties	83		
	Total	142		
TAA2D7 – SAA2D7	Negative Ranks	13	22.50	292.50
	Positive Ranks	33	23.89	788.50
	Ties	96		
	Total	142		
a. negative ranks: teacher assessment ratings are lower than self-assessment ratings				
b. positive ranks: teacher assessment ratings are higher than self-assessment ratings.				
c. ties: teacher assessment ratings equalled to self-assessment ratings.				

Self- and teacher ratings on the language use of argumentative essays

Table 6.11 showed significant agreements in eight out of the 14 descriptors (p< .05). However, a mean of 0.18 Kappa value for the eight descriptors suggested slight agreements between self- and teacher ratings.

Table 6.7 Kappa inter-rater agreement between teacher and self-ratings for the language use of argumentative essays

Descriptors	N of valid case	Measure of Agreement: Kappa value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
D1	140	.194	.049	3.643	.000*
D2	140	.137	.067	2.300	.021*
D3	140	.129	.073	1.868	.062
D4	138	.127	.070	1.834	.067
D5	140	.075	.059	1.291	.197
D6	140	.133	.052	2.648	.008*
D7	140	.141	.067	2.123	.034*
D8	139	.084	.061	1.517	.129
D9	139	.198	.074	3.448	.001*
D10	140	.247	.057	4.506	.000*
D11	140	.119	.068	1.880	.060
D12	140	.055	.053	1.174	.240
D13	138	.109	.055	2.398	.017*
D14	140	.245	.067	3.862	.000*
a. Not assuming the null hypothesis.					
b. Using the asymptotic standard error assuming the null hypothesis.					
*statistically significant differences					

Table 6.12 showed a significant difference in twelve of the fourteen descriptors (p< .05).

Table 6.8 WST: differences between teacher- and self-ratings for the language use of argumentative essays

	TAA2D1 - SAA2D1	TAA2D2 - SAA2D2	TAA2D3 - SAA2D3	TAA2D4 - SAA2D4	TAA2D5 - SAA2D5	TAA2D6 - SAA2D6	TAA2D7 - SAA2D7	TAA2D8 - SAA2D8	TAA2D9 - SAA2D9	TAA2D10 - SAA2D10	TAA2D11 - SAA2D11	TAA2D12 - SAA2D12	TAA2D13 - SAA2D13	TAA2D14 - SAA2D14
Z	-7.437[b]	-4.418[b]	-1.485[b]	-4.299[c]	-5.894[b]	-6.745[b]	-1.641[b]	-3.389[b]	-3.960[b]	-5.728[b]	-2.693[b]	-6.099[b]	-5.968[b]	-3.501[b]
Sig. (2-tailed)	**.000***	**.000***	.138	**.000***	**.000***	**.000***	.101	**.001***	**.000***	**.000***	**.007***	**.000***	**.000***	**.000***

*Statistically significant differences

Note: TAA2= teacher assessment of the language use of argumentative essays; SAA2 = self-assessment of the language use of argumentative essays

Table 6.13 further suggested that more assignments received higher teacher than self-ratings for thirteen descriptors. In addition, 52% of the assignments received the same teacher and self-ratings on twelve descriptors (i.e., ties).

Table 6.9 WST: congruence between teacher and self-ratings for the language use of argumentative essays

		N	Mean Rank	Sum of Ranks
TAA2D01 - SAA2D1	Negative Ranks[a]	3	33.50	100.50
	Positive Ranks[b]	65	34.55	2245.50
	Ties[c]	72		
	Total	140		
TAA2D02 - SAA2D2	Negative Ranks	15	29.50	442.50
	Positive Ranks	49	33.42	1637.50
	Ties	76		
	Total	140		
TAA2D03 - SAA2D3	Negative Ranks	27	32.20	869.50
	Positive Ranks	38	33.57	1275.50
	Ties	75		
	Total	140		
TAA2D04 - SAA2D4	Negative Ranks	52	34.77	1808.00
	Positive Ranks	16	33.63	538.00
	Ties	70		
	Total	138		
TAA2D05 - SAA2D5	Negative Ranks	13	38.00	494.00
	Positive Ranks	65	39.80	2587.00
	Ties	62		
	Total	140		
TAA2D06 - SAA2D6	Negative Ranks	8	29.00	232.00
	Positive Ranks	68	39.62	2694.00
	Ties	64		
	Total	140		
TAA2D07 - SAA2D7	Negative Ranks	27	34.52	932.00
	Positive Ranks	41	34.49	1414.00
	Ties	72		
	Total	140		
TAA2D08 - SAA2D8	Negative Ranks	19	34.03	646.50
	Positive Ranks	48	33.99	1631.50
	Ties	72		
	Total	139		
TAA2D09 - SAA2D9	Negative Ranks	11	25.50	280.50
	Positive Ranks	39	25.50	994.50
	Ties	89		
	Total	139		
TAA2D10 - SAA2D10	Negative Ranks	9	30.50	274.50
	Positive Ranks	55	32.83	1805.50
	Ties	76		
	Total	140		

