A Statistical Analysis of Liberal Arts English Course Grading Practices at a Japanese National University: Towards More Equitable Grading

日本の一国立大学における教養教育英語科目成績評価状況の統計的分析：より公正な成績評価を目指して

Masashi HASHIMOTO, Dale BROWN, Lewis MURRAY, and Kana OYABU
橋本 将、デール・ブラウン、ルイス・マリー、大藪 加奈

概要

共通教育（教養教育）英語科目は、ほとんどの学生が履修する科目であり、共通教育英語科目の成績を公正なものにすることは、大学が厳正な成績評価を目指す上で重要である。しかし、多くの教員が多様な学生を同一科目名のもとに教えるという科目の性格上、成績評価の統一には困難もつきまとう。金沢大学では、2016 年度より共通シラバスによる授業や成績評価分布の予備研究、2017 年度より共通ルブリックや教員マニュアルの配付による成績評価統一に向けた努力を行ってきた。本論では、共通教育英語科目の成績評価統一に向けて、共通教育英語科目 EAP コースの全教員による全学生の成績評価のデータ 1 年半分 10,284 件を分析し、成績評価のばらつきを調べた。分析の結果、教員の所属による有意なばらつきはないが、個々の教員に由来するばらつきと学生の所属に由来する成績のばらつきがあること、教員由来のばらつきは 2016 年度と 2017 年度を比較すると EAP I で 9 ポイント減少し 6%となったことがわかった。本論は 1 年半のデータによる分析であるので、ばらつきの減少傾向を明らかにするためには、今後もデータの収集と分析を続ける必要がある。

1. Introduction

English language is a compulsory subject in most Japanese universities (Terauchi, 2017). Unlike with secondary education, however, there have not been strong governmental guidelines or requirements for university English education until recently (Lu, 2008). Even in governmental initiatives that give direction for English education such as “An Action Plan to Cultivate ‘Japanese with English Abilities’” (the Ministry of Education, Culture, Sports and Technology [MEXT], 2003), “Five Proposals and Specific Measures for Developing Proficiency in English for International Communication” (Commission on the Development of Foreign Language Proficiency, 2011), or “English Education Reform Plan corresponding to Globalization” (MEXT, 2013), school English education has been the main focus.

Nevertheless, the MEXT has started to show a more committed approach to English education in higher education, as it has initiated major globalization projects, such as Global 30 and the Top Global University Project, in which English-medium instruction and English language education are important components (Rose & McKinley, 2018; Tada, 2016). Currently, a larger-scale nationwide 5-year university reform is also nearing completion (MEXT, 2012), and education quality assurance and stakeholder accountability have come to be often discussed in Japanese tertiary education as a result. These days, the wording of curricula is often scrutinized and more emphasis on communication as well as active-learning is encouraged by the MEXT. The days when university English language education was left to universities and instructors may be fast disappearing.

One area that needs examination when one talks about education quality assurance and accountability to stakeholders is assessment and grading practices. However, it may be said that there has been "no general agreement in higher education regarding how student performances should be graded" across the board (Yorke, Bridges, & Woolf, 2000, as cited in Beenstock & Feldman, 2018). Also, as Bloxham, Boyd and Orr (2011) have pointed out, completely standardized assessment is hard to achieve even with detailed written assessment criteria when it comes to assessing social science written work. Perhaps because of such difficulties, there is, to our knowledge, no research publicly available on the grading of an entire cohort of students by their instructors as they take compulsory English classes in a Japanese university, even though almost all Japanese universities teach English to almost all students. In this paper, we report on a statistical analysis of grades given to all first-year students by instructors in liberal arts English language courses at a Japanese university over an 18-month period, in order to examine what affects student grades, and whether there are differences in grading by instructor.

2. Background

2-1. The institution and the courses

Kanazawa University is one of 86 Japanese national universities. It has about 8,000 undergraduate and 2,000 postgraduate students (Kanazawa University, 2017). In 2013, it was selected as one of the 24 Type B universities in the Top Global University project. This means that it is designated as one of the "Global Traction Universities," which are "innovative universities that lead the internationalization of Japanese Society" (Top Global University Japan, https://tgu.mext.go.jp/en/index.html). This designation and the desire for education and university reform led to the creation of the Institute of Liberal Arts and Science in April 2016, and, at the same time, the implementation of major curriculum changes. These changes included the establishment of a four-quarter system (each quarter lasting eight weeks), the streamlining of liberal arts courses to 30 subjects, and the introduction of compulsory English language programs called "Kanazawa University Global Standard Language Courses" (GS Language Courses).

