Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 12, June 2022 ISSN 1531-7714

A Mixed-methods Approach for Assessing Student Learning Gains in English Listening Comprehension[1,2]

Yingchen Wang, *School of International Studies, Shandong Youth University of Political Science*[3]
Xiaoxin Wei, *Cambium Assessment Inc.*
Yamin Liu, *School of International Studies, Shandong Youth University of Political Science*
Tiffany Chiu, *Department of Education, Health, and Behavior Studies, University of North Dakota*

The growth of literature in student learning gain in recent decades has posed challenges to address the issue within the classroom. To further shed light on this scientific body of knowledge, the current study implemented a sequential explanatory mixed-methods design on a sample of 76 students. First, test data were analyzed using the interval approach to calculate gain scores for the group. The learning gain was small. Second, individual learning gains were analyzed using the pseudo-anchors identified based on estimated Rasch item parameters. *T*-tests were performed on the stacked pseudo-anchors. While most students exhibited insignificant improvement, 10 students exhibited significant improvements in their anchors. Third, a Wright map was obtained to assess the student's self-reported gains. Specifically, learning gain was the highest when the students used VOA special English materials. Fourth, learning data were investigated using the multi-facet Rasch model. In the qualitative phase, follow-up interviews were administered and student learning logs were investigated. Thematic analyses and learning pattern examinations revealed the different motivation levels and learning strategies among the students. The connection of quantitative and qualitative analyses provided a more conceptual understanding of student learning gain. The mixed-method approach can be implemented and generalized to other settings, such as the classroom.

Keywords: listening comprehension; learning gain; sequential explanatory mixed-methods

Introduction

Classroom assessment is vital to teaching and learning, motivation and instruction (McMillan, 2013). To promote a student-centred learning experience in educational settings, assessment is instrumental to support student learning under different types of instructional delivery modes (Liang & Creasy, 2004; Ziegenfuss & Furse, 2021). However, classroom assessment is complex. It is more complex to quantifiably assess the desired learning gain in classroom, which is defined as growth or change in knowledge or skills (Rogaten et al., 2019). Moreover, to measure learning gain in the classroom, datasets are

[1] The authors would like to thank Mike Linacre for his help on the interaction analysis.
[2] The authors would like to thank Yunxia Han and Shujun Wang for providing their expertise to confirm the validity of test items.
[3] The first author would like to thank Siyue Guo and the coauthors for their constructive feedback on Figure 8.

frequently collected in settings with wide-ranging sample sizes from approximately 20 to even 200 to 300, reinforcing the idea that measuring learning change is "a nasty challenge" (Wright, 1996). In the face of these complex and dynamic research issues, Johnson and Onwuegbuzie (2004) advocates the use of mixed-method approach (i.e., methods utilizing both statistical and qualitative techniques) with its methodological pluralism. Mixed-methods research designs often lead to "superior research," in comparison with monomethod research (Johnson & Onwuegbuzie, 2004, p.14). The current research implemented a mixed-method approach to investigate the learning gains provided by various classroom constructed measurements.

Despite the application of mixed-methods approaches in social sciences, there seems to be limited research utilizing this "relatively new approach" (Creswell & Plano Clark, 2017) on academic learning gain in the context of classroom assessment. Given this under-researched topic (i.e., student learning gain) (Mathers et al., 2018), the current research distinguished from the existing body of literature on student academic learning gain by applying the mixed-methods design to properly measure learning gain in the classroom. What is more, it exemplified the need for follow-up explanations in classroom assessment to capture the complexity of academic learning gains. The follow-up phase allowed us to collect qualitative data to understand the rationale for differential learning gains and trajectories. Finally, this study mapped the conceptual understanding of learning gain utilizing the linked quantitative and qualitative results.

Literature Review

Learning gain is defined as a growth or a change in knowledge or skills demonstrated by students in relation to learning outcomes (Rogatenet al., 2019). It is another term for "value added" concept in education (Rogaten et al., 2019). The study of learning gain involves longitudinal data, at least differences or changes between two data collection points—pre- and post-measures. Learning gain is a key indicator for teaching excellence, student success, and quality education. Conceptually, learning gains may include academic, affective, cognitive and behavioural gains (e.g., Rogaten et al., 2019; Zhao et al., 2017). The current study focused on academic gains to better understand student learning outcomes in academia.

Assessment of Learning Gains

The practices of assessing learning gains have been a topic of debate in psychology and education for the past several decades. More specifically, researchers have endeavoured to investigate academic learning gains and determine influential factors for learning gains, using large-scale datasets across institutions. For example, Anaya (1999) used a subset of a national representative sample of newly enrolled freshmen. A number of variables were reported to significantly impact student self-reported learning gains. These variables included institutional characteristics, non-academic activities, learning environment and learning activities inside and outside of classroom. Furthermore, Cabrera et al. (2001) analyzed a large dataset from 7 universities. Their study identified several instructional practices that had positive associations with students' reported academic learning growth. These practices included interaction and feedback, collaborative learning, and clarity and organization, which emphasized the social-cognitive approach to student learning and motivation.

Researchers have also aimed to investigate learning change/gains at the institutional level. Terenzini and Wright (1987) probed into factors influencing student academic growth at a large, public university. They reported that social integration (including extracurricular activities, peer relations and social activities) was influential in students' reported academic growth during the junior and senior years. However, academic integration (i.e., frequencies of contact with faculty, relationship with faculty, participation in classroom activities) had a direct effect on students' reported academic skill growth each year. To understand learning gains at a tertiary education institution, Zhao et al. (2017) compared summative scoring approach with the Rasch modeling. Their findings indicated that the latter methodology revealed more significant gains. To clarify the effect size of student assessment, Mathers et al. (2108) reported that students' learning gains in science were small and not practically meaningful. The researchers concluded that faculty did not always receive assistance on practical ways to use assessment effectively for student learning.

Unlike the research in the preceding paragraphs, in which learning gain was the purpose, some

researchers have aimed to calculate learning gains to explore the effectiveness of different learning conditions. Sonbul and Schmitt (2013) set up three different types of input conditions for participants of native and non-native English speakers to examine how they affected collocational knowledge acquisition. All three conditions yielded significant short- and long-term gains in the explicit knowledge. In the non-native speakers group, the gains in the second condition were significantly superior to the first condition. To investigate the effect of TV viewing on L2 learning single words and formulaic sequences, Puimège and Peters (2019) measured learning gains using a form recall test, a meaning recall test and a form recognition test. The results indicated that there were significant gains with TV watching, but item characteristics and prior vocabulary knowledge mediated the learning outcomes.

The above findings demonstrated the role of assessment on learning gains from multiple sociocultural perspectives. It was pivotal to promoting and enhancing student learning in relation to the learning goals and educational outcomes. It was not only instrumental to teaching efficacy and competency, but also useful in identifying effective pedagogical practices. In fact, its role in educational reforms was vital with assessment reforms serving as "the very foundation of general educational reforms" (Cizek, 1997, p.8). Given the limited research in this line of work, the current research contributed to the growing body of literature on assessment of learning gain by adopting a mixed-methods approach to make use of the strengths of both paradigms and minimize the weaknesses of one paradigm (Johnson & Onwuegbuzie, 2004).

Mixed-Methods Research with Learning Gains as a Heuristic

According to Creswell and Plano Clark (2017), mixed-methods research has been applied in various disciplines including social, behavioral and health sciences. In education, limited research has applied this approach to study learning gains as a research goal. In some research, it is a heuristic for the research purpose. For example, to compare the effectiveness of online cooperative learning strategies in discussion forums with traditional online forums, Kupczynski et al. (2012) conducted a one-way ANOVA and found non-significant results. Therefore, they collected qualitative data, and conducted thematic analysis. Qualitative analysis revealed that participants in cooperative learning reported more learning benefits than those in the traditional group.

In addition, Liu et al. (2021) compared several input conditions and working memory groups (high vs. low) to examine the effects of attentional manipulations on language vocabulary learning. Based on the study, simple input enhancement for internal attentional manipulations (i.e., varying the contextual supports for the target expressions) was as effective as the compound input enhancement for internal attentional manipulations (i.e., capitalizing and underlining the target words). The compound input enhancement had higher gains, but it did not unambiguously bring about greater gains than the external manipulations in all cases. Liu et al. (2021) depended on MANOVA and ANOVA tests as well as on interviews with participants. In the interview responses, the participants confirmed that manipulating frequency of test input was effective, but such a manipulation may have negative impact in cases of excessive exposure.

