Article

Conceptions of Assessment as an Integral Part of Language Learning: A Case Study of Finnish and Chinese University Students

Magdalini Liontou [1,2]

1 Centre for Applied Language Studies, Faculty of Humanities, University of Jyväskylä, 40014 Jyväskylä, Finland; malionto@jyu.fi
2 Languages and Communication, Extension School, University of Oulu, 90570 Oulu, Finland

Abstract: Assessment is viewed as an internal and pivotal part of learning, where cultural factors, previous experiences, and future aspirations affect learners' perceptions. In recent years, an increasing number of western universities have established their campuses or "dual" programmes in China. In the first Sino–Finnish programme, 293 Finnish and Chinese students participated in the same English for Specific Purposes (ESP) course. This study investigated students' perceptions of assessment through an adapted version of the "Students' Conceptions of Assessment" inventory, and it explored if the responses on each conception differ between the groups. The self-report inventory included statements based on four main conceptions of assessment: improvement, external factors, affect/benefit, and irrelevance, while open-ended questions were also included. The analysis of the open-ended questions raised the issues of teacher fairness, learner autonomy, and feedback. Additionally, differences appeared between the role of assessment and its relation to future aspirations, as well as the role of the parents. This study is a starting point for exploring the conceptions for distinct groups of students regarding assessment, providing a better understanding of students' perceptions and discussing the implications for the language classroom.

Keywords: assessment and learning; students' conceptions; intercultural awareness

1. Introduction

Assessment is an internal and pivotal part of learning. As Saville and Khalifa (2016) point out, the continuous effects of language assessment are visible on two levels. On the macro-level, it involves language policies in the society and consequently in various educational systems. On the micro-level, assessment is related to the experiences of individuals participating in it. Hence, it is vital to involve students as one of the main stakeholders who encounter language assessment practices, and they could face its long-lasting effects on their education and careers. For example, high-stakes exams could determine the life decisions and plans of test-takers.

Several studies have investigated how students' attitudes of assessment affect their learning behaviour, pointing out four major themes/conceptions with regards to assessment. Initially, the conception "assessment improves learning" focuses on assessment as a tool (1) for students to evaluate their performance and improve learning and (2) for teachers as a way to understand their students' performance and consequently improve their instructions (Brown 2011). These students think that assessment could have a formative nature and inform them for their next steps of their learning; additionally, by showing self-responsibility, they tend to achieve greater results (Brown et al. 2014; Brown and Hirschfeld 2008). Secondly, research has shown that the students' locus of control affects assessment outcomes. Students with an inner locus of control believe that academic success derives from their own actions, and, subsequently, they tend to perform better (Kirkpatrick et al. 2008; Albert and Dahling 2016). In comparison, the ones who

associate assessment practices with an external locus of control have demonstrated poorer academic achievements and self-handicapping (Brown and Hirschfeld 2008; Stewart and George-Walker 2014). For example, an external locus of control could be associated with external factors regarding intelligence, future plans, parents' expectations, and school performance. Thirdly, the conception "assessment is irrelevant" (irrelevance) taps into the notion that assessment can be considered as unfair or subjective by the students (Brown 2011). Finally, statements regarding personal enjoyment for the classroom environment have been addressed to investigate the "assessment is liked" (affect/benefit) conception. The "irrelevance" and "benefit" conceptions have previously been linked to lower grades (Brown and Hirschfeld 2008).

1.1. Students' Perceptions of Language Assessment

For more than twenty years, various researchers have discussed the concept of context in language assessment. Chalhoub-Deville (2003, 2019) has discussed extensively how students' learning experiences could affect their decision on the resources likely to be used in a given situation, such as language assessment and the integration of local theories in language assessment practices. Moreover, Saville (2009) stresses the importance of local context and the attitudes of participants when designing assessment tasks with a focus on impact, while Jenkins and Leung (2019) refer to the participants taking part in language assessment as "local speakers" with unique needs and circumstances instead of generally "non-native speakers of English."

In the last decade, a few studies have investigated students' perceptions in terms of language assessment focusing on either secondary or higher education (HE). A quantitative study in Albania revealed that students in secondary education were divided into two almost equal subgroups who either considered tests as a motivating force or as stressful and paralysing. The majority of the responses indicated teacher feedback was given only after the tests, while there were almost no opportunities to practice peer- and self-assessment in the English-as-a-foreign-language (EFL) classroom (Vavla and Gokaj 2013). Additionally, a large-scale needs-analysis study (1788 participants) investigating learners' perceptions of language assessment practices was conducted across four European counties, Cyprus, Germany, Hungary, and Greece, as part of the Erasmus project "TALE." An interesting finding was the contrast between the teachers' and students' responses regarding both the types of assessment and feedback used in the classroom. The students considered classroom participation, testing, and writing as assessment forms that could promote learning. However, these were the assessment practices they were exposed to, and differentiation of the classroom assessment could potentially trigger different results (Vogt et al. 2020).

Perceptions of students regarding assessment in the EFL classroom have been investigated to some extent in HE. The consistency of assessment and learning and the employment of alternative assessment such as self- and peer-assessment were valued by HE students in a quantitative study carried out in Cyprus (Tsagari 2013). Another study conducted between three Chinese universities focusing on the students' perceptions of assessment tasks and assessment environment pointed out similar results. A learning-oriented environment was promoted when planned learning was in agreement with assessment tasks and was well-communicated to the students. Moreover, the teacher had a central role in language assessment, while students' efforts did not seem significant in their final assessment results. This is an important finding since the quality of assessment is affected by the students' views regarding their learning (Cheng et al. 2015). Both Cheng et al. (2015) and Tsagari (2013) mention that university degrees affect how EFL students perceive the purposes of assessment and the classroom assessment environment.

As it is evident from the perceptions of students in the empirical studies above, the local context has influenced to some extent their preferences and the way they experience language assessment. However, there is a lack of comparative research between different institutions across countries. The practical difficulties to accomplish this are the lack of common curricula and assessment criteria, as well as the various educational traditions

in different countries. Taking this into consideration, an opportunity to produce comparative research has arisen through an ESP course taught both in Finland and China. Both universities offered this course as part of the first Sino–Finnish double-degree engineering programme established in 2018. Acknowledging that culture and previous experiences of assessment practices could affect learners' conceptions of assessment, the aim was to explore these prior to the course in order to understand students' current perceptions better. This research study is the first step to clarify how the students' preconceptions of assessment affected their learning in the ESP course. Data has also been collected after the completion of the course. In the future, a comparison of their perceptions will be conducted. In order to understand better the two educational systems in which the students were exposed prior to tertiary education, a brief description of the Finnish and Chinese educational systems with a focus on language assessment is provided below.

1.2. Language Assessment Practices in Finland: Setting the Scene

Studying foreign languages is compulsory in secondary education in Finland. The Finnish students have, in principle, various options; nevertheless, English remains the most common one. In 2019, 99.7% of all students attending upper-secondary level education and 99.5% of students of vocational education selected to study English (OSF 2019). According to the national curriculum for foreign languages for upper-secondary education (EDUFI 2015), the target for English language learning is the B1-B2 CEFR-level descriptors. This variation is based on the number of courses the students choose during their studies. High-stakes exams do not dominate the Finnish educational system, while most of the assessment is classroom-based produced by individual teachers in primary and secondary education (Pollari 2016; Tarnanen and Huhta 2008).

The only high-stakes national exam, the Matriculation Exam (ylioppilastutkinto), is administered at the end of upper-secondary education. During this exam, the students are tested in their mother tongue (Finnish) and have to choose a minimum of three additional exams from four options: Swedish (as a second national language), mathematics, another foreign language (often English), and a course based on the humanities/natural sciences (Ylioppilastutkinto.fi n.d.). The foreign language exam consists of reading, listening, writing, and vocabulary and structures, while its grading is norm-referenced. Speaking is not included as part of the exam (Tarnanen and Huhta 2008). The performance of the majority of students who participated in the English-as-a-foreign-language exam (Englanti A path) was found to achieve or exceed level B2.1 (Huhta and Hildén 2016). Since 1919, the Matriculation Exam signifies the end of upper-secondary school, but it does not mark an instant acceptance to higher-education institutions (HEI) (Kaarninen and Kaarninen 2002). However, nowadays, the impact of the matriculation exam has grown, and it affects greatly university entrance policies. For example, at the official webpage of the University of Oulu (2020), it is stated that 80% of the students studying their undergraduate degrees in the fields of education, economics, and technology have been offered a position based on the results of this exam. Thus, the matriculation exam plays a significant role in the future plans/careers of the students.

Upon admission to an undergraduate degree, the students are required by law to participate in language courses. In particular, Chapter 1 (Section 6) of the "Government Decree on University Degrees and Professional Specialisation Programmes" (Finlex.fi 2014) focuses on language proficiency in HEI, and it is mentioned that "the student shall demonstrate attainment of the following standards in the studies included in the programme for a Bachelor's or Master's degree, or in some other manner: . . . 2) proficiency in at least one foreign language that enables students to monitor progress in their own field and operate in an international setting." Each university and department is responsible for interpreting and implementing the law since the universities specify the number of courses needed and the type of assessment to fulfil the language requirements (University of Oulu 2020). The types and administration of assessment in these language courses rest upon the university teachers who facilitate low-stakes classroom-based assessment.

Concluding, even though the Finnish undergraduate students have experienced a high-stakes exam and its potential washback effect, they are not expected to participate in a national high-stakes language exam during their undergraduate studies or immediately after completing their studies.