The GS Language Courses are divided into two distinct sets of courses; English for Academic Purposes (EAP), and TOEIC test preparation (TOEIC Prep.), and are one of the biggest clusters of courses taught under the same name at the university. There are about 300 EAP courses (298 in the 2018 academic year) taught by over 40 instructors (42 in the 2018 academic year), and about 240 TOEIC Prep. courses (240 in the 2018 academic year) taught by about 30 instructors (29 in the 2018 academic year). Students are divided into five registration blocks based on their departments, and each registration block consists of three departments: the Arts 1 block includes the humanities, law, and international studies departments; Arts 2 includes the economics, education, and regional development departments; Sciences 1 includes the mathematics and physics, biology, and chemistry departments; Sciences 2 includes the civil, electronic, and mechanical engineering departments; and Medicine includes the medicine, pharmacy, and health science departments.

The EAP courses and TOEIC Prep. courses have very distinct characteristics. The EAP courses, which are the focus of this paper, are designed to improve students' ability to take part in courses held in English, both within and outside Kanazawa University. The EAP courses consist of lessons in paragraph writing (EAP I), public speaking (EAP II), summarizing and responding to academic texts orally and in writing (EAP III), and carrying out a mini research project culminating in the writing of a five-paragraph essay (EAP IV). All EAP classes are mixed-ability, and students are allocated to a particular class randomly within their registration block. The grading of the EAP courses is conducted solely by individual instructors.

The TOEIC Prep. courses, on the other hand, are designed to help students gain better marks in the TOEIC test. They focus on listening and reading comprehension skills, as well as test-taking skills. Students are divided into three ability-based levels and assigned to a particular class on this basis. In the TOEIC Prep. courses, 20% of a student's grade is determined by the individual instructor, and 80% by TOEIC-style tests sat by all students.

2-2. A greater need for grade standardization

The curriculum changes described above have created a greater need for grade standardization across required English courses. In addition, changes to the admissions system and in the administrative climate around the university have also increased the need for equitable grading. The influence of each of these factors is described below.

Regarding the curriculum changes, until 2016, students had greater freedom in choosing language courses because such courses were required elective subjects in humanities and social science departments, so English was just one of the languages offered to those students. Even in departments where English was compulsory, students could choose from several English courses offered in the same timetable period. Although each student was required to take four types of English language courses, namely, Writing, Reading, (Oral) Communication, and Listening, instructors individually determined the syllabus for their class and selected textbooks, so there was

greater choice for students in terms of course content. In contrast, students in the current curriculum are allocated to pre-determined classes, so cannot choose course content or instructors. The random nature of class assignment in the current system means different approaches to grading can be perceived as unfair, since if a student had been assigned to a different class, a different grade may have been received.

The second factor encouraging greater standardization of grading is that from the 2018 academic year Kanazawa University is introducing an admission system whereby 402 out of 1,726 first-year students will start university without belonging to a specific department. The students' first-year grades will in part determine which department they ultimately join. In addition, even for students whose departments are already determined at admission, their first-year grades are still used to determine which course within their department they can belong to. Consequently, the grades given for English courses have a clear impact on students and so grade standardization is necessary to allow fair treatment of all students.

Thirdly, as one of the national universities, Kanazawa University has been directly affected by the MEXT's University Reform Action Plan (MEXT, 2012). In the action plan, changes in education methods and education quality assurance are important components. The EAP curriculum, which emphasises active-learning methods, is in line with the action plan's desire for universities to promote more student-centered education. However, after examining the grading practices of English language courses as well as other courses across the university, in December 2017 Kanazawa University Education Management Board (教育企画会議) directed all departments to look into their grading practices from the point of view of education quality assurance. According to the board, diverse grading practices are still seen across the university despite the fact that the Central Council of Education (中央教育審議会) recommended the introduction of stricter grading practices (based on mutual understanding among teaching staff) in 2008 (Central Council of Education, http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/gijiroku/08103112/003/004.htm). This recommendation points out that university assessment is often based on the discretion of individual instructors, and that little organizational effort has been made to regulate grading practices. The same recommendation states that universities with a greater focus on internationalization need to introduce more accountable foreign language assessment systems. Although the recommendation and the action plan mainly focus on using external English tests, such as TOEIC and TOEFL, to accomplish this goal, particular efforts are deemed necessary for the university's English language course assessments, too, as one of the Global Traction Universities.

2-3. Promoting standardization of grading in EAP courses

Continued efforts have been made to standardize the grading of GS Language Courses, particularly with respect to the EAP courses. As EAP courses are assessed solely by individual instructors, different grading practices can affect students' overall grade more in EAP courses than in TOEIC

Prep. courses. Consequently, the EAP course management committee produced grading guidelines and rubrics for the 2016 academic year, and these are distributed in the form of an EAP Teacher's Guide, which also includes the course syllabi and guidance on teaching the courses. The committee also offers three to four seminars for instructors every year, explaining the syllabi, reporting good teaching practice, and discussing problems and difficulties with instructors.

Apart from the above approaches directed towards all instructors, the committee also examined the grades given by individual EAP instructors. Instructors who had distinctively unusual assessment practices, for instance, over 60% of students receiving S (≥90%) and A (89−89%) grades in all their classes, or a majority of students receiving Fail or C (60−69%) grades in the same course in the 2016 academic year, were asked to reexamine their grading practices for 2017 by reviewing the rubrics and the grade distribution guideline.

