Measuring the Power of Learning.®

Research Report
ETS RR–16-37

Pilot Testing the Chinese Version of the *ETS*® Proficiency Profile Critical Thinking Test

Ou Lydia Liu

Liyang Mao

Tingting Zhao

Yi Yang

Jun Xu

Zhen Wang

December 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Pilot Testing the Chinese Version of the *ETS*® Proficiency Profile Critical Thinking Test

Ou Lydia Liu,[1] Liyang Mao,[2] Tingting Zhao,[3] Yi Yang,[3] Jun Xu,[1] & Zhen Wang[1]

1 Educational Testing Service, Princeton, NJ
2 IXL Learning, San Mateo, CA
3 Beihang University, Beijing, China

Chinese higher education is experiencing rapid development and growth. With tremendous resources invested in higher education, policy makers have requested more direct evidence of student learning. However, assessment tools that can be used to measure college-level learning are scarce in China. To mitigate this situation, we translated the critical thinking test from the *ETS*® Proficiency Profile (EPP) into Chinese. EPP has been widely used in the United States to assess general college learning outcomes. We pilot tested the EPP–Chinese test with students from a university in China. Results suggest that (a) the test is unidimensional and therefore is sufficient to report a total score from a practical standpoint; (b) the total score reliability is satisfactory; (c) most items showed moderate correlations with the total score, but the translation of one item needs additional revision; (d) the test is correlated with related constructs (e.g., the Chinese college entrance examination and a national English test); and (e) no item showed differential item functioning or was found to be biased toward any subgroup. In summary, the Chinese version of the critical thinking test showed potential as a suitable assessment tool for Chinese college students.

Keywords Critical thinking; student learning outcome; assessment; psychometrics

doi:10.1002/ets2.12123

China has one of the fastest growing higher education systems in the world. In 1949, when the People's Republic of China was first founded, approximately 117,000 students were enrolled in undergraduate education (Pepper, 1980). During the Cultural Revolution (1966–1976), China's higher education system stalled as many students were forced to terminate their study for various political reasons. The Chinese government has made an unparalleled commitment to develop higher education since the ending of the Cultural Revolution. In 1966 and 1978, the numbers of undergraduate students in China were 600,000 and 800,000, respectively, showing an increase of only 200,000 over the two decades (Yu, 2009). The number increased to 1.63 million in 1995 and reached 27 million in 2008 (Ministry of Education of China, 2014). Similarly impressive, from 1952 to 2014, the number of colleges and universities in China expanded more than 10 times, from 201 to 2,542 (Ministry of Education of China, 2015; National Bureau of Statistics of China, 1998). Despite the tremendous governmental investments in higher education, insufficient attention has been paid to the quality of higher education and the outcomes of student learning. The critical question of how much learning students achieve through college is often neglected in evaluations of the effectiveness of higher education in China. Instead, such evaluations tend to focus on indicators such as faculty's research productivity, graduation rates, academic atmosphere, and compliance with policy, which do not necessarily provide direct evidence of student learning (Liu, 2013; Rasmussen & Zou, 2014).

In addition to the growth of higher education, two other factors in the changing landscape of China's higher education, globalization of higher education and a competitive workforce, have brought student learning outcomes into the spotlight. With the aim of capitalizing on advanced models of higher education in other parts of the world, in particular those in Western countries, Chinese higher education has steadily globalized through the adoption of student exchange programs, study abroad programs, joint programs, and, in some cases, joint institutions. Examples include New York University's campus in Shanghai and Duke University's branch in Kunshan (Feng, 2012; Kirby, 2014; Sexton, 2012). Such global integration necessitates the evaluation of Chinese students' learning outcomes in a way that facilitates global comparison. For instance, key stakeholders would be interested to know the degree to which Chinese college students have equipped themselves with 21st-century skills deemed important universally, regardless of their chosen discipline and location of study.

Corresponding author: O. L. Liu, E-mail: lliu@ets.org

The competitive labor market in China is another factor that has propelled employers to focus more on the competencies that college graduates can demonstrate rather than the brand name of their alma mater. A college degree in China 40 years ago almost always led to a decent job, high income, and social prestige; none of these is guaranteed in the current Chinese job market. According to the *Chinese College Graduates' Employment Annual Report* (MyCOS Institute, 2014), among the 7 million college graduates in 2013, approximately 600,000 were still having trouble finding a job after 6 months. This climate makes it all the more important for students to demonstrate the knowledge and skills required to succeed in the global workforce in a direct and measureable way.