1.3. Language Assessment Practices in China

Assessment is deeply embedded in Chinese history and culture as traces of the first summative assessment practice, the imperial examinations (Keju), can already be found in ancient times (Ko 2017). As Cheng (2008) points out, the long tradition of summative assessment still continues to influence the Chinese society. This is evident in all levels of education, starting from the age of 4 when a child takes the entrance exam to enter kindergarten, and regions, such as from school- to municipal- and national-level exams. The students are familiar with different types of assessment of learning (summative assessment) as they have to participate in a series of "graduation examinations" during their primary and secondary education (Hu and West 2014). Moreover, the use of self- and peer-assessment is limited and interpreted under the Chinese cultural prism (Chen 2017; Poole 2016). For example, formative assessment is often expressed as attendance and/or grading rather than feedback (Chen et al. 2014; Chen 2017).

Regarding English language learning and assessment, a clear interest in foreign language teaching and learning was apparent in the mid-1990s when EFL was introduced in the 3rd grade of primary school, and it was, and still is, systematically tested as one of the main courses in entrance examinations starting from primary to tertiary education. In particular, all high-school students take the National College Entrance Examination (Gaokao), a high-stakes exam that includes English and partially dictates university entrance (Hamp-Lyons 2016). Similar to Finnish university practices, Chinese university students need to be competent in foreign languages. Thus, many Chinese undergraduate students undertake the College English Test (CET) in order to prove their language proficiency upon their graduation. For example, in 2017, nearly 10 million people participated in the CET exams. The CET test follows the College English Teaching Syllabus (CETS), which recently has also included formative assessment through self-and peer-assessment, as part of its requirements (Chen et al. 2014). The CETS was already drafted in 1962, but due to the Cultural Revolution (1966–1976) and the events that followed, its reformation and implementation started in 1980, focusing mainly on grammatical teaching and the four language skills (reading, listening, speaking, and writing). The CET test is administered by the National College English Testing Committee of the Higher Education Department, Ministry of Education (MoE). The exam is divided into three standardised tests (CET-4, CET-5, and CET-6) focusing on listening and reading comprehension (35% each of the overall score), cloze or error correction (10% of the overall score), and writing and translation (20% of the overall score) (Zheng and Cheng 2008). Both the National College Entrance Examination and CET are high-stakes national language exams that affect the future plans of Chinese undergraduate students and are indicators of a successful academic and professional career.

2. Methods

This study belongs to a larger research project that aims to provide an insight on the assessment culture of the students participating in the ESP course as part of the first-ever Sino–Finnish double-degree programme, which was established as a collaboration between a Finnish and a Chinese university. The sample consisted of 293 students participating in the ESP course offered in Finland and China. The course was for first-year students pursuing an engineering degree, and the target CEFR level of the course was B2–C1. The participants of this study consisted of more than 95% of the Finnish and Chinese student population taking part in the engineering degrees of that year. Of these, 91 (33.2%) were Chinese, and 202 (69.1%) were Finnish students. Additionally, of the overall sample, 197 (67.6%) participants were males, while 95 (32.4%) were females. The mean years of studying

English were 9.9 years for the Chinese group, while the Finnish group reported a slightly higher but still similar mean (10.2 years). Finally, the vast majority of all participants studied English at a public school (*n* = 279, 95.5%). However, it seems that apart from schooling, one-third of Chinese (*n* = 31, 34%) had also used another form of tutoring to advance their language skills compared to Finns (*n* = 8, 4%). It is worth noting that students could select more than one option under the category "opportunities to study English." The differences in the Chinese and Finnish samples were due to the different admission results and annual intake of the two universities. Bearing in mind the two distinct student groups and assessment traditions, the following research questions were investigated:

1. How did Chinese and Finnish students perceive assessment prior to the ESP course?
2. What are the similarities and differences between Chinese and Finnish students' perceptions regarding the role of the teacher in the assessment process?
3. What are the similarities and differences between Chinese and Finnish students' perceptions regarding their role in the assessment process?

In order to address these research questions, a mixed-method approach was employed. Since the focus was on the entire course population, a survey was given to all course participants. Additionally, open-ended questions were included at the beginning of the survey in order to collect qualitative data and enhance our understanding of potential differences between students. As Creswell and Clark (2018) point out, there are various reasons for a researcher to utilize mixed-methods research, particularly to secure more complete data results, since often the results might be more complex and even contradictory between different methods. Bryman (2012) also emphasizes that quantitative results could show various relationships between variables, while the qualitative part of the research could explain these relationships more sufficiently to enhance completeness and to address different research questions. Finally, the underlying philosophical paradigm of mixed methods, pragmatism, is in line with this type of large research project. Specifically, the core aspect of pragmatism, which emphasizes that "knowledge is action," and the emphasis of interpretating a phenomenon based on the practical consequences (Brinkmann 2017) is in accordance with this research project, where the students' interpretations of assessment were examined before and after the ESP course, in order to investigate potential changes. According to Tashakkori and Teddlie (2003), mixed-methods categories could be divided in the level of methods by combining qualitative and quantitative methods and in the level of sequence based on the various stages of the research. In this study, the former category was used through an online self-report including quantitative and qualitative data-analysis methods where students' pre-existing perceptions of assessment were collected. In the future, the latter category will also be employed to compare these results to the data collected after the completion of the ESP course.

2.1. Instrument

In the first section of the online self-report, three statements were given, and students' opinions were asked:

1. Describe in a few words teacher's responsibilities in the language assessment process.
2. Describe in a few words student's responsibilities in the language assessment process.
3. What do you think about monitoring your own progress? Describe in a few words.

Additionally, 15 definitions of assessment were included, while the students were instructed to choose the ones they associated with assessment. The starting point was based on the Brown et al. (2009) definitions of assessment, while three more statements were included: the teacher gives me informal feedback; I monitor my own performance; and my classmates give me informal feedback. The reasons these statements were added in the original 12 statements were twofold:

a. The researcher considered that the dimension of Assessment for Learning (AfL) was not fully covered with the statements "My classmates score or evaluate my performance" and "The teacher observes me in class and judges my learning." The

use of words such as "score" and "judge" indicated a closer relation to grades and summative assessment, and even though evaluation could imply the use of informal feedback, the researcher considered that the word was not explicit enough. Thus, the phrase "give informal feedback" was considered closer to the notions of interactive–informal practices through scaffolding and assessment as part of students' active participation in their learning (Rivers and Lomotey 1996).

b. The students would also engage in these forms of assessment (self-assessment, peer-assessment, and teachers' feedback) during their course, so it would be beneficial to investigate in advance if they actively associated them with the assessment.

The primary data instrument used in this research study was an adapted version for Students' Conceptions of Assessment-V inventory (SCoA-V) for tertiary education. Various versions of the inventory have been used to investigate students' conceptions of assessment (Brown and Harris 2012; Brown et al. 2009; Brown and Hirschfeld 2008; Brown and Wang 2013). The inventory was given in English, and the students filled it out electronically prior to the first lesson of the ESP course using their university student number and password. SCoA-V responses were given on a six-point scale, and 45 statements were included based on four main conceptions of assessment: assessment improves learning, students' accountability and assessment, assessment is irrelevant, and assessment is liked (Brown et al. 2009; Weekers et al. 2009). Context-specific adaptations included words related to tertiary education, such as the word "school," which became "university."

2.2. Data Analysis and Synthesis: Methods

For the analysis of the collected data, various methods were employed. Confirmatory factor analysis (CFA) was used for the SCOA-V inventory under the structural equation modelling (SEM) framework. Compared to exploratory factor analysis (EFA), where there are no fixed factors and instead the researcher explores the relationships among variables, in CFA, the researcher seeks to validate an a priori number of factors based on previous published research. In particular, in CFA, the questionnaire items form a scale that loads on an unobserved factor, which is also called the latent variable/factor. In this research study, the latent variables were the following: improvement, including two subfactors: the teacher's instructions and the students' self-improvement; external factors; irrelevance; and affect. Measurement invariance was employed for the analysis of the quantitative data and was conducted by the statistical tool for conducting SEM, Mplus. Measurement invariance focuses on whether the meaning of the scores in a construct changes based on, for example, population, time of measurement, etc. (Meade and Lautenschlager 2004). There are four types of measurement invariance ranging from configural to weak to strong to strict invariance. Each type shows a stricter and more restrictive hypothesis about the invariance (Wu et al. 2007). In particular, when the hypothesis of configural invariance is accepted, it means that the same factors are established in both groups; however, the factor loadings of the observed variables are somewhat different in each group. Therefore, means cannot be directly compared, and, instead, the focus would be on how the loadings differ. A weak factorial model or metric invariance presents similar factor loadings since they are constrained. However, the means are allowed to be different between groups. So, because the loadings are the same, mean differences can be examined. A strong configural model, scalar invariance, assumes a weak invariance hypothesis. That means that both the factor loadings and means are similar between the two groups. Finally, a strict configural model assumes the strong invariance hypothesis and further restricts the variances to be equal between groups (Kline 2016).

Thematic analysis was utilized for the open-ended questions due to the flexibility this qualitative method offers in order to create emerging categories from the raw data. Thus, descriptive codes for each group of students were created separately. Next, based on these codes, the answers were grouped, and the categories were refined to create analytic codes. Afterwards, each theme was cross-referenced between the two groups of students in order to unveil possible similarities and differences and to reveal potential patterns. Finally, two

graphs were created: one for each group based on each research question and another spiderweb diagram comparing the emerged themes of the groups.