Learning gain has also been applicable to students' professional development. For instance, to investigate PharmD students' professional development, Peeters and Vaidya (2016) adopted assessment for learning gain approach. Paired *t*-test and Cohen's *d* revealed a positive growth. Qualitative analysis was performed to triangulate quantitative results. Two types of data confirmed that all students seemed to have improved and those with less development at initial time improved more than others overtime. Formative assessments for learning guided students in their professional development.

Each of the mixed-methods research in this section analyzed learning gains to study pedagogical practices or program outcomes. The paucity of mixed-methods research in assessing learning gains as the central research focus is possibly due to the recency of the approach (Creswell & Plano Clark, 2017). Additionally, for most research studies, one type of data, either quantitative or qualitative, may be deemed as sufficient. When one data type is insufficient for research hypotheses, mixed-methods is justified (Creswell & Plano Clark, 2017). The current research focused not only on whether the students changed over time and who had changed, but also on why

students demonstrated a differential pattern of changes. Thus, quantitative and qualitative data were warranted.

Research Purpose and Questions/ Hypotheses

The purpose of this study was to assess and increase understanding of student academic learning gains in the classroom in a manner that allowed us to explore academic gains in different perspectives. Specifically, we sought to investigate whether learning gains occurred at the group and individual levels, using multiple constructed measurements. Moreover, we sought to understand why learning gains differed for each individual. The nature of "gain" indicated the need to quantify student academic learning. The statistical methods produce numerical results and provide probabilistic events. The "why" question indicated the need for a follow-up qualitatively-informed approach. The qualitative method described the complex phenomenon of academic learning growth by studying a few cases in depth. In the current sequential explanatory mixed-methods research (Creswell & Plano Clark, 2017), the integration of quantitative and qualitative analyses complemented each other and minimized errors of a single paradigm to capture a holistic picture of learning gains (Johnson & Onwuegbuzie, 2004).

Quantitative Research Hypotheses

1. We hypothesized that some students would show learning gains significantly. Regardless of probabilistic events, students were expected to progress towards the desired learning outcomes at different rates. Those with faster pace might produce significant gains, whereas others with slower pace might not.

2. We hypothesized that students as a group would demonstrate a significant academic growth. Results at this step were only global numbers. Regardless of the statistical results, it would be meaningful to investigate the students' perceptions about their academic growth using a self-assessment instrument.

3. In a self-assessment survey, we hypothesized that students would report more pronounced gains in some aspect(s) of listening comprehension than in others.

Four students reported no gains in some aspects of their listening comprehension. The 'no-growth' cases and a different change pattern warranted the need to conduct follow-up interviews with the students. Subsequent qualitative data were collected to explain the phenomenon in the quantitative analyses.

Qualitative Research Questions

1. Why did some students assess that they had made no gains? This question was only addressed to the students who reported 'no growth'.

2. What factors facilitated or hindered their learning gains in listening comprehension?

3. How did the students define "learning gains in listening comprehension"?

Mixed-Methods Question

1. How can the findings from quantitative and qualitative analyses inform us about student learning gains in listening comprehension?

Research Methodology

This study adopted a sequential, explanatory mixed-method design (Creswell & Plano Clark, 2017). This is a two-phases design—quantitative and then qualitative—separately (Figure 1). At each stage, a single type of data was collected and analyzed with predominately quantitative approaches. The two methods were integrated during the final interpretation.

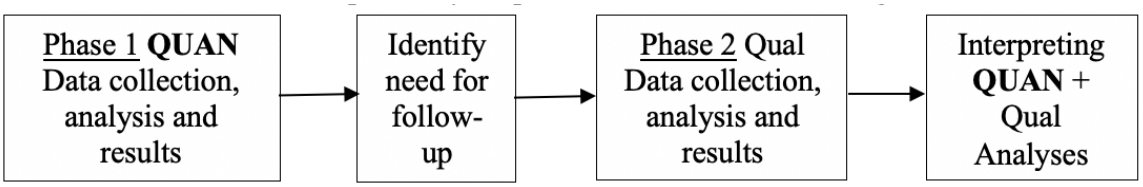
Sample

The participants in the current study were 76 English majors (70 females and 6 males) at a Chinese university. Data were collected from their English listening comprehension courses at two time-points, namely when they were freshmen and sophomores. Prior to the study, the researchers obtained approvals from the Institutional Review Board and written informed consents from the participants.

Quantitative Data Collection and Analyses

Data Collection. Three types of quantitative data were collected. The first two were summative tests collected from two consecutive end-of-term examinations administered to all students. These tests

Figure 1. Explanatory Sequential, Mixed-Method Design



Note: Bolded and capitalized letters denote the main focus.

comprised two parts—multiple-choice questions (40 and 55 items for the first and second tests; the total of 80 for all multiple-choice items on each test) and one dictation (the score of 20). For the multiple choice formatted items, each item was worth 2 points on the first test. On the second test, there were 30 1-point items and 25 2-point items. The second type of data was student learning performance data (hereafter referred to as "learning data"). The learning data was comprised of different tasks, which included learning log, class notes, assignments and quizzes. The first instructor required students to submit their in-class notes, which were used as an indicator of learning performance. She also evaluated students based on their outside-class notes, which detailed the students' learning activities and self-reflections. The in-class notes and outside-class notes of the first instructor's students were different tasks in the two semesters due to the varying teaching materials. The second instructor gave the students two different types of tasks each semester. In the first semester, students had assignments and quizzes (combinations of multiple-choice and open-ended questions). In the second semester, quizzes were retained. Assignments were changed to personalized learning logs, which students provided when using listening apps of their choices. Thus, each instructor evaluated students based on two tasks per semester. Teacher 1 rated the students using four categories: "4" representing scores $\geq$ 90 to 100; "3", 80–89; "2", 70–79; and "1", 60–69. Teacher 2 rated the students using the aforementioned four categories and one additional category, namely category "0", which represented scores of <60.

The third type of data was obtained from a self-assessment instrument. One of the instructors developed the self-assessment questionnaire for collecting information on learning gains in listening comprehension. The questionnaire was sent to some students for pilot testing. Feedback was solicited from the students. A student suggested removing a demographic item regarding the high school type. Another suggested adding an item assessing the overall improvement in English proficiency level. The suggestions were incorporated into the instrument items of the final version. The topics in the questionnaire included: (1) Challenges and factors that influenced learning progress upon entering the university; (2) Self-evaluation on learning gains in benchmark areas between the first and second semester; and (3) Amount of learning efforts for normal-speed and Voice of America (VOA) special English materials. The current research utilized the second portion of the instrument, which had a Cronbach's alpha of 0.839. Please refer to detailed item information in the "Learning Gains from Self-Assessment" section.

Originally, the survey items were on a five-point Likert scale (i.e., 'no gain', 'limited gain', 'a little gain', 'gain' and 'substantial gain'). Due to the sparse responses in some extreme categories, the items were combined into 2 categories ('limited or a little gain' and 'gain or substantial gain') with each category having ≥10 subjects. This was done, following the suggestion by Linacre (2020), who recommended the minimum sample size of 10 respondents per category for rating scale data. This size yields item and person parameter estimation precision with ±1 logit and 99% confidence interval. As a result of categories collapsing, four subjects who selected 'no gain' for some items were removed, reducing the sample size to 72. The four students deleted from the Rasch analyses were interviewed by the researcher.

Data Analyses. Table 1 shows that the analyses for group and individual learning gains using both observed and latent scores from the Rasch models. Rasch models, as measurement models, produce a consistent interpretation of the examinees' abilities (Stemler & Naples, 2021). Besides, studies have

confirmed the robustness of Rasch models for a small sample size. For instance, Wright and Stone (1979) applied dichotomous Rasch model to 18 items with 34 candidates successfully. A number of researchers have applied Rasch models in the context of classroom (e.g., Davidson & Henning. 1985; Han, 2018; Newhouse, 2014; Yan, 2020). In our study, the estimated *p*-values from the Rasch models were used to investigate if individuals and the group had significant gains. As presented in Table 1 and "Analyses for Group Gains" section, some group level indices were built on individual index.