Evaluation of Higher Education Learning Outcomes in China

As Chinese higher education undergoes rapid expansion, quality control becomes a central concern for policy makers, educators, and researchers. In 2002, with the aim of improving quality and increasing efficiency, the Chinese Ministry of Education initiated an effort, called the Quality Assessment of Undergraduate Education (QAUE), to undertake the evaluation of institutional effectiveness. Under the provisions of this project, institutions will be evaluated every 5 years on a rolling basis (Liu, 2013). This state-run effort set eight broad categories of criteria for evaluation, with student learning outcomes as one of the target features. However, there is a remarkable lack of transparency in terms of the tools used to measure student learning outcomes and the comparison of such results across institutions. As a result, QAUE remains controversial and has been criticized by scholars as failing to improve quality (Liu, 2013; Zhou, 2010).

One of the important issues that has limited the success of the QAUE effort is the lack of standardized tools that can be used to evaluate student learning outcomes among Chinese higher education institutions. Unlike in the United States, where many institutions use commercial or in-house assessments to assess students' general competencies, such as critical thinking, written communication, and quantitative literacy, the concept of assessing generic learning outcomes among college students is still novel in China. One of the few standardized assessments that provides direct evidence of students' learning is the Chinese Test of English Band 4 (CET-4), which tests college students' English proficiency in listening comprehension, reading, English–Chinese translation, and essay writing (Zheng & Cheng, 2008). Since its launch in 1987, CET-4 has been broadly used in China as a requisite for college graduation and is currently valued by many employers when evaluating job candidates' English skills.

Another noteworthy effort in the evaluation of higher education in China is the translation and adaptation of the National Survey of Student Engagement (NSSE). NSSE is the most widely used survey in the United States for evaluating students' engagement and experiences in college. Through its self-report student survey, the NSSE collects information in five categories: (a) participation in educational activities, (b) institutional requirements and the nature of college course work, (c) perceptions of the college environment, (d) perceptions of educational and personal growth since entering college, and (e) students' backgrounds and demographic information (Kuh, 2002). As of 2014, it had been used by more than 1,500 institutions in the United States.[1] In 2007, researchers from Indiana University (Bloomington) and Tsinghua University in China initiated the translation and adaptation of the NSSE into Chinese (Ross, Luo, & Shen, 2008). Researchers have provided reliability and validity evidence for the NSSE–China survey (Luo, Ross, & Shen, 2009; Shi, Wen, Yifei, & Jing, 2014) and confirmed the suitability of using such a tool among Chinese college students.

Another instrument used by Chinese universities is the Student Experience in the Research University (SERU) survey, developed by researchers at the University of California, Berkeley (Brint & Cantwell, 2010; Kim & Sax, 2009), which aims to measure students' engagement and experiences at research universities, such as academic engagement, global skills and awareness, community and civic engagement, and student development. Using a Chinese version of the SERU, Gong and Lv (2012) found that compared to students at Berkeley, Chinese students at Nanjing University obtained lower scores on three dimensions of the SERU: (a) classroom discussion and initiative, (b) peer collaboration and interaction, and (c) critical reasoning and creative thinking.

Despite the usefulness of survey tools such as the NSSE–China and SERU, they do not provide direct evidence of student learning. While the aggregate data may provide insight into the overall academic environment that an institution is able to offer, the survey results are not of much value to individual students, as the results do not speak directly to a student's competency in any area.

Objectives

Given that a standardized tool is urgently needed to provide direct evidence of college student learning in China, we report on an effort to translate, adapt, and investigate the psychometric properties of the critical thinking test from the *ETS*® Proficiency Profile (EPP). The purpose of this study is to investigate the psychometric properties of the EPP critical thinking test among Chinese college students. Specifically, we address the following research questions:

1. What are the psychometric qualities of the critical thinking test in terms of reliability, item–test correlation, and dimensionality?
2. How does the critical thinking test correlate with related constructs such as the Gaokao (i.e., college admission test in China) and CET-4?
3. Is there any differential item functioning (DIF) with regard to gender, major, and socioeconomic status (SES)?
4. Is there any performance difference on the critical thinking test across subgroups of interest?

This preliminary investigation sheds light on the suitability of using such a test to assess Chinese college students' critical thinking skills. As Chinese institutions are increasingly emphasizing quality and accountability, due to the influence of domestic and international factors, identification of assessment tools that can be used for such evaluation becomes a priority on the research agenda (Huang, 2005).