3. Results

Initially, the configural model showed poor fit, indicating that the full factor structure was not the same for both groups (CFI = 0.668, χ^2 = 3726.714, df = 1870). Even when certain modifications were performed, no acceptable model fit was achieved. This indicates that the factor structures for these two groups were different. To investigate those differences, configural invariances were conducted between groups for each factor separately, which allowed us to achieve our primary purpose to investigate how different aspects of assessment behave between the two groups. To determine the level of invariance between the groups, we compared nested models for each factor using a Chi-square difference test. For each factor, we first fit the configural model and then compared it with the weak, then the strong, and then the strict model, stopping the comparison when the difference between the currently held model and the next model was determined to be significant. This was determined by a "Chi-square difference test" to assess whether each model differed from the subsequent nested model.

In Table 1, the model that was determined (at alpha= 0.05) to best represent the level of invariance between the two groups was underlined as the acceptable model fit for configural invariance. According to Kline (2016), chi-square should be evaluated and presented along with other fit indices such as the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) to avoid possible sensitivity of fit indices with regard to the size of the sample, model complexity, etc. In this study, the use of CFI was selected as an additional method to assess model fit. CFI can take values from 0–1, meaning that, for example, a value of 0.9 would be a good model fit as it represents that the model is 90% accurate (Kline 2016). There is not a consensus among scholars of a possible benchmark; however, scholars agree that a value over 0.95 is a strict benchmark, while others propose that a value over 0.9 presents an adequate model (Schumacker and Lomax 2010). In this study, the goal was to achieve values over 0.9 as a first step and over 0.95 whenever possible (e.g., benefit and irrelevance latent variables). Finally, the suggested modifications for each configural invariance are also presented in the table in order to achieve the accepted model fit.

Table 1. Configural invariance modifications and accepted model fit.

Latent Variables	Model Fit	CFI	χ^2	df
Irrelevance	Configural	1	37.282	40
	Weak	0.994	49.746	47
	<u>Strong</u>	<u>0.967</u>	<u>69.464</u>	<u>55</u>
	Strict	0.763	167.779	63
Improvement: Student (Modification: IMPR8 with IMPR2)	Configural	0.958	94.062	52
	Weak	0.954	106.879	60
	<u>Strong</u>	<u>0.949</u>	<u>120.653</u>	<u>69</u>
	Strict	0.885	194.223	78
Improvement: Teacher's instructions (Modification: IMPR1 excluded; IMPRT38 with IMPRT33; IMPRT37 with IMPRT32; IMPRT37 with IMPRT14)	Configural	0.924	108.624	64
	<u>Weak</u>	<u>0.909</u>	<u>126.397</u>	<u>73</u>
	Strong	0.852	170.077	83
Benefit (Modification: B18 with B5)	Configural	0.958	75.483	38
	<u>Weak</u>	<u>0.952</u>	<u>87.961</u>	<u>45</u>
	Strong	0.894	147.970	53
External factors (Modification: EXT45 excluded; EXT34 with EXT17; EXT10 with EXT7; EXT17 with EXT4)	<u>Configural</u>	<u>0.918</u>	<u>68.504</u>	<u>34</u>
	Weak	0.885	89.157	41

Note: CFI = comparative fit index; χ^2 = chi-square test; df = degrees of freedom.

Both the quantitative and qualitative findings are presented below based on the research questions in order to avoid a potential limitation of a mixed-methods design, e.g., a scattered presentation of the results, due to the complexities of merging quantitative and qualitative types of data (Creswell and Clark 2018). Thus, regarding the quantitative part of the research, the factors "irrelevance," "benefit," and "external factors" are presented in the first research question, while the "improvement: teacher's instructions" and "improvement: student" are used for the second and third questions, respectively.

3.1. RQ1. How Did Chinese and Finnish Students Perceive Assessment Prior to the ESP Course?

Initially, both groups were asked to select statements they would associate with assessment (Table 2). The vast majority of Finnish students picked various forms of summative assessment such as exams and written tests made by the teacher as the primary forms of assessment. The Finns also considered important teacher's feedback and written in-class assessment compared to the Chinese students. Like Finnish students, half of the Chinese students agreed with the statements "a three-hour examination" and "grading written work." However, Chinese students underscored written tests as an important form of assessment, since less than a quarter opted for these two statements (written tests made by the teacher/someone else).

Table 2. Definitions of assessment.

	Definitions of Assessment	Finns (%)	Chinese (%)
1.	An examination that takes one to three hours.	72%	48%
2.	I score or evaluate my own performance.	51%	50%
3.	I monitor my own performance.	50%	59%
4.	My classmates score or evaluate my performance.	32%	41%
5.	My classmates give me informal feedback.	35%	31%
6.	The teacher asks me questions out loud in class.	12%	22%
7.	The teacher grades or marks or scores the written work I hand in.	77.7%	50%
8.	The teacher grades me on a written test that he or she made up.	68%	27%
9.	The teacher grades me on a written test that was written by someone other than the teacher.	35%	16%
10.	The teacher observes me in class and judges my learning.	36%	38%
11.	The teacher scores a portfolio of work I have done over the course of a term or school year.	39%	32%
12.	The teacher scores me on an in-class written essay.	52%	25%
13.	The teacher gives me informal feedback.	59%	36%
14.	The teacher scores my performance after meeting or conferencing with me about my work.	33%	48%
15.	The teacher uses a checklist to judge my in-class performance.	19%	25%

Both groups strongly valued self-assessment. It is worth noting that this was the most-selected statement for Chinese students. Additionally, nearly a third of Finnish and Chinese students selected statements related to peer-assessment. Finally, both groups seemed to agree that in-class performance and assessment criteria were not considered a significant form of assessment since only a small minority of students selected these two forms of assessment.

Regarding the irrelevance latent variable, the strong factorial model was considered as the acceptable one showing that both factor loadings and means were similar between the two groups (Table 3). Strict invariance was not achieved. This indicates a difference

in the variances of the items, which was expected given the difference in group size. Given the strong invariance, both groups showed disagreement with the general idea that assessment is irrelevant to their learning since both groups presented low means of the observed variables.

Table 3. Irrelevance latent variable—strong factorial model.

	Chinese Group					Finnish Group			
IR	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value	IR	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value
IR6	0.758	0.053	14.292	0.000	IR6	0.671	0.044	15.411	0.000
IR12	0.556	0.055	10.046	0.000	IR12	0.620	0.054	11.383	0.000
IR19	0.360	0.078	4.616	0.000	IR19	0.266	0.055	4.844	0.000
IR25	0.650	0.063	10.338	0.000	IR25	0.500	0.049	10.215	0.000
IR31	0.578	0.061	9.476	0.000	IR31	0.507	0.054	9.474	0.000
IR36	0.749	0.052	14.458	0.000	IR36	0.620	0.046	13.591	0.000
IR40	0.707	0.054	13.031	0.000	IR40	0.603	0.049	12.404	0.000
IR44	0.540	0.067	7.996	0.000	IR44	0.382	0.053	7.245	0.000
				Means -Intercepts					
IR	Mean	S.E.	Est./ S.E.	Two-Tailed *p*-value	IR	Mean	SE.	Est./ SE.	Two-Tailed *p*-value
IR6	1.493	0.112	13.336	0.000	IR6	2.269	0.125	18.224	0.000
IR12	1.330	0.102	13.023	0.000	IR12	2.547	0.132	19.225	0.000
IR19	1.916	0.144	13.312	0.000	IR19	2.437	0.134	18.163	0.000
IR25	2.422	0.171	14.155	0.000	IR25	3.197	0.163	19.661	0.000
IR31	2.304	0.172	13.425	0.000	IR31	3.474	0.169	20.525	0.000
IR36	1.652	0.123	13.415	0.000	IR36	2.349	0.127	18.431	0.000
IR40	1.381	0.106	13.025	0.000	IR40	2.020	0.113	17.845	0.000
IR44	2.220	0.162	13.704	0.000	IR44	2.700	0.145	18.596	0.000
				Residual Variances					
IR	Residual variance	S.E.	Est./ S.E.	Two-Tailed *p*-value	IR	Residual Variance	S.E.	Est./ S.E.	Two-Tailed *p*-value
IR6	0.425	0.080	5.284	0.000	IR6	0.550	0.058	9.399	0.000
IR12	0.691	0.061	11.240	0.000	IR12	0.616	0.068	9.120	0.000
IR19	0.871	0.056	15.546	0.000	IR19	0.929	0.029	31.714	0.000
IR25	0.578	0.082	7.070	0.000	IR25	0.750	0.049	15.359	0.000
IR31	0.666	0.070	9.464	0.000	IR31	0.743	0.054	13.677	0.000
IR36	0.439	0.078	5.667	0.000	IR36	0.615	0.057	10.873	0.000
IR40	0.499	0.077	6.502	0.000	IR40	0.637	0.059	10.866	0.000
IR44	0.709	0.073	9.736	0.000	IR44	0.854	0.040	21.181	0.000

Next, for the "affect/benefit" latent variable (Table 4), the modification of B5 "Assessment encourages my class to work together and help each other" correlated with B18 "Assessment motivates me and my classmates to help each other," which occurred in order to achieve good model fit.

The weak factorial model was accepted, indicating that the factor loadings of observed variables did not differ between the two groups of students. However, a mean difference was observed. Specifically, the Chinese group of students presented quite high mean scores in B11 "Assessment is an engaging and enjoyable experience for me," as well as B5 and B18 compared to the Finnish group. It should also be pointed out that, for B43, "I find myself really enjoying learning when I am assessed," the factor loading was low, and the residual variance was very high in both groups, implying that while this measured variable was significant, it is not a strong indicator for this latent factor. Overall, there were differences among means between the groups, with the Chinese group showing more overall agreement with the benefit factor.