Data Analyses - Analyses of Individual Gain. Current applications of the analyses utilized two distinct approaches central to the research purpose: the observed approach and the latent approach. For the observed approach, we calculated the individual gain on the test data, using $gain_i=[(post_i) - (pre_i)]/[100 - (pre_i)]$, where $post_i$- and pre_i-measures are individual scores (Bao, 2006). For the latent approach, Rasch analyses were conducted using self-assessment data, learning data and the test data. Four phases of evaluation/analyses were implemented. The initial phase was to evaluate the fit of the data to the models. Next was to identify pseudo-anchors from the test datasets and to perform *t*-test. The third phase was to examine the related *p*-values from the model as well as the relative measures for each student across semesters. The last phase was to obtain the visualization of student growth using the Rasch models.

In the initial phase, we evaluated item and person fitst statistics using jMetrik (Meyer, 2018) and MINIFAC (Linacre, n.d.). jMetrik is a free computer program for analyses with classical and modern psychological models. It was used to analyze test data and self-assessment data. MINIFAC is a free computer program with a limited capacity for diagnosing rating datasets with different facets. MINIFAC was used to analyze student learning data with the Multi-facet Rasch model

Table 1. Evaluation Indices for Group and Individual Gains

Source \ Level	Group Gains (Data)	Individual Gains (Data)	Note
Observed scores	1. Magnitude of average of gains (pre-post tests) 2. Magnitude of normalized gains (pre-post tests) 3. *p*-value from *t*-test with confidence interval (pre-post tests)	Size of individual gains for each student (pre-post tests)	
Rasch model	1. χ^2 test of compounded *p*-values for pseudo-anchors (pre-post tests) 2. *p*-value from fixed χ^2 test for students × time interaction term (learning data) 3. χ^2 test of compounded *p*-values for every student (learning data)	1. Wright map from Rasch (self-assessment gains) 2. Wright map from Rasch (learning data) 3. *p*-values of *t*-tests for each student from pseudo-anchors (pre-post tests) 4. Individual *p*-values for *t*-tests from interaction terms (learning data)	Identification of pseudo-anchors was presented in the next section.

Note: *p*-value was set at 0.05. A Wright map is a visual presentation.

Compound *p*-values was obtained using $\chi^2 = -2\log(p_1 p_2 p_3 \dots p_n)$

(MFRM; Linacre,1989, 1994) to obtain item and person fit indexes of the learning data. The means of item and person infit and outfit statistics for the datasets were close to 1.0 and the means of the standardized fit values were close to 0.0. Thus, the data exhibited a good fit to the Rasch models. About 5% of the items did not have satisfactory fit indices. Consultation with content experts confirmed the validity of these items.

In the second phase, to place students within one unambiguous numerical framework, pseudo-anchors were identified using the estimated item parameters of the test data from jMetrik. *T*-tests were performed on the academic ability estimates from pseudo-anchors. For this purpose, we adopted the recommended procedures (Luppescu, 2005; Linacre, personal communication, January 22nd, 2020; Mallinson, 2011; Zhao et al., 2017; Wright, 1996, 2003). Systematic steps were followed. Step 1 involved identifying pseudo-anchors. We ran the Rasch model separately on the test datasets and rank-ordered the estimated difficulty parameters before pairing the items according to content and item difficulty. The rationale of anchoring procedure is based on the invariance property of item parameters of the Rasch model. When common items are estimated separately in different datasets, the item parameters should theoretically remain invariant. Thus, when items of similar content are close in item difficulty, they can serve as potential pseudo-anchors. Additional methods are presented in the research by Wright (2003) and Longford (2015). To examine how well the pairing was, the correlation between paired items was calculated. If the matching was successful, an identity line was observed on the scatter plot.

Initially, all items of the test data were entered into jMetrik, but the program dropped the dictation items from the analyses. For more items entering into the anchoring process, attempts were made to split dictations into sets of polytomous items as well as sets of binary items. All combinations of the recoded dictation on test 1 failed to converge, suggesting that the dictations not enter into the anchoring process. Consequently, anchoring was performed on the multiple-choice items, among which 19 pairs of items were successfully paired.

Step 2 involved adopting a single-group design, stacking (i.e., combining data vertically) the responses with the paired items and running the Rasch model on the stacked anchor data (Mallinson, 2011; Wright, 2003; Zhao et al., 2017). The ability estimates were then obtained by fixing the item difficulty for anchors at Time 1 because the interest was change at Time 2 (Wright, 2003).

Step 3 involved examining whose latent learning scores had changed significantly. For this purpose, the statistic $t_{T1,T2} = (\beta_{n_{T1}} - \beta_{n_{T2}})/\sqrt{SE^2_{n_{T1}} + SE^2_{n_{T2}}}$ with $df = (2I_a - 2)$ was computed, where I_a was the number of pseudo-anchors. $\beta_{n_{T1}}$ and $\beta_{n_{T2}}$, $SE_{n_{T1}}$ and $SE_{n_{T2}}$ were the ability estimates and standard error for Time 1 and Time 2, respectively, which were determined from the stacked data analyses in Step 2 (see the previous paragraph).

In the third phase, MFRM analysis was performed on the learning data. To run MINIFAC program, a connected design is essential. Therefore, the group means of all the tasks and semesters were fixed at 0 (Linacre, 2012). The MFRM in the current study was a four-facet model (student, task, rater, and time) with an interaction term between student and time. The interaction term for student and time was integral to this study because one of research goals was the changes in student learning outside the classroom across semesters. MINIFAC produces *t*-tests with *p*-values for interaction terms at individual levels. A significant term implies that the student daily learning had changed significantly over time. Non-significance implies that the daily learning had not progressed significantly. The size of the interaction term larger than 2.0 logit is a signal for further investigation. In addition, the program produces relative measures for the interaction terms, which could be utilized as approximate estimates for the student ability in each semester (Linacre, personal communication, December 2nd, 2020).

In the last phase of the latent approach, a Wright map was obtained from Rasch analysis conducted on self-assessment data to reveal the areas in which students thought they had gained. The Wright map from MINIFAC was presented to examine the visual change in latent abilities of the students rated over time.

Data Analyses - Analyses for Group Gains. For the observed scores, we calculated dependent *t*-test with 95% confidence interval. Two types of gains were

calculated (Bao, 2006). Normalized gain was calculated using $gain_{normalized} = \frac{[(post)-(pre)]}{[100-(pre)]}$, where *post* and *pre* are class averages (Hake, 1998). Next, average of gains was obtained using $gain_{average} = \sum_{n=1}^{N} \frac{[(post_i)-(pre_i)]}{[100-(pre)]}/N$, where *N* is the total number of students and $post_i$ and pre_i are individual scores. For the pseudo-anchors (see the preceding section), individual *p* values were calculated and compounded into a group-level index by using the following equation: $\chi^2 = -2\log(p_1 p_2 p_3 \ldots p_n)$ with $df = 2N$ (Fisher, 1932, as quoted in Anselmi et al. 2015). MFRM on the learning data produced not only *t*-test for each individual, but also fixed χ^2at global level. The fixed χ^2 tests whether the elements of the facets are heterogeneous. A significant value suggests that the elements are heterogeneous. Individual *p* values from MFRM were also compounded into a group-level χ^2.

Qualitative Data Collection and Analysis

Data Collection. Students demonstrated a different pattern of learning gains (see *"Student Learning Gains at Individual Level"*). Four participants reported no learning growth in some aspects of their listening comprehension. To examine the different change pattern and no-growth cases, qualitative data were collected from three different sources. The first were interviews with students who reported no growth. The second source was the additional students we reached out, and 19 of them agreed to participate. The convenience sample of students (17 females and 2 males) were selected based on two criteria: their availability and academic performances. Their performances varied from high, intermediate to low proficiency levels. Due to the conflict of academic schedules, individual interviews were performed via QQ (i.e., a Chinese social media platform) for data collection. The third source derived from the students' learning logs and classroom notes.[4]

The focus of the interviews varied. For the no-growth students, the interview concerned the rationale for making no progress. The interview questions for the additional 19 students included, "What factors facilitate or hinder your improvement in listening comprehension?", and "How do you define learning improvement in listening comprehension?" Although the word "gain" implies a quantitative result, it is impossible for students to quantify their learning. Thus, we used "improvement" instead of "learning gain" in the interview.

Data Analyses. Qualitative data analysis was performed in three steps. The first two steps were qualitative analyses. The last step involved linking the results of quantitative and qualitative analyses.

First, coding and thematic analysis was performed on data from student interviews (Creswell, 2003; Creswell & Plano Clark, 2017). The first author had training in qualitative research methods and performed the coding and thematic analyses on interview data. When new ideas emerged, coding was refined and merged with other similar ideas.