Method

Instrument

The original English critical thinking test is one of the four sections included in the EPP, a college-level general skills assessment that measures critical thinking, reading, writing, and mathematics. The critical thinking test measures students' ability to (a) distinguish between rhetoric and argumentation in a piece of nonfiction prose, (b) recognize assumptions, (c) recognize the best hypothesis to account for information presented, (d) infer and interpret a relationship between variables, and (e) draw valid conclusions based on information presented (ETS, 2010). The 27-item critical thinking test consists of 15 stand-alone items and six testlets (i.e., a set of items based on a common reading passage) with two items within each testlet. Although the EPP critical thinking test targets general skills, meaning that specific disciplinary knowledge is not required to answer the items, the assessment items are embedded in three broad domains: arts and humanities (nine items), social sciences (nine items), and natural sciences (nine items).

In the United States, EPP has been widely used as a tool for accreditation, accountability, and institutional internal improvement (Liu, 2011a). EPP has also been researched extensively in terms of its internal structure (Lakin, Elliott, & Liu, 2012), relationship to tests of similar constructs (Klein et al., 2009), predictive validity in terms of predicting GPA and other college outcomes (Hendel, 1991; Liu & Crotts, 2013; Marr, 1995), DIF with regard to language status (Lakin et al., 2012), suitability for measuring value-added learning (Liu, 2011b; Liu, Bridgeman, & Adler, 2012), and the effect of students' test-taking motivation on EPP scores (Liu et al., 2012; Liu, Rios, & Borden, 2015; Rios, Liu, & Bridgeman, 2014).

Translation Procedure

The critical thinking test of the EPP was translated into Chinese through the following steps: (a) Two groups of translators in China translated the test separately; (b) the two groups of translators then met and discussed the discrepancies in the translation until an agreement was reached; (c) a panel of assessment experts who were bilingual in English and Chinese reviewed and refined the translation; (d) a user acceptance study was conducted with a small group of college students in China to see if any of the translated items could be improved in terms of clarity and also to determine the proper amount of testing time for Chinese students; (e) a final version of the test was created based on all the previously described steps.

Sample

A total of 1,009 college students from a local institution in China voluntarily participated in this pilot study in fall 2014. This institution is a nonelite (Tier 2) university. The participants were a stratified sample (i.e., a random sample separately

Table 1 Demographic Information

	Local institution	
Demographic information	*N*	%
Gender		
Female	684	68
Male	267	26
Missing	58	6
Major		
Arts and humanities	274	27
Social sciences	185	18
Natural sciences	352	35
Business	140	14
Missing	58	6
Class status		
Freshman	266	26
Sophomore	261	26
Junior	248	25
Senior	176	17
Missing	58	6
Parental education		
Associate degree or above	145	14
High school or below	663	66
Missing	201	20

drawn from freshmen, sophomores, juniors, and seniors). The demographic information of the sample is shown in Table 1. In addition to the demographic information, participants also reported their CET-4 test scores, Gaokao scores, and where they took the Gaokao, as the content of the test varies across provinces in China.

Analysis

Analyses were conducted to address the four research questions. Cronbach's alpha and the standard error of measurement (SEM) of the raw total score (i.e., number correct score) were used to evaluate the test's reliability. Item difficulty and item-total biserial correlation were then calculated to evaluate the performance of each item. In terms of dimensionality, we evaluated the model fit of a unidimensional model based on confirmatory factor analysis in Mplus 7.3 (Muthén & Muthén, 2012). The model fit was indicated by the goodness-of-fit indices, including χ^2, the comparative fit index (CFI), the Tucker – Lewis index (TLI), and the root mean square error of approximation (RMSEA). A good model fit was suggested by a nonsignificant χ^2 at the .05 level, CFI and TLI values greater than .90, and an RMSEA value less than .05 (Cheung & Rensvold, 2002).

In addition, we obtained the Pearson correlation between the critical thinking scores with CET-4 scores and Gaokao scores. As previously discussed, CET-4 is a standardized English proficiency test for college students in China. It is administered twice a year, and scores from different test administrations are equated and therefore can be used interchangeably. The internal consistency reliability for the CET-4 was consistently over .90 (Zheng & Cheng, 2008). Gaokao, an annual achievement test for high school graduates, is a prerequisite for entrance into undergraduate education for almost all higher education institutions in China, and "it is thought to be the most typical and standard test with moderate difficulty, high reliability and validity in China" (Chen, Cheng, Chang, Zheng, & Huang, 2014, p. 218). Unlike CET-4, Gaokao varies across provinces in China, and the scores are not comparable across years. We only used data from students who took Gaokao in Jiangsu Province in 2014 because the majority of the students (88%) who reported their Gaokao scores were from Jiangsu. The Jiangsu form was developed based on China's national high school curriculum (Ministry of Education Testing Center, 2014). It covered five subject areas: Chinese, English, mathematics,[2] science, and liberal arts. All students took the Gaokao Chinese, English, and Mathematics. Depending on their intended major field (e.g., science or liberal arts), each student takes a fourth test, either Science or Liberal Arts. In this study, Pearson correlations were obtained between the critical thinking score and each subtest score for Gaokao.