TAA2D11 - SAA2D11	Negative Ranks	23	33.00	759.00
	Positive Ranks	44	34.52	1519.00
	Ties	73		
	Total	140		
TAA2D12 - SAA2D12	Negative Ranks	10	33.00	330.00
	Positive Ranks	62	37.06	2298.00
	Ties	68		
	Total	140		
TAA2D13 - SAA2D13	Negative Ranks	9	33.00	297.00
	Positive Ranks	58	34.16	1981.00
	Ties	71		
	Total	138		
TAA2D14- SAA2D14	Negative Ranks	16	28.50	456.00
	Positive Ranks	42	29.88	1255.00
	Ties	82		
	Total	140		
a. Teacher assessment Argumentation 2 rating < Self-assessment Argumentation 2 rating				
b. Teacher assessment Argumentation 2 rating > Self-assessment Argumentation 2 rating				
c. Teacher assessment Argumentation 2 rating = Self-assessment Argumentation 2 rating				

Discussions and implications for practice

The statistical analysis between self- and teacher ratings based on Kappa tests and Wilcoxon Signed Ranks Test revealed that students tended to assign either the same or lower ratings compared to their tutors; in addition, the congruence between self- and teacher ratings varied from tasks and assessment descriptors. The findings provide useful implications for utilising self-assessment in EAP writing and instruction.

Firstly, the slight to fair agreements between self- and teacher ratings across the four tasks and assessment descriptors and the significant differences between self- and teacher ratings echoed the existing concern about the reliability of self-ratings (Zhao, 2010, 2018). The incongruence between self and teacher assessment existed in both macro- and micro-aspects of producing summaries and argumentative essays, namely: how to construct them and the language use of both genres.

A question for instructors to ask is whether it is still worthy of integrating self-assessment in their writing instruction. For one thing, over half of the students assigned the same ratings for the majority of descriptors as their tutors. This showed that those students assessed themselves as effectively as their tutors did. Additionally, Falchikov and Boud (1989) stipulate that the success of student-led assessment should be moved beyond the decontextualized degree of agreement with teacher assessment but take account of the learning benefits of self-assessment. The benefits of self-assessment activities using the 'I can do statements' were stipulated by learners and tutors in Zhao and Zhao (2020): fostering metacognition of their knowledge gap in writing, understanding writing beyond the accuracy of language use and increasing learning motivation to fill in the knowledge gap. This is particularly important for language learners to develop their learning autonomy with the assistance of action-oriented assessment criteria. The benefits of self-assessment were echoed by writing tutors who wished to keep involving learners in the assessment process (Zhao & Zhao, 2020). Introducing self-assessment also raises the roles of students in assessment and brings about positive teacher-student relations in the classroom settings (Dörnyei, 2009).

Secondly, the different K values and Z values across assessment descriptors alongside the descriptive statistics generated by the Wilcoxon Signed Ranks Test showed that the congruence of self- and teacher assessment varies from tasks and assessment aspects: e.g. regarding the language use, those requiring a lower level of cognitive knowledge (e.g. grammatical and spelling accuracy) received higher self- than teacher-ratings. This suggests that instructors need to beware of the potentially different roles of self-assessment in different aspects of writing and use it selectively, depending on assessment focuses and learners' writing proficiency of these focuses. Sadler and Good (2006, p. 23) state: "without students awarding exactly the same grades, a teacher is obliged to add some oversight to the process of student-grading". Tutors need to train students in assessing different writing focuses with varied frequencies and strategies, taking into consideration (a) students' familiarity with them, (b) their writing proficiency, (c) their previous and present assessment experience and (d) assessment literacy relating to these aspects and in general.

Last but not least, instructors need to address individual differences when introducing self-assessment, suggested by the number of assignments received higher, lower and the same teacher ratings in comparison with self-ratings. As reported in Zhao and Zhao (2020) with the same group of students, some students reported difficulties in understanding and using assessment descriptors. These students need more support than their peers to develop their confidence and competence in conducting self-assessment. Individual differences in self-assessment also reveal the necessity of individualised assessment methods. Tutors could use self-assessment more frequently with those students who could assess themselves as effectively as them and replace teacher written assessment with other assessment formats: e.g. tutorials focusing on under-achieved aspects in the assessment grids. Students having a low level of agreements with teachers could resort to teacher assessment more often than self-assessment but increase the use of the latter when they are more capable of doing it.