Next, searches for patterns of motivation and cognitive learning strategies were conducted, using available student learning logs and notes. Each time a learning pattern or a set of learning strategies was identified with a student, we cross-validated this trend with other logs and notes as well as the student's logs in another semester.

The third step involved integrating the separate alignments of the quantitative results with the qualitative results. Given the limited space, we only utilized 'no-growth' cases in self- assessment data and anchor results from the quantitative analyses for a number of reasons. First of all, the four 'no-growth' cases necessitated further investigation. Secondly, the interest was in learning gains. Individual gains from anchor results fit the purpose of the study better. Lastly, Figure 6 under "Quantitative Results" showed that the rating biases might exist with two-thirds of the students above zero logit and that positive biases suggested that grading for passing might exist. The first alignment was to match the students' GPA in the first semester with the results from previous qualitative analyses for their learning profiles. Students were classified into four profiles, each with its English proficiency level, motivation and learning strategies.

[4] The university requires that each instructor submits at least one-third of the assignments per semester for official documentation purposes.

The second alignment was to link 'no-growth' cases with their profiles. The third alignment was to connect the results of anchor analyses with the profiles. The final result was a graph connecting quantitative and qualitative analyses. We sought feedback from university faculty members, who mentioned the importance of personal characteristics. It was added to the final graph.

Results

Quantitative Results

The correlation between multiple choices and dictation was 0.699 and 0.634 in test 1 and test 2, respectively. The reliability of test 1 was 0.728 for the whole test and 0.792 for the multiple-choice items. This index of test 2 was 0.702 for the whole test and 0.738 for the multiple-choice items.

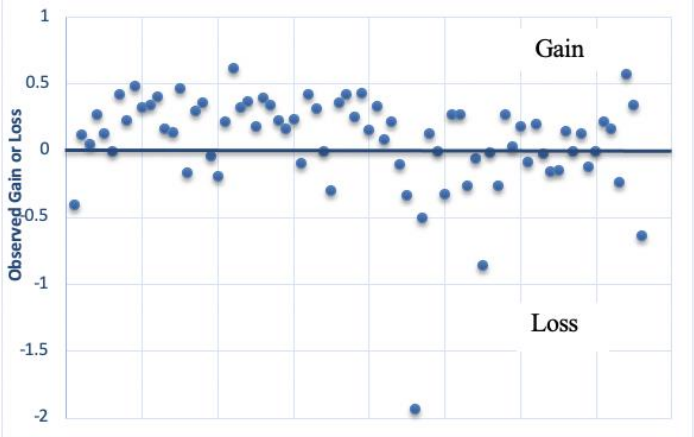
Student Gains at Individual Level - Individual Gains. Figure 2 displays student learning gains at the individual level. A total of 23 students exhibited lower scores in the posttest than in the pretest (below the line), five students exhibited the same scores in the pretest and posttest (on the line) and 48 students exhibited higher scores in test 2 than in test 1 (above the line).

Student Gains at Individual Level - Anchoring Approach. Table 2 presents the virtually identified 19 items. To avoid the confounding effect of item formats, only items with the same format and similar content were aligned. A total of 19 items were paired. They represented all the contents of the multiple-choice items. The item difficulty parameters estimated from two calibrations ranged from approximately 3.3 to −3.0. Each anchor item from test 1 was matched with a corresponding item with the similar content from test 2 (Table 2)—passage-based, number, or vocabulary. Each pair of passage-based items was similar with respect to the characteristics of the question.

The 19 paired items achieved a correlation coefficient of 0.994. The scatter plot of the pseudo-anchor items showed almost a straight line (Figure 3). The correlation, content and the number of anchors indicated that the pseudo-anchors could be considered a mini version of test forms with respect to content and statistical representation.

Figure 4 displays the scatter plot for ability change obtained from the stacked data. Determined on the basis of the anchored items, the abilities at Time 1 and Time 2 exhibited a correlation of 0.509. The line passing through (0, 0) in the plot indicated that more students exhibited ability gains and some exhibited ability losses. According to the *t*-test results, 10 students exhibited significant gains ($p<0.05$) and 9 students exhibited insignificant downward progression. A total of 45 students exhibited insignificant upward progression and 12 students progressed neither upwards nor downwards.

Figure 2. Plot of student individual learning gain



Note: Each dot represents a student.

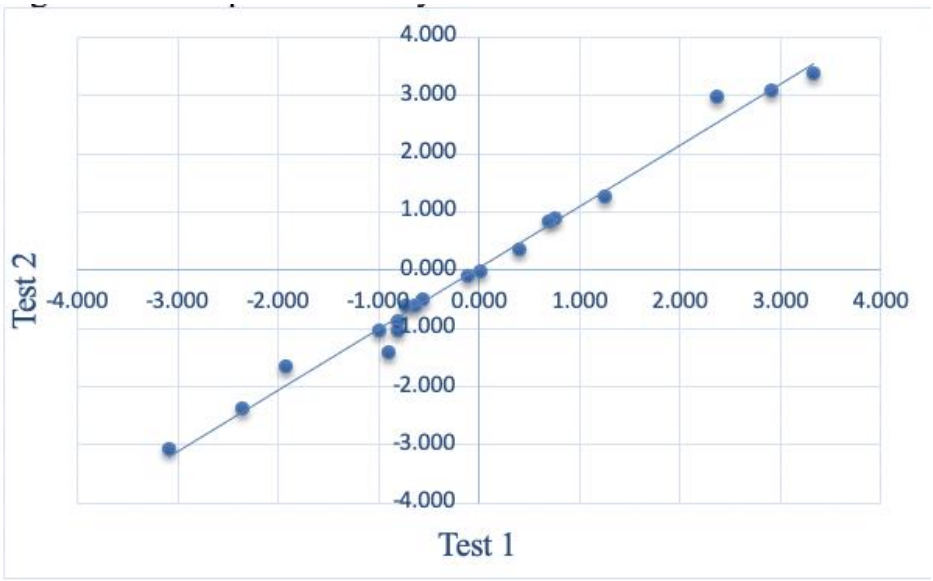
Table 2. Virtually identified common items

Test 1			Test 2		
Item	Content	*b*-parameter	Item	Content	*b*-parameter
Item 1	Vocabulary	3.335	Item 1	Vocabulary	3.387
Item 2	Inference*	2.902	Item 2	Inference*	3.075
Item 3	Vocabulary	2.366	Item 3	Vocabulary	2.982
Item 4	Vocabulary	1.252	Item 4	Vocabulary	1.254
Item 5	Number	0.759	Item 5	Number	0.894
Item 6	Summary*	0.698	Item 6	Conclusion*	0.831
Item 7	Number	0.394	Item 7	Number	0.366
Item 8	Detail*	-0.56	Item 8	Detail*	-0.504
Item 9	Vocabulary	-0.121	Item 9	Vocabulary	-0.108
Item 10	Number	0.013	Item 10	Number	-0.022
Item 11	Number	-0.642	Item 11	Number	-0.62
Item 12	Vocabulary	-0.726	Item 12	Vocabulary	-0.62
Item 13	Detail*	-0.814	Item 13	Detail*	-0.884
Item 14	Inference*	-0.814	Item 14	Inference*	-1.037
Item 15	Number	-1.002	Item 15	Number	-1.037
Item 16	Detail*	-0.905	Item 16	Inference*	-1.41
Item 17	Fact*	-1.92	Item 17	Fact*	-1.651
Item 18	Fact*	-2.361	Item 18	Fact*	-2.374
Item 19	Vocabulary	-3.082	Item 19	Vocabulary	-3.075

Note: The item sequence number in this table is unrelated to the test item sequence.