Furthermore, we conducted DIF analyses to see if subgroups with matched ability levels performed differently on the critical thinking test. As DIF may be an indicator of potential unfairness, items with DIF need further review to determine if they are unfair toward certain subgroups. In this study, we analyzed DIF between men and women, between natural science majors and other majors, and between high- and low-SES groups as indicated by parents' education level. The Mantel–Haenszel (MH) method (Mantel & Haenszel, 1959) was employed to perform the DIF analyses. We classified all the items as Category A, B, or C, depending on the magnitude of the MH delta difference (MH D-DIF) statistic and its statistical significance (Dorans & Holland, 1992; Holland & Thayer, 1988). Category A, B, and C indicates negligible, moderate, and large DIF, respectively. Zwick (2012) provided a more detailed description of these categories. To be consistent with the operational practice of the EPP program (ETS, 2010), this study only flagged items of Category C DIF for fairness review.

Last, when examining the performance difference by gender, major, and parents' education, we reported the descriptive statistics, test significance, and effect sizes in Cohen's d (for t-test) or η^2 (for ANOVA). Cohen's d expresses the mean differences in standard deviation units (Cohen, 1988). For η^2, an effect size of .01 is a small effect, .06 is a medium effect, and .14 is a large effect (Cohen, 1988).

Results

Psychometric Qualities of the ETS Proficiency Profile Critical Thinking (Chinese) Test

Reliability

The Cronbach's alpha for the EPP critical thinking test (Chinese) was .71, and the overall SEM of the raw total score was 2.31, indicating a satisfactory reliability of the test. The reliability of the EPP English version is .78 (ETS, 2010), slightly higher than for the Chinese version.

Item Analysis

Table 2 presents the item analysis results. All the items, except for Item 5, showed small to moderate correlation with the total score. The negative correlation for Item 5 suggests that this item was not functioning properly on the test. The detailed item analysis results for Item 5 are shown in Table 3. The percentage of the students who selected C (the correct answer) for the top 20% students is lower than the percentage for all students, which suggests poor discrimination for this item. One possible reason is related to how a particular word was translated into Chinese (for test security purposes, we cannot reveal that word here). The translation was correct, but the translated word assumed a different meaning in the context, which may have affected students' responses. Because Item 5 showed negative correlation with the total score, it was removed from all the following analyses. The new total score based on the remaining 26 items was calculated, and the new item-total correlation is shown in the last column of Table 2. After removing Item 5, the correlation for some items slightly increased.

Dimensionality Analysis

The goodness-of-fit indices suggested acceptable model fit for the unidimensional model, $\chi^2(299, N = 1009) = 583.16$, $p < .001$ (CFI = .91; TLI = .90; RMSEA = .03). Therefore it is sufficient only to report a total score from a practical standpoint.

Relationship With Gaokao and the Chinese Test of English Band 4

The correlations between the EPP critical thinking test (Chinese) and Gaokao are shown in Table 4. As previously discussed, we only used the scores from the Jiangsu form. The scores from the EPP critical thinking test (Chinese) showed moderate correlations with scores from Gaokao English and small correlations with Gaokao Mathematics (liberal arts track). The correlation between scores from the EPP critical thinking test (Chinese) and the Gaokao Mathematics is

Table 2 Item Analysis Results

Item no.	% correct	*SD*	Item-total correlation with Item 5	Item-total correlation without Item 5
1	.80	.40	.10	.11
2	.76	.43	.40	.42
3	.66	.48	.21	.22
4	.63	.48	.27	.27
5	.18	.39	−.17	–[a]
6	.19	.39	.05	.05
7	.76	.43	.31	.31
8	.80	.40	.12	.12
9	.46	.50	.10	.10
10	.62	.49	.29	.30
11	.55	.50	.18	.19
12	.60	.49	.28	.27
13	.46	.50	.18	.18
14	.58	.49	.33	.33
15	.54	.50	.25	.25
16	.78	.42	.46	.46
17	.63	.48	.27	.27
18	.71	.45	.28	.28
19	.40	.49	.11	.11
20	.45	.50	.19	.19
21	.56	.50	.39	.39
22	.81	.40	.36	.35
23	.44	.50	.24	.23
24	.62	.49	.37	.37
25	.65	.48	.39	.39
26	.53	.50	.34	.34
27	.68	.47	.22	.22

[a]No data were available in this category.