Practice brief

This study has demonstrated the congruence and variance between self- and teacher assessment in terms of constructing summaries and argumentative essays and the language used in them. The results have revealed the importance for writing tutors to beware that self-assessment results differ from focuses of assessment, with a higher level of consistency between teacher and self-assessment in macro- than micro-aspects of writing. This implies that in practice, writing tutors need to provide more explicit explanations of the descriptors of the micro-aspect of writing (i.e. the language use) through either demonstrating how they assess these aspects of writing and/or eliciting and refining students' assessment literacy of self-assessing themselves in terms of those aspects. It would also be ideal to spread different aspects of language use across a few self-assessment sessions to increase the reflection time and thus more accurate self-assessment of these aspects. The different levels of consistency between self- and teacher ratings across assessment aspects and individual assignments also suggest the necessity of selectively using self-assessment for different purposes with different learners, shifting from the current misconception of self-assessment as a one-size-fits-all assessment tool. Equally important to accommodate self-assessment with careful assessment design (e.g. why, when and how to implement self-assessment), it is essential to foster a culture of self-assessment in the classroom settings to develop students' and tutors' affective (e.g. appreciating self-assessment as an effective learning tool) and behavioural (e.g. effectively carrying out self-assessment) systems to maximise the value of self-assessment for writing. This would raise learners' confidence and competence in carrying out self-assessment and thereby raise the congruence with teacher assessment.

References

Bouzidi, L. H., & Jaillet, A. (2009). Can Online Peer Assessment be Trusted? *Journal of Educational Technology & Society*, *12*(4), 257-268. http://www.jstor.org/stable/jeductechsoci.12.4.257

Chang, C.-C., Tseng, K.-H., & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers and education*, *58*(1), 303-320. https://doi.org/10.1016/j.compedu.2011.08.005

Chang, C. Y.-H. (2016). Two decades of research in L2 peer review. *Journal of Writing Research*, *8*(1), 81-117.

Denies, K., & Janssen, R. (2016). Country and Gender Differences in the Functioning of CEFR-Based Can-Do Statements as a Tool for Self-Assessing English Proficiency [Article]. *Language Assessment Quarterly*, *13*(3), 251-276. https://doi.org/10.1080/15434303.2016.1212055

Dörnyei, Z. (2009). Individual Differences: Interplay of Learner Characteristics and Learning Environment. *59*(s1), 230-248. https://doi.org/https://doi.org/10.1111/j.1467-9922.2009.00542.x

Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of educational research*, *59*(4), 395-430. https://doi.org/10.2307/1170205

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, *70*(3), 287-322. https://doi.org/10.2307/1170785

Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications.

Larsen–Freeman, D. (2019). On Language Learner Agency: A Complex Dynamic Systems Theory Perspective. *The Modern language journal (Boulder, Colo.)*, *103*, 61-79. https://doi.org/10.1111/modl.12536

Ministry of Education of the People's Republic of China. (2018). *China's Standards of English Language Ability*.

Rahimi, M. (2013). Is training student reviewers worth its while? A study of how training influences the quality of students' feedback and writing. *17*(1), 67-89. https://doi.org/10.1177/1362168812459151

Sadler, P. M., & Good, E. (2006). The Impact of Self- and Peer-Grading on Student Learning. *Educational assessment*, *11*(1), 1-31. https://doi.org/10.1207/s15326977ea1101_1

Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a web-based self- and peer-assessment system. *Computers and education*, *45*(2), 187-202. https://doi.org/10.1016/j.compedu.2004.07.002

Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing writing*, *15*(1), 3-17. https://doi.org/10.1016/j.asw.2010.01.002

Zhao, H. (2014). Investigating teacher-supported peer assessment for EFL writing. *ELT Journal*, *68*(2), 155-168. https://doi.org/10.1093/elt/cct068

Zhao, H. (2018). Exploring tertiary English as a Foreign Language writing tutors' perceptions of the appropriateness of peer assessment for writing. *Assessment & Evaluation in Higher Education*, 1-13. https://doi.org/10.1080/02602938.2018.1434610

Zhao, H., & Zhao, B. (2020). Co-constructing the assessment criteria for EFL writing by instructors and students: A participative approach to constructively aligning the CEFR, curricula, teaching and learning. *Language teaching research*, 136216882094845. https://doi.org/10.1177/1362168820948458

Word count: 4993