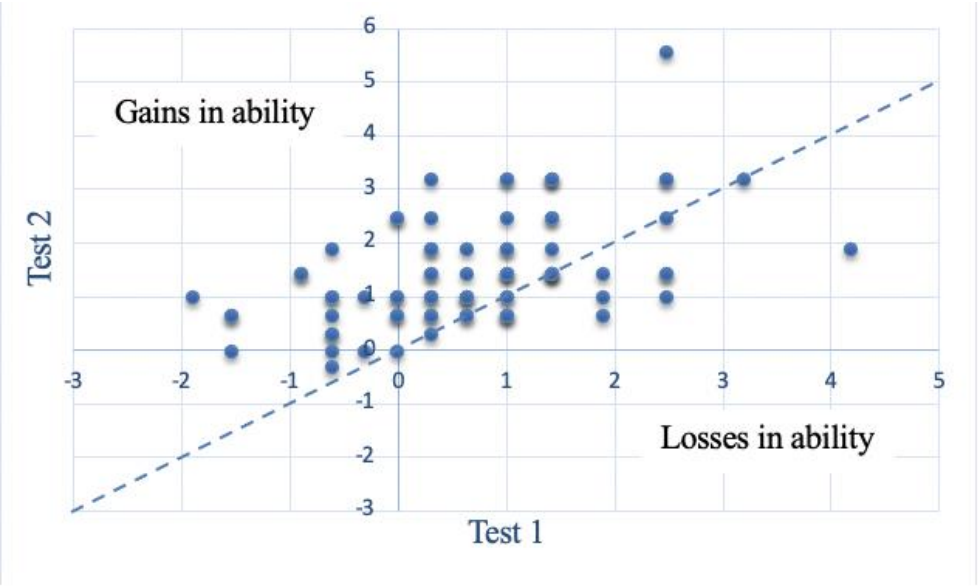
*Passage-based multiple-choice items.

Figure 3. Scatter plot of virtually identified anchor items



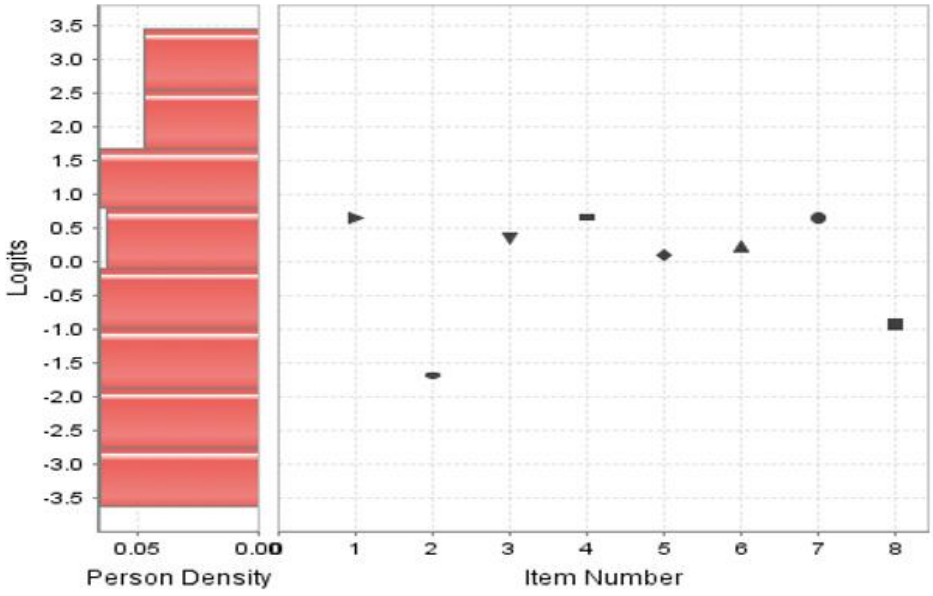
Note: Each dot represents a pair of items.

Figure 4. Scatter plot of ability change based on anchors



Note: Each dot represents a student.

Figure 5. Wright map for self-assessment



Note: From left to right, each shape represents growth in a certain area.

Student Gains at Individual Level - Leaning Gains from Self-Assessment. Figure 5 displays the Wright map for the self-assessment. In Figure 5, each shape from left to right represents gains in some areas: (1) understanding miscellaneous listening materials; (2) VOA special English; (3) normal-speed English; (4) response time; (5) vocabulary; (6) English listening skills; (7) overall English proficiency level, and (8) overall listening comprehension. The students tended to endorse their learning gains in VOA special English.

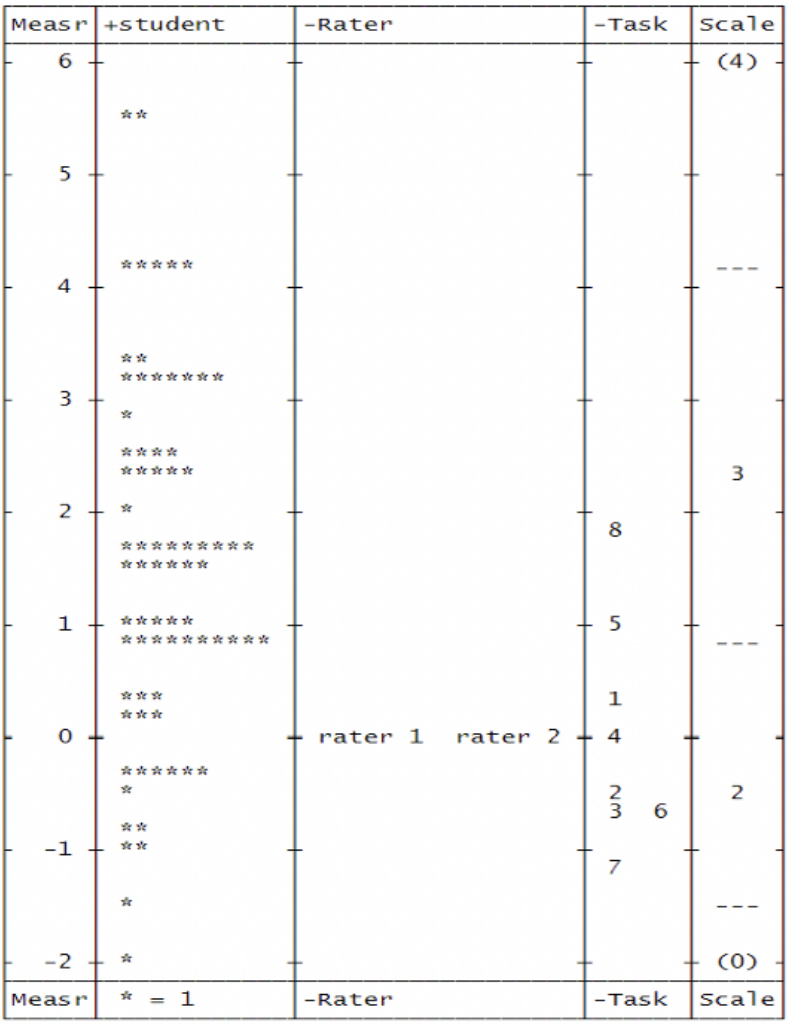
They were also likely to agree that they had made some progress in overall listening comprehension. However, they reported less gains in their overall English level, quickness to respond and their understanding of miscellaneous listening materials.

Student Gains at Individual Level - Learning Gains from MFRM Analysis. Figure 6 displays the Wright map for the learning data. Few students were at the extreme ends of the latent ability distribution. Most students were located between -1.0 and +3.0 logits. The strata value for the students (2.37) indicated the existence of two clusters of students.

Figure 7 displays the interaction plot for the students by semesters. The individual relative measures were mapped as proxy data of the estimated student ability in each semester. When reading this figure, one should start at each dot in semester 1 and refer vertically for its counterpart in semester 2. Forty students moved above the zero line. By contrast, 36 students exhibited negative progression. For instance, the relative measure for student 66 was 2.87 in semester 1, but the estimate dropped to -1.92. The relative measure for students 50 was 2.13, but it went down to -1.56 in the second semester. Dividing the relative measure by its standard error yielded the *t*-test statistic. None of the test statistics was significant.