3. Objectives

The objective of this paper is to assess the extent to which the EAP course grading is equitable. It presents a statistical analysis of grades given for six EAP courses delivered in the 2016 and 2017 academic years. The grades examined are for all the EAP courses (EAP I to EAP IV) in the 2016 academic year, and the EAP I and EAP II courses in the 2017 academic year.

The specific questions investigated are as follows:

1. Are there differences in grading between instructors?
2. Are there differences in grading depending on the affiliation of instructors?
3. Are there differences in grading depending on the registration blocks by which the cohort of students is divided?

Differences between instructors might occur as different instructors may have different ideas of how to grade students. As Beenstock and Feldman (2018) have commented, differential grading "is ubiquitous and seems to be the norm rather than the exception" (p. 114) in higher education, and since EAP courses have run for only two years so far, a consensus on grading may not have formed sufficiently. It is important to examine and eliminate differences in grading by instructor as much as possible from the point of view of quality assurance, as well as fairness for those students who have not been allocated to departments at admission. As students cannot choose their EAP instructors, the existence of overly lenient or overly strict instructors may offend students' sense of fairness.[1]

Differences depending on affiliation may arise as EAP courses are taught by instructors with various affiliations. These have been divided here into the following three categories: instructors belonging to the Institute of Liberal Arts and Science, instructors belonging to a different department

[1] It should be mentioned, however, that in the two years during which the EAP courses have been delivered, the main complaints placed by students taking EAP courses have been about differences between instructors in the amount of homework set rather than grading.

within Kanazawa University, and part-time instructors. Differences in grading in accordance with affiliation could occur if instructors base their grading on comparisons of the performance of students in their EAP classes with the performance of students in other classes they teach, whether within Kanazawa University or elsewhere. In other words, it is possible that instructors give grades based on the relative performance of their EAP students, rather than grading students objectively against the assessment criteria. For example, in the case of part-time instructors, the students in their EAP courses may be of higher proficiency than students in classes they teach at other institutions, and so may receive higher grades than perhaps merited.

Differences depending on registration block may exist since there may be differences in the English proficiency of students in the departments which make up each block. These differences in proficiency may translate into different levels of performance in the EAP courses.

4. Method

4-1. Data

In this study, data of grades S, A, B, C, and F for EAP I, II, III and IV in 2016 and EAP I and II in 2017 were analyzed. The number of students who received a letter grade S, A, B, C, or F for these EAP courses is shown in Table 1, where students are classified based on registration block.

Table 1. Number of students who received each letter grade

	Registration block					
Course	Arts 1	Arts 2	Sciences 1	Sciences 2	Medicine	Total
2016 EAP I	394	365	272	336	380	1,747
2016 EAP II	392	367	272	336	382	1,749
2016 EAP III	372	357	258	334	343	1,664
2016 EAP IV	365	349	258	330	333	1,635
2017 EAP I	384	375	270	328	388	1,745
2017 EAP II	384	374	270	329	387	1,744

The number of instructors involved in the data is shown in Table 2. Here, instructors are divided into three groups: (1) instructors who belong to the Institute of Liberal Arts and Science at Kanazawa University (ILAS instructors); (2) instructors who belong to another institute or college at Kanazawa University (non-ILAS instructors); and (3) part-time instructors (non-KU instructors).

Table 2. Number of instructors

	Affiliation			
Course	ILAS	non-ILAS	non-KU	Total
2016 EAP I	10	3	13	26
2016 EAP II	10	3	13	26
2016 EAP III	9	9	19	37
2016 EAP IV	9	9	19	37
2017 EAP I	10	5	10	25
2017 EAP II	8	4	15	27

The letter grades given by instructors were converted to grade points (GPs) following Table 3.[2,3]

Table 3. Conversion of letter grades to GP

Letter grade	GP
S	4.0
A	3.0
B	2.0
C	1.0
F	0.0

4-2. Models

The grading data have a hierarchical structure where students are nested within groups (i.e. classes taught by individual instructors), and instructors are nested in terms of their affiliation or with respect to the registration blocks they teach. It is parallel to the structure which is observed in studies of assessment of students across schools (e.g., Goddard & Goddard, 2001; Goldstein et al., 1993). In the data which these studies analyze, students are nested within classes (teachers), which are nested within schools. Such hierarchical data do not satisfy the assumption of independence and so the application of ordinary single-level models to them leads to a problem of under-estimation of standard error. Also, it is difficult to tease apart effects of different levels if single-level models are

[2] While instructors report letter grades for students, these grades nominally represent particular parts of a 100-point scale: S (100−90), A (89−80), B (79−70), C (69−60), and F (59−0). The parts of the scale represented by each letter grade are not then of equal size, with a grade of F in particular being quite different from the others. The grade points assigned to the letter grades do not take account of this. However, since the number of students who received a grade of F was very small, this was not considered consequential for the statistical analyses.

