Data Description of the "Chinese EFL Learners' Writing Evaluation by ChatGPT"

The data mainly provide ChatGPT's rating on 82 Chinese EFL learners' writings with scores and comments as well as the scores by reliable manual rating. With the data, researchers can do quantitative or qualitative research on the reliability of EFL writing evaluation with ChatGPT by taking reliable manual ratings as a reference. It includes two parts: 1) ChatGPT's rating with scores and comments, and 2) statistics on overall, average, and specific scores of manual and ChatGPT's rating.

1. EFL Writings with ChatGPT's Rating

There are 270 EFL expository compositions in the *Spoken and Written Corpus of Chinese Learners Version 2.0*. (Wen et al., 2008) written by 270 Chinese EFL learners within a time limit of 30 minutes. Their IDs are from "WEXP0001" to "WEXP0270". The following is the instruction for the writing task.

Expository writing task (100 marks, 30 minutes)

You are going to give a presentation about the development of KFC and MacDonald's over a ten-year period in China. Use the information in the following two graphs and write a report in English (150-180 words) for your presentation. Write your report on the separate answer sheet.

Table 1. Number of stores of KFC and MacDonald's over a ten-year period in China

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
KFC	45	72	131	216	292	327	400	534	902	1000	1200
MacDonald's	6	11	53	122	145	195	214	353	543	573	600

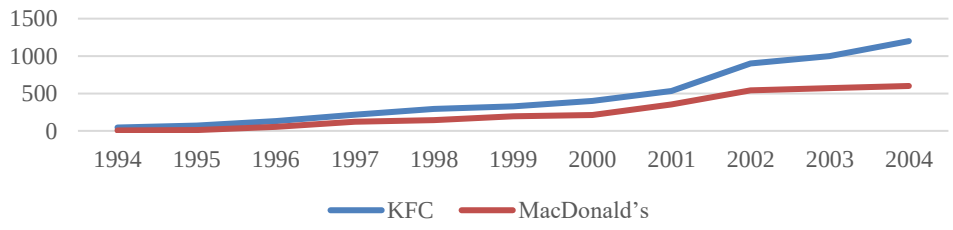

Figure 1. Number of stores of KFC and MacDonald's over a ten-year period in China

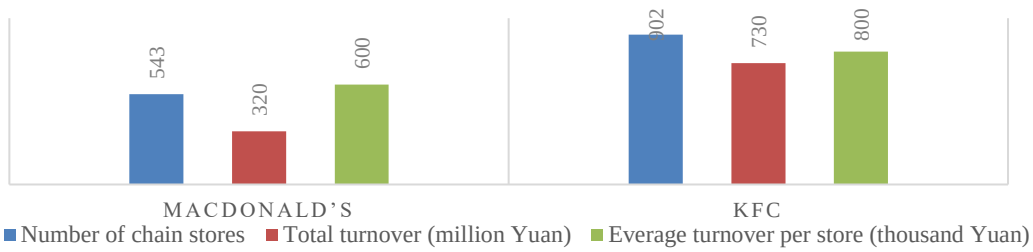

Figure 2. Turnovers of MacDonald's and KFC in 2002

Eighty-two compositions are randomly sampled from the 270 compositions. The sample size is determined by the power analysis software *G*Power* (Faul et al., 2009; Faul et al., 2007). A set of random 82 numbers from 270 are generated by using the *Random Numbers Generator*[1]. The random 82 numbers are as follows:

- `Set 1: 1, 2, 3, 4, 5, 12, 18, 23, 24, 28, 34, 36, 37, 42, 45, 47, 54, 60, 64, 65, 66, 71, 72, 74, 80, 81, 83, 84, 89, 90, 95, 97, 102, 105, 107, 112, 113, 114, 116, 120, 128, 131, 141, 143, 151, 152, 153, 157, 164, 170, 171, 173, 180, 186, 192, 195, 206, 208, 211, 213, 215, 217, 220, 224, 226, 233, 236, 243, 244, 245, 246, 247, 249, 251, 252, 253, 254, 256, 257, 260, 261, 266`

The ChatGPT's rating is generated by inputting the following prompt in the chat box to ask ChatGPT to rate the 82 EFL writings one by one. The next day, the same 82 writings were rated by ChatGPT again with the same prompts to obtain another set of scores.

#WEXP00XX
"..." (EFL writing of the above ID)

The above is a piece of writing by a Chinese university student, who is allowed to write a report of 150-180 words in 30 minutes about the development of KFC and MacDonald's over a ten-year period in China with the reference of the following table. Please rate the writing from aspects of "language" (40 marks), "content" (30 marks), and "organization" (30 marks) and give marks for each aspect and the overall mark.

Table. Number of stores of KFC and MacDonald's over a ten-year period in China

Year	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*
KFC	*45*	*72*	*131*	*216*	*292*	*327*	*400*	*534*	*902*	*1000*	*1200*
MacDonald's	*6*	*11*	*53*	*122*	*145*	*195*	*214*	*353*	*543*	*573*	*600*

2. Scores of Manual and ChatGPT's Rating

The spreadsheet provides not only ChatGPT's rating on the EFL compositions with overall and specific scores but also corresponding scores of manual rating. For the manual rating, the compositions were rated by three experienced raters on aspects of language (40%), content (30%), and organization (30%) and the total score was the sum of the three parts. Then the average scores of the total score and scores of each aspect from the three raters were calculated. When rating the compositions, rubric for rating College English Test Band 6 (CET6) is used as a reference. The following is its rubric.

[1] https://www.random.org/

Language (mark)	Content (mark)	Organization (mark)
There are fragmented language and errors in most sentences, and most of the errors are serious ones. (5/40)	The thinking is totally disordered. (4/30)	The writing is totally disorganized. (4/30)
There are unclear expressions and many serious language errors. (13/40)	It is basically to the point. (10/30)	The coherence is poor. (10/30)
There are a few unclear expressions and language errors, and some errors are serious ones. (21/40)	It is basically to the point. (16/30)	There is bare coherence. (16/30)
There are a few language errors with clear expressions. (29/40)	It is to the point. (22/30)	There is a coherent organization. (22/30)
There are clear expressions and basically no language errors. (37/40)	It is exemplarily to the point. (28/30)	There is a cohesive and coherent organization. (28/30)

*The rubrics are adapted from rubrics for rating CET6 writing in China. The ratter may add or subtract scores based on the five levels at discretion.

The inter-rater reliability analysis between scores from every two raters was conducted. The result showed that they have significant ($p < 0.01$) and high inter-rater reliabilities, which were from 0.710 to 0.785.

References

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Wen, Q., Wang, L., & Liang, M. (2008). *Spoken and Written English Corpus of Chinese Learners (Version 2.0)*. Foreign Language Teaching and Research Press.