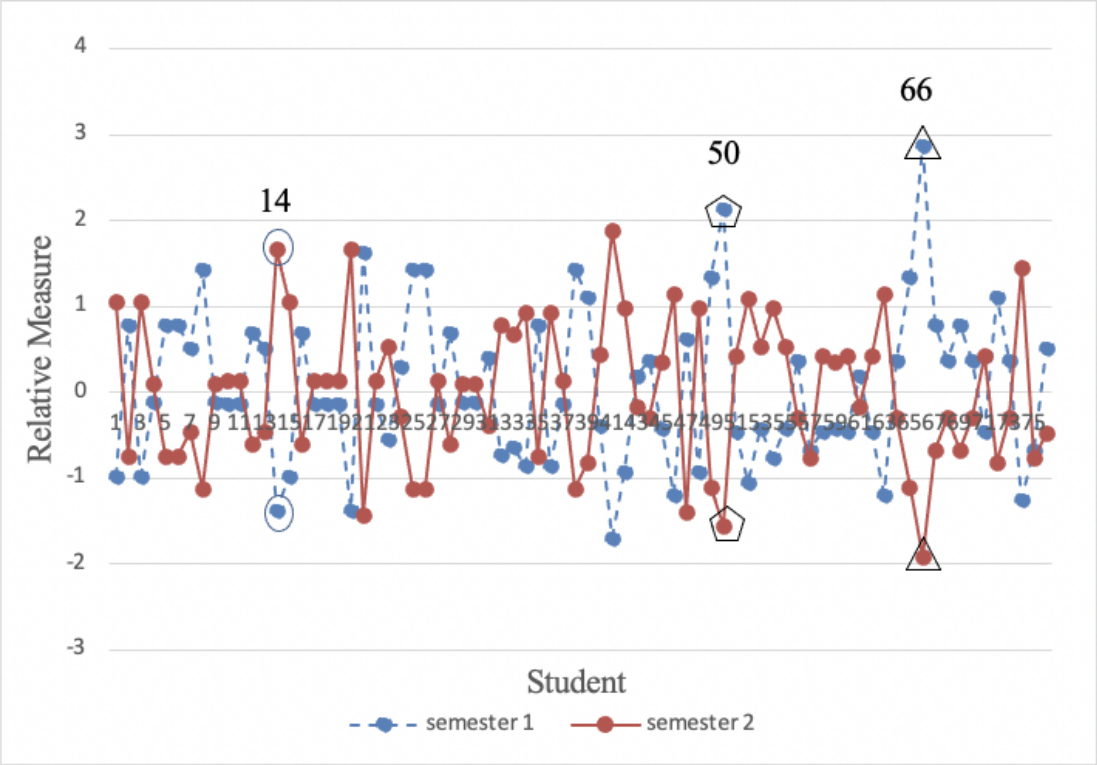
Student Learning Gain at Group Level. Table 3 shows that the average of gains was 0.079, smaller than the normalized gain of 0.158. When the. average of gains is less than 0.3, the gain is small (Hake, 1998). However, *t*-test produced a significant statistic of 4.465 with df=75 (p<0.01). 95% confidence interval ranged from 6.337 to 2.427, indicating a significant gain over time.

Figure 6. Wright map for learning data



Note: Each star represents a student.

Figure 7. Interaction plot of relative measures for students by semester



Note: Each dot represents a student.

Table 3. Test statistics for group growth

Indices	Value	Significance	Effect Size/Note
Average of gains	0.079	NA	Small
Normalized gain	0.158	NA	Small
t-test	4.465 with *df*=75	$p<0.01$	CI: 6.337 to 2.427
Compounded χ^2 from anchor	100.823 with *df*=152	Nonsignificant	
Fixed χ^2 from MFRM	79.7 with *df*=152	Nonsignificant	
Compounded χ^2 from MFRM	27.29 with *df*=152	Nonsignificant	

Note: Fixed χ^2 is obtained from the MINIFAC output. NA=not applicable

Compounded χ^2 was calculated using $\chi^2 = -2\log(p_1 p_2 p_3 \ldots p_n)$

A non-significant result was obtained in the χ^2 test at the group level for anchored items. The fixed $\chi^2$for interaction term was insignificant, which confirmed that the elements in the semesters were homogenous statistically. The compounded χ^2 from MFRM was also insignificant. These mean that students as a group did not change significantly over time.

Considering all test results, we concluded that students grew positively as a group. However, their individual learning gains were different. Moreover, each of the above statistical analyses had its limitations. For instance, *t*-test and gains (normalized gain and average of gains) were only able to capture the group-level gain. They both were based on observed scores, which were much affected by sample characteristics. The analyses from the Rasch models utilized latent scores, but only from two time-points. The latent change analysis from time-period 1 to time-period 2 included random elements. Self-assessment is

prevalent in literature but its drawbacks is self-reported. Qualitative analyses would complement the shortcomings of quantitative analyses.

Qualitative Results

We asked participants to define learning improvement in order to determine factors of facilitating and hindering learning improvement. The qualitative analyses led to the emergence of four themes from interview data with 19 students. Four different rationales arose for the interviews with four 'no-gain' cases. From the log files, three learning patterns emerged, revealing the complexity of cognitive strategies and motivation in learning. Four different learning profiles emerged when we linked the qualitative results with their GPA in the first semester. The following sections discussed these qualitative results.

Results of Thematic Analyses – Factors Facilitating or Hindering Learning Improvement. Table 4 presents the findings on what helped or hindered learning improvement. The first theme was the teachers' expectations and skills. Students mentioned that teachers in other courses required them to use some application softwares for vocabulary memorization, which helped with vocabulary expansion. Instructors the listening comprehension course required students to engage in listening apps for more than 10 minutes every time for three to four days per week. The second theme was the students' English proficiency level, which included vocabulary, response speed, spoken English, and knowledge of English culture. The third theme concerned the student motivation to varying types and degrees. For instance, some forms of motivation may be extrinsic due to the requirements of the course. Others may have intrinsic motivation because interest was top priority. The fourth theme was cognition. Students differed with their complexity of perception and learning. For example, some students pointed out that, to improve, they needed inherent interest in the subject to facilitate learning in listening and speaking. Others revealed unidimensional approach to learning. For them, improving only required the implementation of "drills" or "practices."

Results of Thematic Analyses – Definition of Learning Improvement. The 19 students defined the construct of learning improvement in listening comprehension primarily in four different ways. Notice that student ID numbers were independent of those in quantitative results.

Enjoyment or Appreciation. Two students defined learning improvement in listening comprehension as enjoyment of learning or appreciation for intellectual stimulation. "In my opinion, improvement in listening comprehension is to be able to understand what is not understood previously, to find out that I am able to understand it completely. Sometimes, I enjoy it, feeling calm and not agitated. The more agitated, the worse it is" (Student 8). "Whatever materials I listen, I understand, appreciate the culture and feel connected" (Student 13).

Interacting in English**.** Three students defined learning improvement as the ability to interact with others in English in social environments. "Improvement in listening is connected with improvement in spoken English" (Student 1). "For me, the improvement in listening comprehension is not only understanding the speakers, but also being able to express myself immediately in English…. If we don't understand what is being said, naturally we cannot interact with others in English" (Student 2). "Broadly speaking, it should include the ability to communicate with others in English" (Student 17).

Understanding the Materials**.** Almost all the students defined learning growth as understanding the listening materials. Altogether, 16 students endorsed this opinion. "Improvement means understanding" (Student 16). "Improvement means that I can understand the listening materials that fit my level and understand the main ideas" (Student 14). "Improvement means that I can progress from no understanding to understanding some sentences…. As long as I understand the gist, I am improving" (Student 12). "I find myself improved when I go back to the old materials. They are no longer so difficult" (Student 3 and 10). "Compared with my previous level, I understand better. Each time I listen to some material, I can get the main idea" (Student 7). "Getting not only the main ideas, but also the details" (Student 5, 18 and 19). "For extensive listening, follow the speaker. For intensive listening, understand accurately each word" (Student 6).

Table 4. Influential factors that improved and challenged learning growth

Factors	Themes	Examples
What are helping?	1.Teachers' expectations and skills. 2. English proficiency level 3. Motivation 4. Cognition	1. "Another teacher asked us to use one application software to memorize the words." "We were expected to take notes and do autonomous learning. We had to do." 2. "Vocabulary and English proficiency level go hand in hand" "Pronunciation" "Improvement in oral English". 3. "You have to learn yourself." "It depends on each individual motivation" "We had to do the homework." 4. "Interest is of top priority. If you are really interested, you can absorb more".
What are hindering?	1. Lack of English proficiency 2. Amotivation 3. Lack of cognition	1. "Limited vocabulary" "Non-standard pronunciation" 2. "Lack of attention" "Some students are not very motivated" "Lack of self-regulation." 3. "Lack of practice" or "Lack of vocabulary"

Improving the Grades**.** Four students explicitly specified improvement as receiving a better grade. Student 11 and 15 said, "The indicator for improvement is the number of correct answers I choose in a test." "Practically, it should include improvement in academic performance." (Student 13) "Improvement should be reflected in better grades, and the number of correct answers." (Student 17)

Results of Thematic Analyses – Interview Themes with No-growth Cases. One of the students reported that the average amount of time spent on listening was less than 10 minutes per day, whereas most students reported 15 minutes or more. He admitted his lack of self-discipline. The second student concentrated only on VOA special English materials and barely spent any time on normal English work. This student reported no change in the response time. The third student reported that the lack of vocabulary was a major problem. The fourth student admitted that he was considerably behind other students when he was admitted into the programme. In addition, he admitted his lack of motivation to learn.