[3] A grade of F is fundamentally different from S, A, B, and C in that F means no credits earned, and this difference may have an impact on grading practice. In this respect, the assignation of grade points to the letter grades does not fully reflect the meaning of the letter grades.

applied. Hierarchical linear modeling was developed to avoid these problems (Raudenbush & Bryk, 2002; see also Braun, Jenkins, & Grigg, 2006, for a review). As shown in Section 5-1, the variation between instructors in the grade data studied in this paper is somewhat large, so it is desirable to apply hierarchical linear modeling to our data also.

Specifically, three hierarchical linear models were used to analyze the EAP grading data. First, an unconditional model (a null model) was used. The level-1 (student-level) equation of this is as follows:

$$GP_{ij} = \beta_{0j} + r_{ij}$$

Here, GP_{ij} is the GP of student i of instructor j, β_{0j} is the intercept which represents the mean of GPs of the students of instructor j, and r_{ij} is the residual. The level-2 (instructor-level) equation is as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Here, γ_{00} is the grand mean of GPs over instructors and u_{0j} is the residual which indicates the variation between instructors. By comparing the two residuals, r_{ij} and u_{0j}, we can obtain information on how much of the variation in GPs across students lies between instructors, i.e., how much of the variation in GPs is due to instructors, as calculated below.

Second, a two-level hierarchical model of the variables GPs and instructors' affiliation was modeled. Let us call it the affiliation model. This model examines the effect of instructors' affiliation on the mean of GPs of the students. For that purpose, instructors' affiliation is represented using two dummy variables, ILAS and NILAS. When instructor j is an ILAS instructor, $ILAS_j = 1$; otherwise $ILAS_j = 0$. When instructor j is a non-ILAS instructor, $NILAS_j = 1$; otherwise $NILAS_j = 0$. This means that $ILAS_j = NILAS_j = 0$ if instructor j is a non-KU instructor. The level-1 equation for this model is the same as that for the unconditional model:

$$GP_{ij} = \beta_{0j} + r_{ij}$$

The level-2 equation is as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, ILAS_j + \gamma_{02}\, NILAS_j + u_{0j}$$

Here, γ_{00} is the grand mean of GPs over instructors, γ_{01} and γ_{02} are slopes for the dummy variables $ILAS_j$ and $NILAS_j$ respectively, and u_{0j} is the residual which indicates the variation between instructors.

Third, a two-level hierarchical model of the variables GPs and registration blocks was modeled. Let us call it the registration block model. Registration blocks of students were represented using dummy variables, ART1, SCI1, SCI2, and MED, as shown in Table 4. For example, for a student who belongs to Arts 1, ART1 = 1 and SCI1 = SCI2 = MED = 0.

Table 4. Representation of registration blocks using dummy variables[4]

Registration block	ART1	SCI1	SCI2	MED
Arts 1	1	0	0	0
Arts 2	0	0	0	0
Sciences 1	0	1	0	0
Sciences 2	0	0	1	0
Medicine	0	0	0	1

The level-1 equation is as follows:

$$GP_{ij} = \beta_{0j} + \beta_{1j} ART1_{ij} + \beta_{2j} SCI1_{ij} + \beta_{3j} SCI2_{ij} + \beta_{4j} MED_{ij} + r_{ij}$$

Here, GP_{ij}, $ART1_{ij}$, $SCI1_{ij}$, $SCI2_{ij}$, and MED_{ij} are the values of GP and dummy variables ART1, SCI1, SCI2, and MED of student i of instructor j, respectively, β_{0j} is the intercept, β_{1j}, β_{2j}, β_{3j}, and β_{4j} are slopes for $ART1_{ij}$, $SCI1_{ij}$, $SCI2_{ij}$, and MED_{ij}, respectively, and r_{ij} is the residual. The intercept, β_{0j}, represents the mean of GP over students of instructor j. The level-2 equation is as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$

Here, γ_{00} is the grand mean of GPs over instructors, and u_{0j} is the residual which indicates the variation between instructors. Note that in this model variation of the intercept across instructors was taken into account, but variation of the slopes for registration blocks across instructors was not.

The two-level hierarchical models presented in the above section were applied to the GP data of 2016 EAP I through 2017 EAP II using statistical software for hierarchical linear modeling, HLM7 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011).

[4] Registration block Arts 2, whose students got the lowest GP in most EAP courses, was taken as the default (ART1 = SCI1 = SCI2 = MED = 0) to make the interpretation of the model clearer.