As far as the "assessment and external factors" latent variable is concerned, the original model was modified to achieve a good fit—initially, EXT45. "Assessment tells

my parents how much I've learnt" was excluded since this observed variable was not loading for either of the groups. Moreover, EXT34 "Assessment measures the worth or quality of schools" was allowed to correlate with EXT17 "Assessment provides information on how well universities are doing," EXT10 "Assessment results tell future employers how good I am" with EXT7 "Assessment results show how intelligent I am," and, finally, EXT17 "Assessment provides information on how well universities are doing" with EXT4 "Assessment keeps universities honest and up-to-scratch."

Table 4. Benefit latent variable—weak factorial model.

	Chinese Group					**Finnish Group**			
B	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value	B	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value
B5	0.635	0.052	12.130	0.000	B5	0.635	0.046	13.854	0.000
B11	0.626	0.059	10.589	0.000	B11	0.579	0.045	12.822	0.000
B18	0.698	0.050	13.832	0.000	B18	0.685	0.041	16.702	0.000
B24	0.717	0.049	14.733	0.000	B24	0.737	0.036	20.244	0.000
B30	0.703	0.057	12.230	0.000	B30	0.671	0.039	17.394	0.000
B35	0.718	0.048	14.950	0.000	B35	0.768	0.034	22.330	0.000
B39	0.762	0.044	17.129	0.000	B39	0.768	0.034	22.330	0.000
B43	0.342	0.055	6.177	0.000	B43	0.420	0.058	7.181	0.000
				Means -Intercepts					
B	Mean	S.E.	Est./ S.E.	Two-Tailed *p*-value	B	Mean	SE.	Est./ SE.	Two-Tailed *p*-value
B5	4.035	0.290	13.899	0.000	B5	3.599	0.180	20.031	0.000
B11	3.597	0.251	14.336	0.000	B11	2.918	0.156	18.646	0.000
B18	4.138	0.291	14.220	0.000	B18	3.769	0.188	19.999	0.000
B24	3.685	0.263	14.035	0.000	B24	3.449	0.177	19.446	0.000
B30	3.897	0.261	14.935	0.000	B30	3.650	0.194	18.794	0.000
B35	3.947	0.277	14.262	0.000	B35	3.794	0.194	19.547	0.000
B39	4.219	0.303	13.947	0.000	B39	4.083	0.205	19.946	0.000
B43	2.419	0.199	12.163	0.000	B43	3.048	0.162	18.804	0.000
				Residual Variances					
B	Residual Variance	S.E.	Est./ S.E.	Two-Tailed *p*-value	B	Residual Variance	SE.	Est./ SE.	Two-Tailed *p*-value
B5	0.596	0.067	8.953	0.000	B5	0.597	0.058	10.278	0.000
B11	0.608	0.074	8.205	0.000	B11	0.665	0.052	12.710	0.000
B18	0.512	0.071	7.263	0.000	B18	0.531	0.056	9.453	0.000
B24	0.487	0.070	6.981	0.000	B24	0.457	0.054	8.506	0.000
B30	0.506	0.081	6.252	0.000	B30	0.550	0.052	10.642	0.000
B35	0.485	0.069	7.028	0.000	B35	0.411	0.053	7.785	0.000
B39	0.420	0.068	6.191	0.000	B39	0.410	0.053	7.758	0.000
B43	0.883	0.038	23.299	0.000	B43	0.824	0.049	16.780	0.000

A weak invariance was not achieved, and, as a result, configural invariance was accepted (Table 2). This means that some of the observed variables did not have similar factor loadings in both groups, and variables that hold relevance for one of the student groups may not hold relevance for the other. Additionally, it shows that the external factor for each group is fundamentally different and cannot be compared overall. Therefore, we examined the individual factor loadings to determine if there is a pattern in the differences between groups. As it is shown in Table 5, the highest factor loadings for the Chinese students were the observed variables EXT23 "Assessment results predict my future performance," EXT29 "Assessment is important for my future career or job," and EXT34 "Assessment measures the worth or quality of schools," signifying that, for these students, assessment practices affect both their future and university quality. Similarly, statements related to the relationship between assessment and its effect on their future aspirations (EXT23 and EXT29) and employability (EXT10) had a high factor loading in the Finnish group's popu-

lation. However, observable variables related to assessment and the quality of universities (EXT4, EXT17, and EXT34) presented the lowest factor loadings. It is worth noting that EXT22 "Teachers assess me so they can write reports for my parents" showed one of the lowest factor loadings for the Finnish students compared to the Chinese group in which the observed variable presented a better factor loading. Overall, for the Chinese group, the external factor included high loadings on items related to their future achievements, quality of schools, and, to some extent, parents' expectations. In comparison, the Finnish group related only their future aspirations to external factors as it is presented in this configural model of CFA.

Table 5. External factors—configural factorial model.

	Chinese Group					Finnish Group			
EXT	Stand. Factor loading	SE.	Est./ SE.	Two-Tailed *p*-value	EXT	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value
EXT4	0.518	0.093	5.566	0.000	EXT4	0.229	0.085	2.698	0.007
EXT7	0.341	0.105	3.238	0.001	EXT7	0.510	0.080	6.362	0.000
EXT10	0.427	0.098	4.368	0.000	EXT10	0.655	0.072	9.046	0.000
EXT17	0.597	0.087	6.878	0.000	EXT17	0.342	0.079	4.327	0.000
EXT22	0.497	0.092	5.422	0.000	EXT22	0.233	0.088	2.659	0.008
EXT23	0.776	0.063	12.314	0.000	EXT23	0.646	0.071	9.127	0.000
EXT29	0.700	0.071	9.809	0.000	EXT29	0.500	0.075	6.678	0.000
EXT34	0.710	0.068	10.366	0.000	EXT34	0.372	0.077	4.837	0.000
				Means -Intercepts					
EXT	Mean	S.E.	Est./ S.E.	Two-Tailed *p*-value	EXT	Mean	SE.	Est./ SE.	Two-Tailed *p*-value
EXT4	3.811	0.301	12.648	0.000	EXT4	4.969	0.258	19.284	0.000
EXT7	1.868	0.174	10.756	0.000	EXT7	2.447	0.141	17.360	0.000
EXT10	2.670	0.224	11.922	0.000	EXT10	2.768	0.155	17.854	0.000
EXT17	3.658	0.291	12.584	0.000	EXT17	4.336	0.226	19.184	0.000
EXT22	2.422	0.208	11.651	0.000	EXT22	1.820	0.115	15.831	0.000
EXT23	2.451	0.210	11.685	0.000	EXT23	2.667	0.151	17.714	0.000
EXT29	3.397	0.273	12.454	0.000	EXT29	3.462	0.187	18.561	0.000
EXT34	2.817	0.234	12.057	0.000	EXT34	3.321	0.180	18.447	0.000
				Residual Variances					
EXT	Residual Variance	S.E.	Est./ S.E.	Two-Tailed *p*-value	EXT	Residual Variance	SE.	Est./ SE.	Two-Tailed *p*-value
EXT4	0.732	0.096	7.586	0.000	EXT4	0.948	0.039	24.421	0.000
EXT7	0.883	0.072	2.275	0.000	EXT7	0.740	0.082	9.073	0.000
EXT10	0.818	0.084	9.785	0.000	EXT10	0.571	0.095	6.027	0.000
EXT17	0.644	0.104	6.217	0.000	EXT17	0.883	0.054	16.347	0.000
EXT22	0.753	0.091	8.247	0.000	EXT22	0.946	0.041	23.121	0.000
EXT23	0.398	0.098	4.078	0.000	EXT23	0.582	0.092	6.364	0.000
EXT29	0.510	0.100	5.099	0.000	EXT29	0.750	0.075	10.027	0.000
EXT34	0.496	0.097	5.103	0.000	EXT34	0.862	0.057	15.077	0.000

3.2. RQ2. What Are the Similarities and Differences between Chinese and Finnish Students' Perceptions Regarding the Role of the Teacher in the Assessment Process?

Certain modifications were indicated concerning the "assessment and teacher's instructions" latent variable (Table 6) since the original model did not achieve a good model fit. Initially, IMPRT1 "Assessment is checking off my progress against achievement objectives" was excluded and added a correlation of IMPRT38 "Teachers use my assessment results to see what they need to teach me next" with IMPRT33 "My teachers use assessment to help me improve," IMPRT37 "Assessment is comparing my work against set criteria" with IMPRT32 "Assessment is assigning a grade or level to my work," and IMPRT37 with IMPRT14.

Table 6. Improvement/teacher's instructions—weak factorial model.