Motivation, Cognitive Strategies, and Learning Profiles. Learning logs or notes revealed three different learning patterns and cognitive strategies of the students. The first group, was highly motivated. Their learning logs were consistently characterized by textual enhancement, frequent review of materials, focus on areas of challenge, frequent annotations of new expressions in English and attention to details. The second group was somewhat motivated. Textual enhancement and de-contextualization may appear on the same page. More often, Chinese translations were used to explain the new words and expressions. Students were not that attentive to details. The last group of students was the least motivated. Some of them copied and pasted notes from the listening apps. Some produced 1 to 2 pages of written notes for 2 months, whereas most students created more than four pages of notes in the same period. Sometimes, they took no notes or scribbled some notes. Their behaviors were consistent with their interview responses ("no interest", "no self-regulation").

When we connected the above three groups with their GPA in the first semester and with their perceptions of learning improvement, four learning profiles emerged (Table 5). The first group with high English proficiency level at time-period 1 had less room for improvement. Comparatively, the second group with an intermediate English level and high motivation had more room to improve. The third group, somewhat motivated, also had some room for improvement. The last group had substantial room for improvement, but the lack of motivation may potentially lead to academic issues.

Table 5. Learning profiles

Source	1st Group	2nd Group	3rd Group	4th Group
Proficiency at time 1	High proficiency level	Intermediate proficiency level	Intermediate to low proficiency level	Low proficiency level
Definition	Interest or appreciation, better grades, understanding	Interest, better grades, understanding	Grade, understanding	Grades, understanding
Learning logs	Motivated, good learning strategies	Motivated, good learning strategies	Somewhat motivated, some learning strategies	Lack of motivation, limited learning strategies

Note: The proficiency refers to students' GPA at time-period 1.

The definition refers to how students defined the construct of learning improvement.

Discussion

The present study demonstrated the utility of using different techniques for analysing various datasets. The proposed methodology can also be extended to datasets collected at more than two time-points. The research findings were discussed respectively in "Quantitative Findings Unconnected with the Qualitative Analyses" and "Mixed-Methods Findings".

Quantitative Findings Unconnected with the Qualitative Analyses

The observed-score approach revealed that most of the students progressed positively and the learning gains were small. Although the *t*-test was significant with confidence intervals excluding 0, the right bound was 2.427. This magnitude indicates that the practical significance may be limited. The questions is, "Where did the small learning gains occur?" The second test was related to VOA special English materials. The students reported the most gains in response to VOA special English listening materials and less in their overall English level, vocabulary and response speed. Learners of English as a second language live in non-native English environments, where their native language is ubiquitous. Therefore, students chose the materials that best facilitated the process of learning to process and extract input. Consequently, they found more gains in VOA special English.

The MFRM analysis of the learning data revealed that the students did not academically progress in a linear pattern over time. Some students fluctuated more. For example, in Figure 7, student 66 did poorly in one quiz in the second semester. In contrast, student 14 improved with daily notes and in-class notes. More than half of the students demonstrated upward progression. However, based on the probabilities associated with *t*-test statistics, no students changed significantly. The teachers classified the students into two groups. This classification may or may not have accurately reflected the true distribution of student latent ability because most students were positively graded above zero logit. Figure 6 revealed that the instructors might have assigned grades at random more often at the higher end and not have been able to distinguish among grades (especially, 3, 2 and 1). Thus, grading for passing and rater effects might have existed.

Mixed-Methods Findings

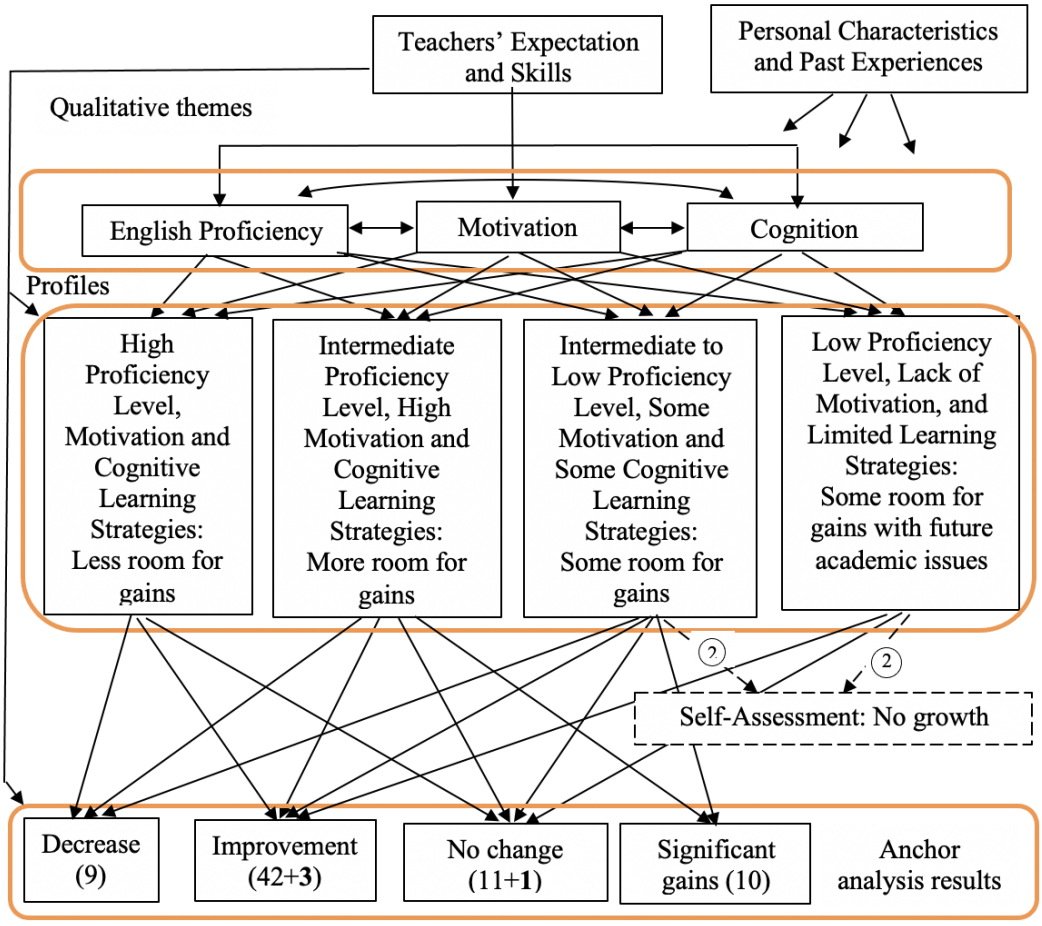
To explain the rationale for individual students making differential gains in listening comprehension, we linked the qualitative and quantitative results (Figure 8). The figure demonstrated the important role of teachers, previous performance, motivation and cognitive learning strategies in the short-term learning gains of students.

Previous English Proficiency, Motivation and Cognition. High achievers at time 1 did not have much room to gain. They could decrease, improve a little or remain unchanged. In comparison, intermediate or low-level achievers at time-period 1 had room to move up. The

amount in which they could gain depended on individual motivation, cognitive learning strategies and performance at test 2. Those motivated intermediate achievers with good cognitive learning strategies could move into the "significant gains" group or "improvement" group. Some of them might also demonstrate no change and decrease because of performance at test 2. Intermediate to low level achievers with some motivation and cognitive learning strategies might improve or gain significantly. Others might either decrease or stay stagnant. Lastly, low-level achievers at time-period 1 with lack of motivation and cognitive learning skills were less likely to have more gains. They demonstrated some improvement. They were also likely to have no change due to their lack of inner drive, performative, and cognitive skills.

Teachers' Expectations and Teaching Skills. The teachers' expectation and teaching skills impacted student learning gains in both direct and indirect ways. As one student put it, "we have to use those apps, take down the notes and reflect because it is the teacher's requirement." The requirements from the teachers possibly stimulated students' motivation (either intrinsic or extrinsic) as well as cognitive learning strategies. The teachers' expectation and teaching skills did influence the student's short-term learning gains both directly and indirectly. However, it remains unclear how much each student could benefit from the teachers' expectations and skills in the long-term. We expect that those with intrinsic motivation will experience more learning gains in the long-term.

Figure 8. Linking Qualitative Results with Quantitative Short-term Outcomes



Note: Each arrow indicates the direct effect.

The bolded numbers indicate the number of students' report of 'no-growth'.

The numbers in the circles indicate the students' report of 'no-growth'.