4-3. Intra-class correlation

Intra-class correlation coefficient (ICC) is a measure of how much variation lies within and between groups. It is calculated from the variance of the level-1 residual r_{ij}, σ^2, and the variance of the level-2 residual u_{0j}, τ_{00}:

$$ICC = \tau_{00} / (\tau_{00} + \sigma^2)$$

ICC = 1 means that the variance of the level-1 residual is zero ($\sigma^2 = 0$), whereas ICC = 0 means that the variance of the level-2 residual is zero ($\tau_{00} = 0$).

Taking the unconditional model as an example, ICC = 1 would indicate that there is no variation in GPs among students of the same instructor and the variation in GPs is completely due to differences among (the mean of GPs across) instructors. On the other hand, ICC = 0 would indicate that there is no difference in the mean of GPs across instructors.

5. Results

5-1. The unconditional model

The unconditional model was applied to the GP data and the ICC was calculated to check the extent of the effect of instructors on the variation in GPs.

For the GP data of 2016 EAP I, the variance of the mean of GPs over instructors was statistically significant: $\tau_{00} = 0.115$, $\chi^2(25) = 263.773$, $p < .001$. This means there was a significant difference in students' grades due to instructors. The variance of GPs among students (σ^2) was 0.666, so ICC = .147. In other words, about 15% of the variation in GPs is between instructors. This somewhat large value of ICC indicates the presence of the hierarchical structure in the grade data and the necessity of applying hierarchical linear modeling.

The unconditional model was applied to the GP data of the other EAP courses similarly. The results are summarized in Table 5.

Table 5. Estimation of variance components under the unconditional model

Course	τ_{00}	*df*	χ^2	*p*-value	ICC
2016 EAP I	0.115	25	263.773	<.001	.147
2016 EAP II	0.112	25	308.617	<.001	.180
2016 EAP III	0.104	36	232.651	<.001	.114
2016 EAP IV	0.170	36	331.974	<.001	.170
2017 EAP I	0.038	24	114.804	<.001	.056
2017 EAP II	0.088	26	296.544	<.001	.135

The table shows that the variance of the mean of GP between instructors is statistically significant for all the EAP courses which were analyzed and it accounts for about 11–18% of the variation in

GPs except for the 2017 EAP I course, whose ICC is less than .10.

5-2. The affiliation model

The affiliation model was applied to the GP data to test whether the affiliation of instructors had an effect on GPs. For the data of 2016 EAP I, the instructor-level intercept γ_{00}, which is the grand mean of GP over instructors, was 2.696, $t(23) = 23.562$, $p < .001$. The data of 2016 EAP I also showed that affiliation of instructors did not have a statistically significant effect: for ILAS, $\gamma_{01} = -0.182$, $t(23) = -1.372$, $p = .183$ and for NILAS, $\gamma_{02} = 0.028$, $t(23) = 0.120$, $p = .905$.

The affiliation model was applied to the GP data of the other EAP courses similarly. The results are summarized in Table 6.

Table 6. Estimation of fixed effects under the affiliation model

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*-value
2016 EAP I					
Intercept, γ_{00}	2.696	0.114	23.562	23	<.001
ILAS, γ_{01}	−0.182	0.133	−1.372	23	.183
NILAS, γ_{02}	0.028	0.235	0.120	23	.905
2016 EAP II					
Intercept, γ_{00}	2.893	0.106	27.385	23	<.001
ILAS, γ_{01}	−0.258	0.127	−2.030	23	.054
NILAS, γ_{02}	−0.429	0.148	−2.893	23	.008
2016 EAP III					
Intercept, γ_{00}	2.539	0.076	33.462	34	<.001
ILAS, γ_{01}	−0.104	0.099	−1.052	34	.300
NILAS, γ_{02}	0.062	0.185	0.333	34	.741
2016 EAP IV					
Intercept, γ_{00}	2.637	0.092	28.815	34	<.001
ILAS, γ_{01}	−0.174	0.122	−1.423	34	.164
NILAS, γ_{02}	−0.013	0.226	−0.059	34	.954
2017 EAP I					
Intercept, γ_{00}	2.507	0.048	52.039	22	<.001
ILAS, γ_{01}	0.035	0.075	0.465	22	.647
NILAS, γ_{02}	−0.230	0.127	−1.816	22	.083
2017 EAP II					
Intercept, γ_{00}	2.549	0.091	27.970	24	<.001
ILAS, γ_{01}	0.030	0.104	0.288	24	.775
NILAS, γ_{02}	−0.260	0.179	−1.449	24	.160

The table shows that the effect of affiliation of instructors is not statistically significant for all the EAP courses except 2016 EAP II. For 2016 EAP II, the slope for the dummy variable NILAS is −0.429 and it is statistically significant at the level of $\alpha = .05$. This means that non-ILAS instructors on average gave about 0.4 points lower GP to students than non-KU instructors. Simply speaking, two out of five students of non-ILAS instructors received a one rank lower letter grade than students of non-KU instructors.

5-3. The registration block model

The registration block model was applied to the data to check the effect of the registration block of students. Results of the registration block model show us that there was an effect of registration block of students and that after taking account of the effect of registration block, there was nonetheless an effect of instructor.