	Chinese Group					Finnish Group			
IMPR	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value	IMPR	Stand.factor loading	SE.	Est./ SE.	Two-Tailed *p*-value
IMPRT3	0.363	0.090	4.019	0.000	IMPRT3	0.236	0.053	4.418	0.000
IMPRT9	0.704	0.056	12.628	0.000	IMPRT9	0.636	0.049	12.879	0.000
IMPRT13	0.609	0.060	10.179	0.000	IMPRT13	0.568	0.055	10.331	0.000
IMPRT14	0.502	0.069	7.243	0.000	IMPRT14	0.440	0.056	7.902	0.000
IMPRT16	0.758	0.052	14.494	0.000	IMPRT16	0.698	0.046	15.150	0.000
IMPRT32	0.512	0.064	8.043	0.000	IMPRT32	0.418	0.059	7.121	0.000
IMPRT33	0.785	0.053	14.705	0.000	IMPRT33	0.453	0.046	9.836	0.000
IMPRT37	0.321	0.075	4.280	0.000	IMPRT37	0.266	0.060	4.418	0.000
IMPRT38	0.608	0.074	8.246	0.000	IMPRT38	0.342	0.048	7.185	0.000
IMPRT42	0.319	0.080	3.987	0.000	IMPRT42	0.212	0.055	3.855	0.000
				Means -Intercepts					
IMPR	Mean	S.E.	Est./ S.E.	Two-Tailed *p*-value	IMPR	Mean	SE.	Est./ SE.	Two-Tailed *p*-value
IMPRT3	2.641	0.207	12.772	0.000	IMPRT3	2.894	0.163	17.714	0.000
IMPRT9	3.867	0.285	13.582	0.000	IMPRT9	6.503	0.314	20.714	0.000
IMPRT13	3.640	0.274	13.281	0.000	IMPRT13	5.872	0.281	20.888	0.000
IMPRT14	3.355	0.246	13.611	0.000	IMPRT14	5.393	0.270	19.965	0.000
IMPRT16	3.693	0.273	13.525	0.000	IMPRT16	6.168	0.305	20.242	0.000
IMPRT32	3.617	0.284	12.733	0.000	IMPRT32	5.348	0.257	20.810	0.000
IMPRT33	5.186	0.383	13.555	0.000	IMPRT33	4.925	0.240	20.530	0.000
IMPRT37	3.062	0.239	12.816	0.000	IMPRT37	4.738	0.238	19.915	0.000
IMPRT38	4.588	0.330	13.906	0.000	IMPRT38	3.973	0.204	19.444	0.000
IMPRT42	2.279	0.195	11.670	0.000	IMPRT42	2.707	0.150	18.059	0.000
				Residual Variances					
IMPR	Residual Variance	S.E.	Est./ S.E.	Two-Tailed *p*-value	IMPR	Residual Variance	SE.	Est./ SE.	Two-Tailed *p*-value
IMPRT3	0.869	0.065	13.282	0.000	IMPRT3	0.944	0.025	37.504	0.000
IMPRT9	0.504	0.079	6.418	0.000	IMPRT9	0.595	0.063	9.456	0.000
IMPRT13	0.629	0.073	8.632	0.000	IMPRT13	0.677	0.062	10.836	0.000
IMPRT14	0.748	0.070	10.737	0.000	IMPRT14	0.806	0.049	16.454	0.000
IMPRT16	0.425	0.079	5.358	0.000	IMPRT16	0.513	0.064	7.975	0.000
IMPRT32	0.737	0.065	11.296	0.000	IMPRT32	0.825	0.049	16.804	0.000
IMPRT33	0.383	0.084	4.572	0.000	IMPRT33	0.795	0.042	19.044	0.000
IMPRT37	0.897	0.048	18.628	0.000	IMPRT37	0.929	0.032	29.030	0.000
IMPRT38	0.630	0.090	7.022	0.000	IMPRT38	0.883	0.033	27.162	0.000
IMPRT42	0.898	0.051	17.548	0.000	IMPRT42	0.955	0.023	40.891	0.000

A weak factorial model was retained, assuming configural invariance. This finding indicates that similar group loadings appeared in both groups; however, there was a mean difference between the two groups. The Finnish group seems to have agreed more with the statements of this factor compared to the Chinese group. Noteworthily, for both groups, and especially for the Finnish group, IMPR42 "Assessment results tell teachers how well I'm doing," IMPR37, and IMPR3 "Teachers use assessment results to put us into learning groups" did not present strong factor loadings. This was also evident due to high residual variances in these measured variables. In conclusion, the different means emerging from the weak factorial model indicate that the Finnish group was still more positive than the Chinese group regarding how assessment could inform teacher's practices.

Both groups considered fairness as one of the main themes with regard to the role of the teacher. In particular, a little less than half of the Chinese students mentioned fairness in their responses, while some of them associated this concept with seriousness, justice, and objectivity. Even though many mentioned these words as qualities without providing any explanation, a few wrote, for example, that "I think teachers should be

responsible and fair to every student"; "I think it is most important to be fair and impartial as a teacher and to evaluate carefully"; and "I think teachers should be fair and objective. And try to understand why students make such an answer," showing that fairness is related with avoiding biases. Fairness was prevalent for the Finnish group of students since almost half of their responses focused on this theme too. They seemed to agree with the quantitative part of this study as they mentioned that assessment informs the teacher's practice (formative assessment practice), e.g., "I think teacher's assessment should be fair and objectively done. After all, assessment plays a significant role in students' motivation and learning. Assessment should help the teacher to obtain better awareness into what students understand in order to prepare instruction." However, the Finnish students linked fairness with equality through diverse forms of assessment to address the classroom's needs as it shown in these examples: "teacher's responsibilities are to assess everyone fairly and in the same way and as widely as possible" and "teacher must be as equal as possible to all the students. Teacher also needs to make sure that there are enough many ways to students to prove that they have studied what needs to be studied in this course." Finally, students considered that criteria should be well-communicated in advance and systematic as well as considering that assessment should focus on various skills and types of assessment. Two representative examples were the following: " … especially when the teachers are assessing students, they should be fair, and their assessing style should be equal for each and everyone, and they should make clear for the students in the beginning of the class what are the criteria and how the course is assessed" and "teachers should evaluate students equally. I also hope that oral skills, grammar, writing skills etc. would all be noticed when evaluating students."

Another theme for both groups was the supportive role of the teacher. Chinese students reported that the teacher should help students, for example, by correcting their mistakes and should utilise teaching skills in order for students to overcome potential issues as well as by providing guidance. For example, comments such as "help us to find our problems" or "helping the students to find out the mistakes" were common among their answers, while another one wrote "the teacher is responsible for guiding the students to make some correct evaluation of themselves." Finnish students also associated support with the concept of feedback, often positive and/or constructive. Some characteristic examples of this theme were the following: "A teacher should help students improve by pointing out what they could do better and motivate them by giving positive feedback when appropriate" as well as "teachers' responsibilities in assessment are giving feedback to the student, not only after an exam or when returning homework but whenever there is time for it. Teachers should point out the strengths and weaknesses of students, so they know what to pay more attention to." Similar to the Chinese responses, they reported that teachers should focus on pointing out potential weaknesses, but also their responses showed a more holistic approach towards learning through motivation and systematic feedback given throughout the course.

3.3. RQ3. What Are the Similarities and Differences between Chinese and Finnish Students' Perceptions Regarding Their Role in the Assessment Process?

Regarding the "assessment improves learning" latent variable (Table 7), strict invariance was not achieved. This indicates a difference in the variances of the items, which was expected given the difference in group size. Thus, the strong invariance was accepted. In order to achieve a good model fit, the IMPR8 "My assessment results are caused by my taking responsibility for my learning" with IMPR2 "I pay attention to my assessment results in order to focus on what I could do better next time" were allowed to correlate. Overall, this factor presented strong factor loadings of the observed variables, apart from the low loading of IMPR41 "Assessment shows whether I can analyse and think critically about a topic." Both groups presented similar means and seemed to similarly agree with the statements of this factor as can be seen by the strong invariance of the latent variable.

Table 7. Improvement/students—strong factorial model.

	Chinese Group					Finnish Group			
IMPR	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value	IMPR	Stand. factor loading	SE.	Est./ SE.	Two-Tailed *p*-value
IMPR2	0.587	0.059	9.923	0.000	IMPR2	0.530	0.046	1.458	0.000
IMPR8	0.612	0.055	11.046	0.000	IMPR8	0.600	0.044	13.671	0.000
IMPR15	0.772	0.041	19.062	0.000	IMPR15	0.777	0.032	24.027	0.000
IMPR20	0.851	0.032	26.623	0.000	IMPR20	0.809	0.029	28.064	0.000
IMPR21	0.782	0.041	19.060	0.000	IMPR21	0.805	0.029	27.689	0.000
IMPR26	0.743	0.049	15.222	0.000	IMPR26	0.587	0.041	14.355	0.000
IMPR27	0.838	0.034	24.905	0.000	IMPR27	0.699	0.037	18.882	0.000
IMPR28	0.494	0.066	7.489	0.000	IMPR28	0.445	0.050	8.903	0.000
IMPR41	0.177	0.065	2.711	0.007	IMPR41	0.173	0.060	2.884	0.004
Means -Intercepts									
IMPR	Mean	S.E.	Est./ S.E.	Two-Tailed *p*-value	IMPR	Mean	SE.	Est./ SE.	Two-Tailed *p*-value
IMPR2	3.804	0.246	15.485	0.000	IMPR2	4.927	0.242	20.322	0.000
IMPR8	3.943	0.252	15.637	0.000	IMPR8	5.540	0.268	20.700	0.000
IMPR15	4.363	0.294	14.856	0.000	IMPR15	6.291	0.299	21.048	0.000
IMPR20	4.311	0.304	14.173	0.000	IMPR20	5.868	0.280	20.944	0.000
IMPR21	3.940	0.260	15.170	0.000	IMPR21	5.811	0.283	20.532	0.000
IMPR26	5.271	0.349	15.090	0.000	IMPR26	5.972	0.284	21.048	0.000
IMPR27	4.375	0.320	13.682	0.000	IMPR27	5.228	0.235	22.243	0.000
IMPR28	3.433	0.221	15.502	0.000	IMPR28	4.425	0.230	19.229	0.000
IMPR41	2.181	0.162	13.497	0.000	IMPR41	3.062	0.169	18.112	0.000
Residual Variances									
IMPR	Residual Variance	S.E.	Est./ S.E.	Two-Tailed *p*-value	IMPR	Residual Variance	SE.	Est./ SE.	Two-Tailed *p*-value
IMPR2	0.656	0.069	9.458	0.000	IMPR2	0.719	0.049	14.638	0.000
IMPR8	0.625	0.068	9.219	0.000	IMPR8	0.639	0.053	12.123	0.000
IMPR15	0.403	0.063	6.441	0.000	IMPR15	0.395	0.050	7.860	0.000
IMPR20	0.275	0.054	5.048	0.000	IMPR20	0.346	0.047	7.408	0.000
IMPR21	0.389	0.064	6.074	0.000	IMPR21	0.353	0.047	7.539	0.000
IMPR26	0.449	0.072	6.190	0.000	IMPR26	0.655	0.048	13.624	0.000
IMPR27	0.297	0.056	5.259	0.000	IMPR27	0.511	0.052	9.858	0.000
IMPR28	0.756	0.065	11.572	0.000	IMPR28	0.802	0.044	18.036	0.000
IMPR41	0.969	0.023	41.998	0.000	IMPR41	0.970	0.021	46.593	0.000