Table 3 Item Analysis Results for Item 5

Response	*N*	Percentage	Mean	*SD*	Top 20%
A	172	17.1	16.7	4.4	22.6%
B	72	7.1	14.7	4.5	2.4%
C (key)	183	18.1	15.1	4.5	16.0%
D	575	57.0	16.1	4.0	59.0%
Other[a]	7	0.7	12.9	5.6	0.0%

[a]Included missing responses and invalid responses such as "E" or "AD."

much lower than that ($r = .46$) between the EPP critical thinking test[3] (reliability $\alpha = .78$; ETS, 2010) and the Collegiate Assessment of Academic Proficiency (CAAP) Mathematics (reliability $\alpha = .85$; ACT, 2008) found by Klein et al. (2009). One reason is that the Gaokao Mathematics measures much more in-depth mathematical knowledge and skill than does the CAAP Mathematics. The Gaokao Mathematics measures high school students' ability to solve complex math problems (Ministry of Education Testing Center, 2014), whereas the CAAP Mathematics measures college students' proficiency in mathematical reasoning (ACT, 2008). It is expected that critical thinking skills show a higher correlation with mathematical reasoning than with the ability to solve complex math problems. In addition, students in this study took Gaokao at the end of their senior year in high school and the EPP critical thinking test (Chinese) in college, whereas students in Klein et al.'s (2009) study took CAAP and EPP simultaneously.

Among all the participants in this study, 497 students reported their CET-4 scores (see Table 5). Because college students in China usually take CET-4 after they finish their freshman year, only one freshman reported a CET-4 score. The correlation between the EPP critical thinking test (Chinese) and the CET-4 was .09, $p < .05$, suggesting a small correlation between critical thinking skills and English proficiency level.

Table 4 Correlation Between ETS Proficiency Profile Critical Thinking (Chinese) and Gaokao

		EPP critical thinking (Chinese)	
Gaokao form	Subject	*N*	Pearson's *r*
Jiangsu	Chinese	172	.12
	English	177	.43***
	Mathematics (liberal arts track)	259	.13*
	Mathematics (science track)	260	.09
	Liberal arts	57	.06
	Science	58	−.17

Note. EPP = ETS Proficiency Profile.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5 Number of Students Who Took the Chinese Test of English Band 4

Major	Freshmen	Sophomores	Juniors	Seniors	Total
Arts and humanities	0	43	48	34	125
Social sciences	0	39	32	28	99
Natural sciences	1	80	66	53	200
Business	0	31	23	19	73
Total	1	193	169	134	497

Differential Item Functioning

DIF analyses were conducted with regard to gender, major, and SES group (Table 6). A negative MH D-DIF suggested that the item showed DIF in favor of men, natural sciences majors, or students of high SES (i.e., parents' highest education is associate degree or above). The presence of DIF is a signal that the item may be unfair, but statistical bias does not necessarily imply that the item is unfair (Penfield & Camilli, 2007). Therefore DIF items need to go through fairness review. If, and only if, fairness review suggests that the presence of DIF is attributed to an unintended item content or property, the item is considered to be unfair (Penfield & Camilli, 2007). The results show that no items were associated with Category C DIF for gender, major, or SES group.

Performance Difference Across Subgroups

The performance gap between subgroups (i.e., gender, major, and SES) is shown in Table 7. The effect size in Cohen's *d* is also presented in the last column of Table 7. Women significantly outperformed men, $t_{961} = 5.23$, $p < .001$, and the effect size of the difference was .39, indicating that women scored .39 standard deviations higher than men on average. Students showed differential performance by major on the EPP critical thinking test (Chinese: $F_{3,959} = 12.45$, $p < .001$, $\eta^2 = .037$). Specifically, students in arts and humanities majors scored lowest among all. A Gender × Major ANOVA was also conducted to clarify the gender and major differences. The results indicate no interaction effect, $F_{3,955} = 2.07$, $p = .10$, partial $\eta^2 = .006$.