The registration block model was first applied to the data from 2016 EAP I. The result shows that students in the Arts 1 and Medicine blocks received significantly higher GPs than those in the Arts 2 block: ART1, $\gamma_{10} = .252$, $t(1717) = 2.36$, $p = .019$; SCI1, $\gamma_{20} = .103$, $t(1717) = 1.17$, $p = .242$; SCI2, $\gamma_{30} = .117$, $t(1717) = 1.91$, $p = .056$; MED, $\gamma_{40} = .418$, $t(1717) = 5.27$, $p < .001$. At the same time, the variance of the instructor-level residual u_{0j}, τ_{00}, was .309, $\chi^2(25) = 240.1$, $p < .001$, and the intra-class correlation coefficient was large: ICC = .127. These results show that GP did vary significantly across instructors after taking account of the effect of registration block.

The registration block model was then applied to the other data. The results are summarized in Tables 7 and 8.

Table 7. Estimation of fixed effects under the registration block model

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*-value
2016 EAP I					
Base, γ_{00}	2.442	0.081	30.241	25	<.001
ART1, γ_{10}	0.252	0.107	2.355	1717	.019
SCI1, γ_{20}	0.103	0.088	1.171	1717	.242
SCI2, γ_{30}	0.117	0.061	1.909	1717	.056
MED, γ_{40}	0.418	0.079	5.265	1717	<.001
2016 EAP II					
Base, γ_{00}	2.734	0.088	31.140	25	<.001
ART1, γ_{10}	0.054	0.070	0.769	1719	.442
SCI1, γ_{20}	−0.000	0.081	−0.006	1719	.996
SCI2, γ_{30}	−0.100	0.077	−1.303	1719	.193
MED, γ_{40}	0.078	0.088	0.894	1719	.371

(continued)

Table 7 (continued). Estimation of fixed effects under the registration block model

Fixed effect	Coefficient	*SE*	*t*-ratio	*df*	*p*-value
2016 EAP III					
Base, γ_{00}	2.426	0.078	31.270	36	<.001
ART1, γ_{10}	0.198	0.007	2.673	1623	.008
SCI1, γ_{20}	0.145	0.067	2.168	1623	.030
SCI2, γ_{30}	0.001	0.120	0.012	1623	.991
MED, γ_{40}	0.197	0.078	2.519	1623	.012
2016 EAP IV					
Base, γ_{00}	2.434	0.009	26.322	36	<.001
ART1, γ_{10}	0.203	0.008	2.644	1594	.008
SCI1, γ_{20}	0.192	0.090	2.135	1594	.033
SCI2, γ_{30}	0.181	0.103	1.757	1594	.079
MED, γ_{40}	0.275	0.069	3.974	1594	<.001
2017 EAP I					
Base, γ_{00}	2.409	0.006	39.156	24	<.001
ART1, γ_{10}	0.184	0.006	2.896	1716	.004
SCI1, γ_{20}	−0.024	0.007	−0.334	1716	.739
SCI2, γ_{30}	0.011	0.068	0.160	1716	.873
MED, γ_{40}	0.147	0.069	2.161	1716	.031
2017 EAP II					
Base, γ_{00}	2.482	0.061	40.771	26	<.001
ART1, γ_{10}	0.167	0.078	2.130	1713	.033
SCI1, γ_{20}	0.010	0.055	0.191	1713	.849
SCI2, γ_{30}	−0.058	0.077	−0.745	1713	.456
MED, γ_{40}	0.074	0.058	1.275	1713	.202

Table 8. Estimation of variance components under the registration block model

Course	τ_{00}	*df*	χ^2	*p*-value	ICC
2016 EAP I	0.096	25	240.109	<.001	.127
2016 EAP II	0.120	25	326.509	<.001	.192
2016 EAP III	0.109	36	238.534	<.001	.119
2016 EAP IV	0.185	36	355.251	<.001	.183
2017 EAP I	0.042	24	141.443	<.001	.062
2017 EAP II	0.095	26	308.018	<.001	.145

Table 7 indicates that Arts 1 and Medicine students obtained statistically significantly higher GPs

than Arts 2 students in all the courses except for 2016 and 2017 EAP II. Table 8 shows that after taking account of the effect of registration block, there was nonetheless an effect of instructor in each case, but the effect was small (ICC = .062) for 2017 EAP I.

Using the estimate of τ_{00} as given in Table 8, comparisons were made between the extent of the effects for instructor in the two sets of grades from the same courses in 2016 and 2017. That is, comparisons were made between 2016 EAP I and 2017 EAP I, and similarly between 2016 EAP II and 2017 EAP II to test whether the effects for instructor were different in 2017. For EAP I in 2016 and 2017, $F(25, 24) = 2.284$, $p = .023$. Hence, the effects for instructor were significantly smaller for EAP I in 2017. For EAP II in 2016 and 2017, $F(25, 26) = 1.271$, $p = .273$. Hence, the effects for instructor did not change significantly for EAP II in 2017.