Over half of the Chinese and Finnish students agreed that performance is the main responsibility of the student in assessment. The Chinese students also tended to use more adjectives to portray their role and emphasised the guidance received by their teachers. For example, many of the students used the adjectives "honest," "hard working"/"work hard," and "serious" to describe themselves when discussing their own role. These adjectives appeared in more than half of their answers. In most of their responses, they just included only these adjectives, portraying the qualities of a student instead of their responsibilities. Some of the more elaborate Chinese responses mentioned that "for one thing, students must be honest. For another, students need to try them best to fulfil a task" and "I believe students should be honest in exam. And work hard in daily study and think more after classroom learning." The Finns also expressed that performance could be accomplished through meeting the academic criteria and tasks set by the teacher, while, for this group of students, performance was also related to punctuality. However, Finnish students tended to focus more on actions instead of personality traits and attitudes by describing certain tasks they had to complete. For example, one student mentioned "participation in classes, homework and ability to speak and understand English. all that are student's responsibilities in the assessment process" while another said, "student is responsible for their progress in the course, returning the assignments and attending the lectures."

Learner Autonomy

As Little (1995) points out, "the basis of learner autonomy is that the learner accepts responsibility for his or her learning." Under this prism, the theme of learner autonomy was created for both groups as their answers rendered similar results to this definition based on their answers to the open-ended questions. For Chinese students, accepting the responsibility of their learning meant evaluating and possibly perfecting their own skills. For example, a Chinese student wrote: "Students should strengthen the ability of self-study in study and life and do a good job of preview before class and review after class"; in this line, another mentioned that they should "Take the initiative to know my weakness and treat it sincerely." Their responses were also closely associated with teachers' guidance and instructions, showing a classroom hierarchy.

In their words, students said: "The responsibility of students is to study independently, and to prepare and review independently under the guidance of teachers" as well as "Personally, I should try my best to achieve teacher's demands and finished these by my own."

In this spirit, Finnish students mentioned self-assessment as a way of taking responsibility for their own learning. For instance, a student responded "Students are the only person who can really affect the general outcome of exams and courses. From the way they prepare for the test to the way they pay attention during lectures," while someone else wrote: "Student has the biggest role in the learning process. Nobody can learn for you." Another representative example was "Students have the responsibility to make sure that they put enough work that they have a good understanding of their coursework, so they have to assess have much time and effort they put to completing that course. Of course, because of that, they have the best understanding of how well they handle the material so if they are truthful, their assessment is the most accurate." Contrary to Chinese responses, Finnish students focused on a more collaborative relationship among themselves, their classmates, and the teacher in assessment. For instance, a student wrote "I think teacher must cooperate with the student and consider the level of the whole group" as well as "Student should monitor their own progress. Also, it's the student responsibility to provide something for the teacher to evaluate. Assessment process should be balanced cooperation between student and teacher." Furthermore, this collaboration involved the assessment procedures in the class, e.g., "Perhaps the teacher could ask students' opinion about the assessment process and then develop it" as well as stating their disagreement regarding assessment, as a student noted: "mostly it is teacher's responsibility, but student can rate themselves as well. If they are not in the same page as teacher, they have to contact teacher and discuss about that."

Apart from the comparison of themes between the two groups, it is worth mentioning that, for Finnish students, feedback, as a theme, was apparent across the other, emerged themes as it appeared both in the "learning autonomy" and "performance" themes. For Finns, feedback was part of the assessment as their answers indicated that receiving and giving feedback to the teacher as well as peer-feedback was linked to learner autonomy and their performance. Some representative examples are the following:

1. "Students should listen and do their best to use given feedback. They should also pay attention to things they find difficult and compare it to feedback given by teachers."
2. "Student has to take feedback as an opportunity to learn more and make itself better."
3. "Students should also give their feedback to the teacher. What felt useful and what felt like it served little purpose, maybe there would have been a better way to do something?"
4. "Complete tasks and exercises which teacher has given and also give the teacher feedback about his/her teaching skills and lessons."

4. Discussion

This study investigated students' perceptions of assessment in two different university settings in Finland and China. These students will participate in the same ESP course as

part of their undergraduate studies. Thus, their previous experiences and educational traditions could affect their attitudes on language assessment. Additionally, there is a practical need for this type of empirical research since these findings could be used as a starting point in order to develop further future Sino–Finnish courses/degrees, a new emerging educational export market in Finland. As a plethora of studies acknowledge, in order to adapt language assessment practices to the local contexts, the unique characteristics of languages, cultures, disciplines, and institutions, among others, should be taken into consideration (Chalhoub-Deville 2019; Jenkins and Leung 2019; Gutiérrez Eugenio and Saville 2017).

One of the main findings of the study was that the overall structure of the whole questionnaire did not fit this student population. This was mainly due to the configural invariance of the "external factor" latent variable, indicating that the factor structure is different and consequently cannot be directly compared between the two groups. For the Chinese students, the external factors were related to their future aspirations and the quality of the universities. Additionally, the findings show that parents seem to still play a role in the assessment process for these students. As Biggs and Watkins (2001) found, Chinese students' motivation was influenced by external factors such as rewards, parents, etc. Similarly, this finding has been shown in other studies where Chinese university students reported that parent expectations were important to them (Peterson et al. 2013; Wang and Brown 2014). It is also worth noting that English language learning has been associated with students' beliefs of upward social mobility in the Chinese context (Liu et al. 2016; Pan and Block 2011). In comparison, the Finnish group related only their future aspirations to external factors and assessment as it is presented in this configural model of CFA. This could be due to the fact that Finnish students associate to some extent academic skills along with other skills with employability (Räty et al. 2019).

Regarding the "irrelevance" factor, the strong factorial model shows that both the factor loadings and means were similar between the two groups. The sampling size difference could explain the difference in the variances of the items. Both groups contradicted the general idea that assessment is irrelevant to their learning since both presented low means of the observed variables. Additionally, both groups endorsed the "benefit" latent variable; however, the weak factorial model indicated a mean difference between the groups. Still, Chinese students' mean scores were lower in the "irrelevance" factor and higher in the "benefit" factor than the Finnish. This possibly reflects the feelings of pride and enjoyment about the notion that the purpose of assessment is a way for students in China to develop further their learning, skills, and character (Chen and Brown 2018). Academic achievements based on summative assessment seem to dictate Chinese students' personal worth and self-value (Brown and Wang 2013; Xiao and Carless 2013). Thus, assessment could be considered highly relevant and beneficial for these students.

The weak invariance of the "assessment improves teacher's instructions" factor indicates that the Finnish group held a more positive stance than the Chinese regarding the effect of assessment on teacher's practices. The Finnish students especially valued formative practices of assessment as presented in the strong factorial loadings and high mean scores. The qualitative part of this study seems to agree with this finding since students considered that assessment could give an insight into students' knowledge and could subsequently improve teachers' future instructions. Additionally, Finnish students associated assessment with teacher-oriented assessment practices, their responses ranging from teacher feedback to exams made by the teacher (Table 2). This finding seems to agree with previous studies exploring Finnish students' perceptions of assessment, in which assessment was still considered traditional and primarily led by the teacher (Hildén and Härmälä 2015; Mäkipää and Ouakrim-Soivio 2020). Pollari (2016) and Tarnanen and Huhta (2008) also indicated that classroom-based assessment is apparent both in primary and secondary Finnish education. Thus, it can be assumed that students' beliefs reflect positively upon previous assessment practices in which the students have been exposed. This positive attitude towards the formative nature of assessment is considered an adaptive

conception and has been related to increased achievement in previous studies (Brown et al. 2009; Brown and Hirschfeld 2008).

Chinese students presented a positive but still a more moderate stance regarding the use of formative assessment made by the teacher compared to Finns. These students focused more on assessment, such as grading their homework; informal discussion made by the teacher; and a three-hour examination choice. However, they did not choose classroom tests made by the teacher or someone else as essential types of assessment. It is possible the classroom-based tests are not perceived as that important in a high-stakes assessment environment due to the series of externally distributed examinations. Additionally, the previously investigated Chinese teacher conceptions reported that while the teachers praised assessment as a way to improve a student's skills and character, the same conception was highly associated with high-stakes examination results for their students (Kennedy 2016). As a result, these findings raise interesting questions of how these students are going to perceive classroom-based summative assessment practices offered as part of this ESP course and if their conceptions of formative and summative assessment possibly collide with the course practices (Poole 2016).