There was no achievement gap (mean difference of .02 standard deviations) between high- and low-SES groups. Students from high-SES backgrounds performed similarly to students from low-SES backgrounds, $t_{817} = .26$, $p = .79$.

Discussion and Conclusion

In this study, we translated the critical thinking test from the EPP into Chinese and pilot tested it with students from one university in China. The EPP critical thinking test (Chinese) showed satisfactory psychometric properties in terms of internal consistency and item – test correlations, except for one item, which was later removed. The dimensionality analysis revealed that the test consists of one general dimension with three subdimensions corresponding to the three contexts of the items (i.e., arts and humanities, social sciences, and natural sciences). The test also showed reasonable correlations with related constructs such as Gaokao and CET-4. No items showed DIF. Our subgroup analysis revealed that female

Table 6 Differential Item Functioning Across Gender, Major, and Socioeconomic Status Groups

	Gender DIF		Major DIF		SES DIF	
Item	MH D-DIF	Category	MH D-DIF	Category	MH D-DIF	Category
S1	0.44		−0.21		0.07	
S2	−0.83		−0.94*		0.25	
S3	−0.37		−0.19		0.41	
S4	0.60		0.38		−0.23	
S6	0.66		−0.04		0.44	
S7	−0.21		−0.10		−0.41	
S8	0.05		0.76		−0.55	
S9	−0.16		0.67*		−0.57	
S10	−0.31		0.25		−0.57	
S11	0.24		0.32		−0.12	
S12	−0.78*		−0.76*		−0.06	
S13	0.03		0.08		0.14	
S14	−1.03**	B	−0.07		0.65	
S15	−0.29		−0.18		0.15	
S16	0.24		0.06		−0.57	
S17	1.10**	B	0.31		−0.19	
S18	−0.09		−0.70		0.30	
S19	−0.56		0.18		0.61	
S20	−0.14		−0.16		0.10	
S21	0.23		−0.74*		0.34	
S22	0.80		0.91*		0.14	
S23	−0.54		0.19		−0.65	
S24	0.62		0.45		−0.28	
S25	−0.30		−0.76		0.13	
S26	0.20		0.14		0.17	
S27	0.74*		−0.03		0.17	

Note. DIF = differential item functioning. MH-DIF = Mantel–Haenszel delta difference. SES = socioeconomic status.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 7 Performance Gap Among Subgroups

Group	*N*	*M*	*SD*	Cohen's *d*			
Gender				−.39			
Female	687	16.18	4.17				
Male	276	14.51	4.61				
Major				AH	SS	NS	BN
Arts and humanities (AH)	274	14.44	4.68	–			
Social sciences (SS)	197	16.60	4.05	.49	–		
Natural sciences (NS)	352	16.21	4.32	.39	−.09	–	
Business (BN)	140	15.66	3.68	.28	−.24	−.13	–
Parental education					.02		
High school or below	674	15.72	4.38				
Associate degree or above	145	15.82	4.39				

students outperformed male students and that students majoring in the arts and humanities did not perform as well as their peers in other major fields of study. There were no SES group differences. In general, results from this pilot provided some preliminary evidence to support the use of this test among Chinese college students.

Our findings reveal differences in critical thinking skills by gender and major field of study among Chinese college students. Given that research on critical thinking is still relatively new in China, not many prior studies are available to triangulate results. Among the few relevant studies that can be identified, Jiang (2012) analyzed Chinese college students' critical thinking skills using the California Critical Thinking Skills Test (CCTST; Facione, 1990) and the California Critical Thinking Disposition Inventory (CCTDI; Facione & Facione, 1992). The participants were from three nonelite universities in Shanghai. Key findings from that study include that (a) more than 50% of students were classified as not

proficient or partially proficient on the CCTST; (b) on average, students scored lower than the U.S. CCTDI norms; (c) those who majored in sciences scored significantly higher than those who majored in liberal arts; (d) juniors significantly outperformed freshmen and sophomores; and (e) men and women showed equal performance. Science majors' superior performance in China may be explained by a number of factors. One is that in China, the kind of major that a student can get into depends on his or her scores on the Gaokao. Typically, science-related majors are more selective than liberal arts majors, and therefore students enrolled in science majors may have higher prior achievement than their counterparts in liberal arts majors. Another reason is that in general, college curricula are more rigorous for science majors, possibly providing more opportunities for students to enhance their critical thinking skills.