6. Discussion

The analyses reported in this paper set out to determine whether the grading of students taking EAP courses is affected by the individual instructor a student happens to be assigned to, by the affiliation of the instructor, or by the registration block a student is in (which stems from their departmental affiliation). While there are some differences between the six sets of EAP grades analyzed, the results show that although instructor affiliation does not generally impact on grading, both registration block and individual instructors do have an impact. Each of these factors will now be considered in turn.

First, as explained in Section 3, it was anticipated that instructor affiliation – that is, whether an instructor belongs to the Institute of Liberal Arts and Science, belongs to another department within Kanazawa University or is a part-time instructor – may have an impact on grading. As was noted, this could occur if instructors base their grading on the performance of students in their EAP classes relative to the performance of students in other classes they teach (whether within Kanazawa University or outside). There was, however, a significant difference in grading by instructor affiliation in only one analysis out of the six analyses conducted. It seems then that the EAP course management committee has largely been successful in communicating the EAP course aims, expectations and grading standards to instructors regardless of their affiliation, and that instructors have on the whole been successful in applying these standards as requested.

Second, the effects for registration block, which were found in five of the six sets of EAP grades, likely reflect differences in English ability among students. That is, the students in some registration blocks have a higher level of English proficiency than those in other groups, and these differences translate into different levels of performance in EAP courses. This interpretation of the results is supported by two observations. First, the differences in EAP grades by registration block largely correspond with the differences between the blocks in terms of TOEIC scores. Thus, the differences in English proficiency between students in different registration blocks identified by the TOEIC test seem to also have an impact on their EAP grades. Second, while significant differences

by registration block were observed in the grades for all other courses, there were no significant differences by registration block in the 2016 EAP II grades and only one significant difference between the blocks in the 2017 EAP II grades. This pattern of results makes sense when one considers the content of the four EAP courses. That is, EAP I, III and IV are all writing-focused courses, particularly in terms of their assessment tasks, and so it might be expected both that their grading would be similar and that their grading would reflect differences in the general English proficiency of the students. EAP II, in contrast, is a public speaking course, and thus, in addition to a base of English proficiency, the course also demands other, non-linguistic skills of students. The differences in EAP grading between registration blocks seem therefore to be quite reasonable and should not be viewed as inequitable in any way.

Finally, the foremost concern of this paper is whether individual instructors affect the grades students receive. An effect for individual instructor was observed, even after taking account of the effects of the registration blocks, in all six sets of EAP grades, and this effect was somewhat larger than that of the registration blocks. These individual instructor effects could stem from two factors. First, they may reflect differences in grading approach between instructors. Some instructors may interpret the grading guidelines more strictly than others and generally give lower grades to students, some may interpret the guidelines more leniently and give generally higher grades, while some may have a rather different interpretation and simply grade on a different basis from the majority of instructors. Second, the effects for individual instructor may reflect differences in the ability of instructors to help their students achieve the learning objectives of the course. That is, some instructors may be more successful in helping students to grasp what is required of them, in fostering an understanding of how to achieve those requirements and in motivating students to achieve the requirements. In other words, the differences in grading may result from differences in the performance of students, brought about by individual instructors.

Which of the two above factors – instructors' grading approach or instructors' success in engendering student achievement – explain the effects for individual instructor revealed by the analyses, or indeed whether a combination of the two factors is at work, is difficult to determine. As Beenstock and Feldman (2018) point out, in most stand-alone courses high grades received by students due to higher instructor quality is not in itself deemed problematic. It is influence from factors other than higher student ability and higher instructor quality affecting the grades that is deemed "unfair." However, both instructors' grading approach and instructors' success in engendering achievement may be considered inequitable for students of EAP courses, since, in each case, a student assigned at random to one instructor may receive a higher or lower grade than he or she would have done if assigned to a different instructor.

It is desirable, therefore, for these types of differences to be reduced as far as possible. Specifically, action should be taken to reduce differences in the interpretation of the grading guidelines and to support all instructors in improving their ability to help students achieve the

learning objectives. Starting from the initial reform of the curriculum, the EAP course management committee has taken a number of steps in this regard, as outlined in Section 2-3, and is continuing to take action with this aim.

There are some signs in the analyses reported above that these actions are beginning to bear fruit. The effect on grading of individual instructors was significantly smaller in the 2017 EAP I grades as compared with the 2016 EAP I grades. In fact, the variation in grading due to instructors was just 6% in the 2017 EAP I grades as compared with 15% in 2016. Nevertheless, there was no significant difference in the magnitude of the effect when comparing the 2016 EAP II and 2017 EAP II grades, and a significant, albeit small, effect for individual instructors remained in the 2017 EAP I grades. Consequently, the EAP course management committee is taking further action for the 2018 academic year.