The concept of fairness also dominated the students' answers in the open-ended questions. The Chinese answers tended to be very short and paired with few words such as seriousness, justice, and objectivity. Consequently, they did not provide much explanation on how fairness was perceived by the students, even though almost half of the students included it in their answers. From the longer statements, it seems that students mentioned interactional fairness focusing on the teacher's behaviour rather than on the assessment instrument (Wallace 2018). The Finnish students addressed the need for diversification of language assessment practices in order to assess students' skills fairly. These answers echo the concept of inclusive pedagogy as perceived in Finnish education through an attempt to meet individual student's needs (Kivirauma and Ruoho 2007). This is considered a preliminary finding, and follow-up interviews with the Chinese and Finnish students were scheduled as one of the future steps of this project.

Regarding students' responsibilities in the assessment process, both groups strongly valued self-assessment, which is apparent from the quantitative part of the research. This finding appeared both in the definitions of assessment, where self-assessment was one of the most-selected statements for Chinese and Finnish students. Additionally, a strong invariance of the "assessment improves students' learning" factor demonstrated that both groups presented similar means and a positive attitude towards this latent variable. In the qualitative part of the research, Chinese students' answers showed that guidance coming from the teacher was apparent both in teacher and student responsibilities. In that sense, self-assessment was accomplished through the teacher's guidance and their hard work. Similarly, hard work, respect for hierarchy, and obedience were important findings for Chinese learners in previous studies (Peterson et al. 2013; Biggs and Watkins 2001). Even though the focus of the open-ended form was on responsibilities, Chinese students tended to reply through adjectives, e.g., "hard-working," "honest," etc. A possible explanation could be that, for Chinese students, learning, and consequently assessment, has a moral aspect as the final goal is to give back to society. Thus, instead of focusing on responsibilities in the form of actions and tasks, they consider personal traits as fuel for their learning (Li 2005).

On the other hand, the Finnish students portraited a somewhat different picture regarding their responsibilities in the open-ended answers. Self-assessment was associated with learner autonomy and feedback. Even though feedback has been associated as teacher-led in Finland (Hildén and Härmälä 2015; Mäkipää and Ouakrim-Soivio 2020), interestingly in this study, the students did not present a hierarchical order of feedback, e.g., from the teacher to the students. Instead, their idea of feedback was circular via collaboration of all classroom entities since everyone in the classroom could provide feedback. This finding seems to agree with a few empirical studies in the Finnish context regarding students' assessment experiences. A small-scale study declared similar results for Finnish

HE students regarding feedback in the language classroom (Károly 2015). In particular, most Finnish students were positive about receiving feedback from different sources, e.g., teachers, classmates, and various forms of self-assessment practices. Moreover, the idea of classroom synergy in assessment echoes another Finnish study where EFL students, especially those who felt disempowered, expressed the need to affect assessment practices and play a more active role in assessment decisions (Pollari 2017). These differences between groups should be taken into account when teachers would like to construct or prepare students for new assessment tasks and/or give feedback, especially in the context of language learning, one of the most well-researched causes of students' anxiety.

5. Conclusions

To sum up, the discussed similarities and differences between the two student groups are used as a method to document and identify potential areas of differentiation and/or scaffolding when teachers introduce classroom assessment practices and their potential impact on students. This could provide an insight on how students understand their own identity and their role with regards to assessment culture in their local context (Chalhoub-Deville 2019). It is evident that assessment to some extent has different implications for the two groups, especially regarding the external factors and delivering feedback. Additionally, students' answers to the open-ended questions showed that assessment for the Finnish students is related to tasks and procedures, while, for the Chinese students, it is linked to their character. This could potentially affect the way students respond both to formative and summative assessment practices.

Finally, some limitations should be addressed. The questionnaire was distributed in English; however, the participants' language admission criteria to the programme were a minimum of a B2 level. Additionally, the latent variables could not be correlated since an overall model fit of the SCOA-V questionnaire was not successful. This could be due to the relatively small sample size for this type of research, even though it consisted of more than 95% of the Finnish and Chinese cohort participating in the ESP course of that year. Noteworthily, differences in sample size between the two groups affected the variances of the groups and, in particular, the comparison of variance between the groups. Additionally, the model inadmissibility could be due to the small sample sizes per group of students, as the estimations tend not to be so robust in samples of less than 400 participants (Boomsma and Hoogland 2001). Unique "local" case studies could provide fertile ground for transferability across contexts and a fruitful discussion between theory and practice. It is worth noting that Bransford and Schwartz (1999) point out two possible ways regarding transferability as a "direct application" in a new situation of what has been previously demonstrated and subsequently learnt or as a form of "preparation for future learning" through the unique circumstances of each case study. The latter concept of transferability could especially benefit and enrich future language courses of joint programmes among countries. It seems that these findings emphasise the argument of assessment as a highly contextualised phenomenon, and as Kennedy (2016) points out, any assessment changes should not just be " . . . about 'doing things differently," but, more importantly, "thinking about things differently."

References

Albert, Melissa A., and Jason J. Dahling. 2016. Learning goal orientation and locus of control interact to predict academic self-concept and academic performance in college students. *Personality and Individual Differences* 97: 245–48. [CrossRef]

Biggs, John B., and David A. Watkins. 2001. Insights into teaching the Chinese learner. In *Teaching the Chinese Learner: Psychological and Pedagogical Perspectives*. Edited by David A. Watkins and John B. Biggs. Hong Kong: University of Hong Kong, Comparative Education Research Centre, pp. 277–300.

Boomsma, Anne, and Jeffrey J. Hoogland. 2001. The Robustness of LISREL Modeling Revisited. In *Structural Equation Models: Present and Future. A Festschrift in Honor of Karl Jöreskog*. Edited by Robert Cudeck, Stephen Du Toit and Dag Sörbom. Lincolnwood: Scientific Software International, pp. 139–68.

Bransford, John D., and Daniel L. Schwartz. 1999. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education* 24: 61–100.

Brinkmann, Svend. 2017. *Philosophies of Qualitative Research*. New York: Oxford University Press. [CrossRef]

Brown, Gavin T. L. 2011. Self-regulation of assessment beliefs and attitudes: A review of the students' conceptions of assessment inventory. *Educational Psychology* 31: 731–48. [CrossRef]

Brown, Gavin T. L., and Lois Harris. 2012. Student conceptions of assessment by level of schooling: Further evidence for ecological rationality in belief systems. *Australian Journal of Educational and Developmental Psychology* 12: 46–59.

Brown, Gavin T. L., and Gerrit H. F. Hirschfeld. 2008. Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy and Practice* 15: 3–17. [CrossRef]

Brown, Gavin T. L., Earl S. Irving, Elizabeth R. Peterson, and Gerrit H. F. Hirschfeld. 2009. Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction* 19: 97–111. [CrossRef]

Brown, Gavin T. L., Reza Pishghadam, and Shaghayegh Shayesteh Sadafian. 2014. Iranian university students' conceptions of assessment: Using assessment to self-improve. *Assessment Matters* 6: 5–33. [CrossRef]

Brown, Gavin T. L., and Zhenlin Wang. 2013. Illustrating assessment: How Hong Kong university students conceive of the purposes of assessment. *Studies in Higher Education* 38: 1037–57. [CrossRef]

Bryman, Alan. 2012. *Social Research Methods*, 4th ed. Oxford: Oxford University Press.

Chalhoub-Deville, Micheline B. 2003. Second language interaction: Current perspectives and future trends. *Language Testing* 20: 369–83. [CrossRef]

Chalhoub-Deville, Micheline B. 2019. Multilingual Testing Constructs: Theoretical Foundations. *Language Assessment Quarterly* 16: 472–80. [CrossRef]

Cheng, Liying. 2008. The key to success: English language testing in China. *Language Testing* 25: 15–37. [CrossRef]

Chen, Qiuxian. 2017. Localized Representation of Formative Assessment in China: A Regional Study from a Sociocultural Perspective. *Frontiers of Education in China* 12: 75–97. [CrossRef]

Chen, Junjun, and Gavin T. L. Brown. 2018. Chinese secondary school students' conceptions of assessment and achievement emotions: Endorsed purposes lead to positive and negative feelings. *Asia Pacific Journal of Education* 38: 91–109. [CrossRef]

Chen, Qiuxian, Lyn May, Val Klenowski, and Margaret Kettle. 2014. The enactment of formative assessment in English language classrooms in two Chinese universities: Teacher and student responses. *Assessment in Education: Principles, Policy and Practice* 21: 271–85. [CrossRef]

Cheng, Liying, Yongfei Wu, and Xiaoqian Liu. 2015. Chinese university students' perceptions of assessment tasks and classroom assessment environment. *Language Testing in Asia* 5: 1–17. [CrossRef]

Creswell, John W., and Vicky L. Plano Clark. 2018. *Designing and Conducting Mixed Methods Research*, 3rd ed. London: Sage.

EDUFI. 2015. ePerusteet. opintopolku.fi. Available online: https://eperusteet.opintopolku.fi/#/fi/lukio/1372910/oppiaine/1383539 (accessed on 5 May 2021).

Finlex.fi. 2014. *Finlex Data Bank*. Available online: https://www.finlex.fi/fi/laki/kaannokset/2004/en20040794.pdf (accessed on 8 September 2021).