These findings provide preliminary evidence for the use of a translated and adapted critical thinking test for Chinese college students. As the Chinese government calls for enhanced accountability and encourages institutions to provide direct evidence of student learning, tools such as the critical thinking test studied here could be used to gather such evidence. Given the standardized nature of the translated critical thinking test, it has potential to produce data that will allow institutions to evaluate student learning by subgroups of interest internally as well as compared to other institutions of similar size and setting. The critical thinking Chinese test also has the potential to document value-added learning as students progress from freshman to senior year. In addition, our finding that arts and humanities students are lagging behind in critical thinking skills provides some indication for the need for institutions to further examine the performance of liberal arts students.

A limitation of this pilot study is that the sample was only from one local, less selective university in China. The findings from this study may not apply to students from other less selective universities or from more selective universities. Our next-step, scale-up study will include a sample that is more representative of the college population in China, balancing geographical regions, selectivity of institution, and student demographics.

Going forward, we would like to gather more evidence to further validate the Chinese version of the EPP critical thinking test. We plan to address the translation issue in Item 5 and pilot test the revised version of the EPP critical thinking (Chinese) test with a larger and more representative sample. We also plan to examine the relationship between critical thinking scores and students' performance in related courses that promote reasoning and analytical skills. In addition, because many Chinese universities sponsor undergraduate research programs aimed at promoting undergraduates' critical thinking skills, it would be important to see if program participants show any improvement on the critical thinking test in a pre- and posttest scenario. Another direction of research would be focused on institution-level learning gain. Prior research has shown that the EPP critical thinking test (English) is able to detect learning gains such as students' progress from freshman to senior year (Klein et al., 2009; Liu, 2011b). Such evidence is needed to determine if the translated version can also be used to measure value added at an institution.

Notes

1 http://nsse.iub.edu/html/about.cfm
2 Note that the mathematics test is different for science- and liberal arts-track students, with items being more difficult for science-track students.
3 The EPP test was formerly named the Measure of Academic Proficiency and Progress (MAPP).

References

ACT. (2008). *CAAP technical handbook*. Iowa City, IA: Author.

Brint, S., & Cantwell, A. M. (2010). Undergraduate time use and academic outcomes: Results from the University of California Undergraduate Experience Survey 2006. *Teachers College Record, 112*, 2441 – 2470.

Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education*, *1*, 213 – 225.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233 – 255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mentel – Haenszel and standardization* (Research Report No. RR-92 – 10). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.1992.tb01440.x

ETS. (2010). *ETS Proficiency Profile user's guide.* Retrieved from https://www.ets.org/s/proficiencyprofile/pdf/Users_Guide.pdf

Facione, P. A. (1990). *The California Critical Thinking Skills Test – college level: Factors predictive of CT skills* (Technical Report No. 2). Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. C. (1992). *The California Critical Thinking Dispositions Inventory (CCTDI)*. Millbrae, CA: California Academic Press.

Feng, Y. (2012). University of Nottingham Ningbo China and Xi'an Jiaotong-Liverpool University: Globalization of higher education in China. *Higher Education, 65*, 471 – 485.

Gong, F., & Lv, L. (2012). 中美研究型大学本科生学习参与差异的研究——基于南京大学和加州大学伯克利分校的问卷调查 [The comparison of undergraduate students' learning engagement between research universities in China and the USA]. *Journal of Higher Education, 33*(9), 90 – 100.

Hendel, D. D. (1991). Evidence of convergent and discriminant validity in three measures of college outcomes. *Educational and Psychological Measurement, 51*, 351 – 358.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel – Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129 – 145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Huang, F. (2005). Qualitative enhancement and quantitative growth: Changes and trends of China's higher education. *Higher Education Policy*, *18*, 117 – 130.

Jiang, J. (2012). *上海地区大学生批判性思维现状调查及其与人格的关系研究* [College students' critical thinking skill in Shanghai area and its relationship with personality] (Unpublished master's thesis). Shanghai Normal University, Shanghai, China.

Shi, J., Wen, W., Yifei, L., & Jing, C. (2014). China College Student Survey (CCSS): Breaking open the black box of the process of learning. *International Journal of Chinese Education, 3*, 132 – 159.

Kim, Y. K., & Sax, L. J. (2009). Student – faculty interaction in research universities: Differences by student gender, race, social class, and first-generation status. *Research in Higher Education, 50*, 437 – 459.

Kirby, W. C. (2014). The Chinese century? The challenges of higher education. *Daedalus, 143*, 145 – 156.