With regard to the grading guidelines, the EAP course management committee has revised the rubrics for all the assessment tasks for each of the EAP courses. The new rubrics give more detail on both the features instructors should base their grading on and on how to give scores for the assessment tasks. In addition, in the syllabi for the 2018 academic year, more detail is provided on the assessment tasks themselves and how they should be implemented to try to ensure that all instructors are asking students to perform essentially the same task in the same way.

With respect to helping instructors support students in achieving the learning objectives of the courses, the EAP course management committee will: (1) continue to hold orientation sessions for instructors teaching the courses for the first time; (2) go on giving regular seminars for instructors providing ideas and advice for teaching the courses; and (3) expand the EAP Teacher's Guide for instructors which includes the syllabi and grading rubrics to also include week-by-week suggestions for teaching the courses along with worksheets and sample materials which teachers can use.

7. Conclusion

In this paper, we have examined liberal arts English course grades given to students by EAP course instructors. One limitation of the analysis is the lack of a second set of EAP III and EAP IV course grades. This was due to the fact that the research needed to be completed in the 2017 academic year so that the plans detailed above to improve the grading guidelines and support teachers could be developed in time for the 2018 academic year. In particular, a second set of EAP III and EAP IV course grades would have allowed comparisons between the effects for instructors in each year. It would be especially useful to know whether the smaller effect for instructors observed in the 2017 EAP I grades as compared with the 2016 EAP I grades was paralleled in the grading of EAP III and EAP IV.

In our statistical analysis, we found that instructors and registration blocks of students do have an influence on student grades, although the magnitude of the effect of instructors has become

smaller in the EAP I course. We also noted that the grades of the writing-focused EAP I, III, and IV courses broadly correspond with the TOEIC Test scores achieved by students in each registration block, whereas for EAP II, a speaking-focused course, this was not the case. Although it is not clear if instructor teaching ability or assessment practice causes the differences between instructors, differences of both types can affect students unfairly, so the EAP course management committee will endeavor to reduce these differences. As the committee has produced more detailed rubrics, teaching plans and teaching material samples for all the EAP courses, it is hoped that the assessments from 2018 will have a smaller instructor effect. This research is an ongoing undertaking. The data analyzed for this paper gives some guidance for devising better ways to make assessment in EAP courses more equitable and accountable for students and the university, and in the future further analyses will be conducted to continue this process.

References

Beenstock, M., & Feldman, D. (2018). Decomposing university grades: A longitudinal study of students and their instructors. *Studies in Higher Education*, *43*(1), 114–133.

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655–670.

Braun, H., Jenkins, F., & Grigg, W. (2006). *Comparing private schools and public schools using hierarchical linear modeling* (NCES 2006-461). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Commission on the Development of Foreign Language Proficiency. (2011). *Five proposals and specific measures for developing proficiency in English for international communication.* Retrieved from http://www.mext.go.jp/component/english/__icsFiles/afieldfile/2012/07/09/1319707_1.pdf

Goddard, R. D., & Goddard, Y. L. (2001). A multilevel analysis of the relationship between teacher and collective efficacy in urban schools. *Teaching and Teacher Education, 17*, 807–818.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19*(4), 425–433.

Kanazawa University. (2017). 金沢大学概要 2017 (Outline of Kanazawa University 2017). Retrieved from https://www.kanazawa-u.ac.jp/wp-content/uploads/2017/06/jp2017.pdf

Lu, J. (2008). A comparative study on college English education between Japan and China: Focusing on systems and social, cultural backgrounds. *Reports from the Faculty of Human Studies, Kyoto Bunkyo University*, *10*, 115–131.

Ministry of Education, Culture, Sports and Technology. (2003). *The National Action Plan to Cultivate 'Japanese with English Abilities'* (The 2003 Action Plan). Overview retrieved from

http://www.mext.go.jp/b_menu/hakusho/html/hpac200201/hpac200201_2_015.html

Ministry of Education, Culture, Sports and Technology. (2012). *University Reform Action Plan.* Retrieved from http://www.mext.go.jp/b_menu/houdou/24/06/__icsFiles/afieldfile/2012/06/25/1312798_01.pdf (Japanese Version), http://www.mext.go.jp/en/news/topics/detail/1372697.htm (English overview)

Ministry of Education, Culture, Sports and Technology. (2013, 2014). *English Education Reform Plan corresponding to globalization.* Retrieved from http://www.mext.go.jp/en/news/topics/detail/__icsFiles/afieldfile/2014/01/23/1343591_1.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk A., Cheong, Y. F., Congdon, R. T., Jr., & du Toit, M. (2011). *HLM7: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Rose, H., & McKinley, J. (2018). Japan's English-medium instruction initiatives and the globalization of higher education. *Higher Education, 75*, 111–129.

Tada, M. (2016). Recent reform to the English education system in Japan. *21th Century Education Forum, 11*, 21–29.