Gutiérrez Eugenio, Esther, and Nick Saville. 2017. The Role of Assessment in European Language Policy: A Historical Overview. Available online: http://www.meits.org/policy-papers/paper/the-role-of-assessment-in-european-language-policy-a-historical-overview (accessed on 15 August 2021).

Hamp-Lyons, Liz. 2016. Purposes of assessment. In *Handbook of Second Language Assessment*. Edited by Dina Tsagari and Jayanti Banerjee. The Hague: De Gruyter/Mouton, pp. 13–28.

Hildén, Raili, and Marita Härmälä. 2015. *Hyvästä paremmaksi—Kehittämisideoita Kielten Oppimistulosten Arviointien Osoittamiin Haasteisiin*. Helsinki: Kansallinen Koulutuksen arviointikeskus.

Hu, Bo, and Anne West. 2014. Exam-oriented education and implementation of education policy for migrant children in urban China. *Educational Studies* 41: 249–67. [CrossRef]

Huhta, Ari, and Raili Hildén. 2016. Kielitutkinnot ja muu laajamittainen kielitaidon arviointi Suomessa. *Kielitaidon Arviointitutkimus 2000-Luvun Suomessa (AFinLA)* 9: 3–26.

Jenkins, Jennifer, and Constant Leung. 2019. From mythical 'standard' to standard reality: The need for alternatives to standardised English language tests. *Language Teaching* 52: 86–110. [CrossRef]

Kaarninen, Mervi, and Pekka Kaarninen. 2002. *Sivistyksen Portti: Ylioppilastutkinnon Historia*. Helsinki: Otava.

Károly, Adrienn. 2015. Feedback on individual academic presentations: Exploring Finnish university students' experiences and preferences. In *Voices of Pedagogical Development—Expanding, Enhancing and Exploring Higher Education Language Learning*. Edited by Juha Jalkanen, Elina Jokinen and Peppi Taalas. Dublin: Research-publishing.net. [CrossRef]

Kennedy, Kerry J. 2016. Exploring the Influence of Culture on Assessment. In *Handbook of Human and Social Conditions in Assessment*. Edited by Gavin T. L. Brown and Lois R. Harris. Abingdon: Routledge. [CrossRef]

Kirkpatrick, Michael A., Kathryn Stant, Shonta Downes, and Leatah Gaither. 2008. Perceived locus of control and academic performance: Broadening the construct's applicability. *Journal of College Student Development* 49: 486–96. [CrossRef]

Kivirauma, Joel, and Kari Ruoho. 2007. Excellence through special education? Lessons from the Finnish school reform. *International Review of Education* 53: 283–302. [CrossRef]

Kline, Rex B. 2016. *Principles and Practice of Structural Equation Modeling*, 4th ed. New York: The Guilford Press.

Ko, Kwang Hyun. 2017. A Brief History of Imperial Examination and Its Influences. *Society* 54: 272–78. [CrossRef]

Li, Jin. 2005. Mind or Virtue. *Current Direction in Psychological Science* 14: 190–94. [CrossRef]

Little, David. 1995. Learning as dialogue: The dependence of learner autonomy on teacher autonomy. *System* 23: 175–81. [CrossRef]

Liu, Na, Chih-Kai Lin, and Terrence G. Wiley. 2016. Learner Views on English and English Language Teaching in China. *International Multilingual Research Journal* 10: 137–57. [CrossRef]

Mäkipää, Toni, and Najat Ouakrim-Soivio. 2020. Finnish Upper Secondary School Students' Perceptions of Their Teachers' Assessment Practices. *Journal of Teaching and Learning* 13: 23–42. [CrossRef]

Meade, Adam W., and Gary J. Lautenschlager. 2004. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods* 7: 361–88. [CrossRef]

Official Statistics of Finland (OSF). 2019. Subject Choices of Students. Subject Choices of Completers of Upper Secondary General School Education 2019. Available online: http://www.stat.fi/til/ava/2019/01/ava_2019_01_2020-11-26_tie_001_en.html (accessed on 9 September 2021).

Pan, Lin, and David Block. 2011. English as a "global language" in China: An investigation into learners' and teachers' language beliefs. *System* 39: 391–402. [CrossRef]

Peterson, Elizabeth R., Gavin T. L. Brown, and Richard J. Hamilton. 2013. Cultural differences in tertiary students' conceptions of learning as a duty and student achievement. *International Journal of Quantitative Research in Education* 1: 167–81. [CrossRef]

Pollari, Pirjo. 2016. Daunting, reliable, important or "trivial nitpicking?" Upper secondary students' expectations and experiences of the English test in the Matriculation Examination. *AFinLA-e: Soveltavan Kielitieteen Tutkimuksia* 9: 184–211. Available online: https://journal.fi/afinla/article/view/60854 (accessed on 9 September 2021).

Pollari, Pirjo. 2017. The power of assessment: What (dis)empowers students in their EFL assessment in a Finnish upper secondary school? *Apples–Journal of Applied Language Studies* 11: 147–75. Available online: http://apples.jyu.fi/article/abstract/529 (accessed on 9 September 2021). [CrossRef]

Poole, Adam. 2016. 'Complex teaching realities' and 'deep rooted cultural traditions': Barriers to the implementation and internalisation of formative assessment in China. *Cogent Education* 3: 1–14. [CrossRef]

Räty, Hannu, Inna Kozlinska, Kati Kasanen, Päivi Siivonen, Katri Komulainen, and Ulla Hytti. 2019. Being stable and getting along with others: Perceived ability expectations and employability among Finnish university students. *Social Psychology of Education* 22: 757–73. [CrossRef]

Rivers, Shariba W., and Kofi Lomotey. 1996. Handbook of Research on Multicultural Education (book). *Journal of Education for Students Placed at Risk (JESPAR)* 1: 193–200. [CrossRef]

Saville, Nick. 2009. *Developing a Model for Investigating the Impact of Language Assessments within Educational Contexts by a Public Examination Provider*. Bedfordshire: University of Bedfordshire.

Saville, Nick, and Hanan Khalifa. 2016. The impact of language assessment. In *Handbook of Second Language Assessment*. Edited by Dina Tsagari and Jayanti Banerjee. The Hague: De Gruyter/Mouton. [CrossRef]

Schumacker, Randall E., and Richard Lomax. 2010. *A Beginner's Guide to Structural Equation Modeling*, 3rd ed. New York: Psychology Press.

Stewart, Martina A., and Linda De George-Walker. 2014. Self-handicapping, perfectionism, locus of control and self-efficacy: A path model. *Personality and Individual Differences* 66: 160–64. [CrossRef]

Tarnanen, Mirja, and Ari Huhta. 2008. Interaction of language policy and assessment in Finland. *Current Issues in Language Planning* 9: 262–81. [CrossRef]

Tashakkori, Abbas, and Charles Teddlie. 2003. *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks: Sage.

Tsagari, Dina. 2013. EFL Students' Perceptions of Assessment in Higher Education. *International Experiences in Language Testing and Assessment*, 117–44.

University of Oulu. 2020. Todistusvalinnan Tulokset Valmistuneet—Opiskelijavalinnat Etenevät Edelleen. Available online: https://www.oulu.fi/yliopisto/uutiset/todistusvalinnan-tulokset-valmistuneet-valinnat-etenevat-edelleen?fbclid=IwAR3VZu-mp_hY6JRlcuJcxkKEVacDuvifRZUzg (accessed on 9 February 2021).

Vavla, Laureta, and Rregjina Gokaj. 2013. Learner's perceptions of assessment and testing in EFL classrooms in Albania. *Mediterranean Journal of Social Sciences* 4: 509–15. [CrossRef]

Vogt, Karin, Dina Tsagari, Ildikó Csépes, Anthony Green, and Nicos Sifakis. 2020. Linking Learners' Perspectives on Language Assessment Practices to Teachers' Assessment Literacy Enhancement (TALE): Insights from Four European Countries. *Language Assessment Quarterly* 17: 410–33. [CrossRef]

Wallace, Matthew P. 2018. Fairness and justice in L2 classroom assessment: Perceptions from test takers. *Journal of Asia TEFL* 15: 1051–64. [CrossRef]

Wang, Zhenlin, and Gavin T. L. Brown. 2014. Hong Kong tertiary students' conceptions of assessment of academic ability. *Higher Education Research and Development* 33: 1063–77. [CrossRef]

Weekers, Anke M., Gavin T. L. Brown, and Bernrard P. Veldkamp. 2009. Analyzing the dimensionality of the Students' Conceptions of Assessment inventory. In *Student Perspectives on Assessment: What Students Can Tell Us about Improving School Outcomes*. Edited by Dennis M. McInerney, Gavin T. L. Brown and Gregory Arief D. Liem. Greenwich: Information Age Press, pp. 133–57.

Wu, Amery D., Zhen Li, and Bruno D. Zumbo. 2007. Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research and Evaluation* 12: 3.

Xiao, Yangyu, and David Robert Carless. 2013. Illustrating students' perceptions of English language assessment: Voices from China. *RELC Journal* 44: 319–40. [CrossRef]

Ylioppilastutkinto.fi. n.d. Structure of the Examination. Available online: https://www.ylioppilastutkinto.fi/en/matriculation-examination/the-examination/structure-of-the-examination (accessed on 26 January 2021).

Zheng, Ying, and Liying Cheng. 2008. Test review: College English Test (CET) in China. *Language Testing* 25: 408–17. [CrossRef]