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., … Steedle, J. (2009). *Test Validity Study (TVS) report.* Retrieved from http://www.voluntarysystem.org/index.cfm?page=research

Kuh, G. D. (2002). *The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties*. Bloomington, IN: Center for Postsecondary Research, Indiana University. Retrieved from http://nsse.iub.edu/pdf/psychometric_framework_2002.pdf

Lakin, J. M., Elliott, D. C., & Liu, O. L. (2012). Investigating ESL students' performance on outcomes assessments in higher education. *Educational and Psychological Measurement, 72*, 734 – 753.

Liu, O. L. (2011a). Outcomes assessment in higher education: Challenges and future research in the context of Voluntary System of Accountability. *Educational Measurement: Issues and Practice, 30*(3), 2 – 9.

Liu, O. L. (2011b). Value-added assessment in higher education: A comparison of two methods. *Higher Education, 61*, 445 – 461.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes assessment in higher education: Motivation matters. *Educational Researcher, 41*, 352 – 362.

Liu, O. L., & Crotts, K. (2013). *Investigating ten-year trends of learning outcomes at community colleges* (Research Report No. RR-13-34). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2013.tb02341.x

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79 – 94.

Liu, S. (2013). Quality assessment of undergraduate education in China: Impact on different universities. *Higher Education, 66*, 391 – 407.

Luo, Y., Ross, H., & Shen, Y. (2009). 国际比较视野中的高等教育测量 — NSSE-China 工具的 开发:文化适应与信度、效度报告 [Higher education measurement in the context of globalization — The development of NSSE-China: Cultural adaptation, reliability and validity]. *Fudan Education Forum, 7*(5), 12 – 18.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719 – 748.

Marr, D. (1995). *Validity of the academic profile*. Princeton, NJ: Educational Testing Service.

Ministry of Education of China. (2014). *Number of students in higher education institutions*. Retrieved from http://www.moe.gov.cn/publicfiles/business/htmlfiles/moe/s8493/201412/182071.html

Ministry of Education of China. (2015). *Number of regular students enrolled in normal and short-cycle courses in higher education*. Retrieved from http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/s8494/201412/182315.html

Ministry of Education Testing Center. (2014). *2014年普通高等学校招生全国统一考试大纲* [Manual for the National Higher Education Entrance Examination 2014]. Beijing, China: Higher Education Press.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus 7 base program.* Los Angeles, CA: Muthén and Muthén.

MyCOS Institute. (2014). *中国大学生就业报告* [Chinese College Graduates Employment Annual Report 2014], Beijing, China: Social Sciences Academic Press.

National Bureau of Statistics of China. (1998). *中国统计年鉴* [China statistical yearbook]. Beijing, China: China Statistics Press.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. *Handbook of Statistics, 26,* 125–167.

Pepper, S. (1980). Chinese education after Mao: Two steps forward, two steps back and begin again? *The China Quarterly, 81*, 1–65.

Rasmussen, P., & Zou, Y. (2014). The development of educational accountability in China and Denmark. *Education Policy Analysis Archives, 22*(121), 1–26.

Rios, J., Liu, O. L., & Bridgeman, B. (2014). Identifying unmotivated examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69–82.

Ross, H., Luo, Y., & Shen, Y. (2008). Assessing Tsinghua University and US universities on learning process indicators: An approach of higher education quality. *Tsinghua Journal of Education*, *29*(2), 36–42.

Sexton, J. (2012). The global network university: Educating citizens of the world. In M. Stiasny & T. Gore (Eds.), *Going global: The landscape for policy makers and practitioners in tertiary education* (pp. 5–12). Bingley, England: Emerald House.

Yu, Y. (2009). *20 世纪的中国高等教育* [The higher education of China in the 20th century]. Beijing, China: Higher Educational Press.

Zheng, Y., & Cheng, L. (2008). Test review: College English Test (CET) in China. *Language Testing, 25*, 408–417.

Zhou, X. (2010). 从政府问责到社会问责：中国高校问责制的内涵、类型与变革 [From government accountability to social accountability: Connotation, types and change of Chinese university accountability]. *Journal of Higher Education, 31*, 34–40.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2012.tb02290.x

Suggested citation:

Liu, O. L., Mao, L., Zhao, T., Yang, Y., Xu, J., & Wang, Z. (2016). *Pilot testing the Chinese version of the ETS® Proficiency Profile critical thinking test* (ETS Research Report No. RR-16-37). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12123

Action Editor: John Sabatini

Reviewers: Daniel McCaffrey and Larry Stricker