Personal Characteristics. Background and personal characteristics played significant roles in learning gains, for example, students defined learning improvement differently. They implemented varying learning strategies and engaged in learning in different ways. Some enjoyed their learning process. Others focused on improving the grades. Some highlighted their notes with different colours and marks, while others took de-contextualized notes more often. Poor performance might have made some suffer from setback, but others might have put it aside, recover and improve. For example, one student said "my performance was poor in this area, and I don't want to learn." Some low-level performers at time-period 1 did not express frustration, and they improved significantly.

Conclusion

This study proposed the sequential mixed-method approach for measuring learning gain. Multiple analytic techniques complement each other and reveal different aspects of learning gains. Based on the results from different datasets and analytic techniques, most students progressed upwards. Learning occurred mostly utilizing VOA special English materials. Many students improved without exhibiting statistical significance. The amount of gains depended on their motivation and cognitive learning strategies. The mixed-method research design is not only essential for assessment of learning for students, but also instrumental for teaching.

The results provided impactful information to teachers about learning gains and assessment activities. The procedure of identifying anchors in the absence of common items and the subsequent Rasch analyses of the stacked data enable the determination of individual learning change, which is difficult to determine using many other methodologies. If the dependency between time-points is a concern, one can follow the procedures of Chien (2008), Wright (2003), or Zhao et al. (2017). The results suggested the necessity of designing a more scientifically sound assessment plan. Formative assessment is absent in the course assessment plan. In practice, there are positive implications for formative assessment to be integrated into the teaching due to the established evidence about the usefulness of formative assessment to promote learning (National Academies of Sciences, Engineering, and Medicine, 2017; National Research Council, 2001).

As the secondary approach, the qualitative analyses complemented quantitative analyses. The qualitative research revealed student learning issues and processes that the primary techniques did not detect. Practically, the qualitative results suggested the necessity of motivating students and teaching them effective learning strategies. The integration of quantitative and qualitative results captured the underlying learning process. The mixed-method approach is superior to assess student learning gains within the classroom setting.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Anaya, G. (1999). College impact on student learning: Comparing the use of self-reported gains, standardized test scores, and college grades. *Research in Higher Education*, *40* (5), 499-526.

Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes, 13* (1), 16-16.

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics*, *74*(10), 917-922, https://doi.org/10.1119/1.2213632

Cabrera, A. F., Colbeck, C. L., & Terenzini, P. T. (2001). Developing performance indicators for assessing classroom teaching practices and student learning. *Research in Higher Education*, *42*(3), 327-352.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of Classroom Assessment: Learning, Adjustment, and Achievement* (pp. 1–32), San Diego, CA: Academic Press.

Creswell, J. W. (2003). *Research design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: SAGE.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2* (2), 164–179.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66* (1),64–74. https://doi.org/10.1119/1.18809

Hambleton, R.K, & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Springer

Han, C. (2018). Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. *Assessment and Evaluation in Higher Education*, *43* (6), 979-994. https://doi.org/10.1080/02602938.2018.1424799

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14-26. https://doi.org/10.3102%2F0013189X033007014

Kupczynski, L., Mundy, M.A., Goswami, J., & Meling, V. (2012). Cooperative learning in distance learning: A mixed methods study. *International Journal of Instruction*. 5(2): 81-90.

Lee, M., & Jung, J. (2021). Effects of textual enhancement and task manipulation on L2 learners' attentional processes and grammatical knowledge development: A mixed methods study. *Language Teaching Research*, https://doi.org/10.1177/13621688211034640

Liang, X., & Creasy, K. (2004). Classroom assessment in web-based instructional environment: Instructors' experience. *Practical Assessment, Research, and Evaluation*, *9* (7). https://doi.org/10.7275/84mr-wp41.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press. (Original work published in 1989)

Linacre, M. (2012). *Many-facet Rasch measurement : Facets tutorial 4 anchoring*. https://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M. (2020). *A User's Guide to WINSTEPS and MINISTEP Rasch-model Computer Programs*. Program manual 4.8.0. https://www.winsteps.com/a/Winsteps-Manual.pdf

Linacre, M. (n.d.). *MINIFAC Rasch Measurement Computer Program* (Demo Version of Facet) [Computer software]. Chicago: https://winsteps.com/minifac.htm

Liu, Y., Nassaji, H., & Tseng, W. (2021). Effects of internal and external attentional manipulations and working memory on second language vocabulary learning. *Language Teaching Research*. 1-41. https://doi.org/10.1177/13621688211030130

Longford, N.T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics, 40* (3), 227-253.

Luppescu, S. (2005). Virtual equating. *Rasch Measurement Transactions, 19* (3), 1025. https://www.rasch.org/rmt/rmt193a.htm

Mallinson, T. (2011). Rasch analysis of repeated measures. *Rasch Measurement Transaction_251* (1), 1317 https://www.rasch.org/rmt/rmt251b.htm

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149–174. https://doi.org/10.1007/bf02296272

Mathers, C. E., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*, *43*(8), 1211-1227. https://doi.org/10.1080/02602938.2018.1443202

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge

Meyer, J. P. (2018). jMetrik (Version 4.1.1) [Computer software]. https://itemanalysis.com/jmetrik-download/

McMillan, J. H. (2013). Why we need research on classroom assessment. In J. H. McMillan (Ed.), *Sage Handbook of Research on Classroom Assessment* (pp. 3-16). Thousand Oaks, CA: Sage

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy*

& Practice, 21 (2), 205-220. http://dx.doi.org/10.1080/0969594X.2013.868341

Onwuegbuzie, A. J & Leech, N.L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies, *International Journal of Social Research Methodology, 8*(5), 375-387. http://dx.doi.org/10.1080/13645570500402447

Peeters, M. J., & Vaidya, V. A. (2016). A mixed-methods analysis in assessing students' professional development by applying an assessment for learning approach. *American Journal of Pharmaceutical Education*, *80*(5): 1-10. https://doi.org/10.5688/ajpe80577

Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: An exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, *47*(4), 424–438. https://doi.org/10.1080/09571736.2019.1638630

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. *University of Chicago Press*. (Original work published in 1960).

Rogaten, J., Rienties,B., Sharpe, R., Cross, S., Whitelock, D., Lygo-Baker, S., & Littlejohn, A. (2019): Reviewing affective, behavioural and cognitive learning gains in higher education, *Assessment & Evaluation in Higher Education*, 44 (3): 321-337 https://doi.org/10.1080/02602938.2018.1504277

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, 63, 121–159. https://doi.org/10.1111/j.1467-9922.2012.00730.x

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research, and Evaluation*, *26* (11), https://scholarworks.umass.edu/pare/vol26/iss1/11

Terenzini, P.T., & Wright, T. M. (1987). Influences on students' academic growth during four years of college. *Research in High Education*, 26, 161–179. https://doi.org/10.1007/BF00992027

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press. https://www.rasch.org/BTD_RSA/pdf%20[publisher]/Best%20Test%20Design.pdf

Wright, B. D. (1996). Time 1 to time 2 (pre-test to post-test) comparison: Racking and stacking. *Rasch Measurement Transactions,* 10 (1), 478. www.rasch.org/rmt/rmt101f.htm.

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, *17* (1), 905-906. www.rasch.org/rmt/rmt171a.htm.

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment and Evaluation in Higher Education, 45* (2), 224 -238. https://doi.org/10.1080/02602938.2019.1629390

Zhao, Y., Huen, J. & Chan, Y.W. (2017). Measuring longitudinal gains in student learning: A comparison of Rasch scoring and summative scoring approaches. *Research in Higher Education, 58* (6), 605-616. https://doi.org/10.1007/s11162-016-9441-z

Ziegenfuss, D. H., & Furse, C. M. (2021). Flipping the feedback: Formative assessment in a flipped freshman circuits class. *Practical Assessment, Research, and Evaluation*, 26 (8). https://scholarworks.umass.edu/pare/vol26/iss1/8.

Citation:
Wang, Y., Wei, X., Liu, Y., & Chiu, T. (2022). A Mixed-methods Approach for Assessing Student Learning Gains in English Listening Comprehension. *Practical Assessment, Research & Evaluation*, 27(12). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/12/

Corresponding Author:
Yingchen Wang
School of International Studies, Shandong Youth University of Political Science
No. 31699 Jingshidong Road, Jinan, Shandong Province, PRC, China, Zip code 250103.
Admission office telephone: (86-0531)58997707 ； School telephone：(86-0531)58997000

Email: wang.graceful [at] yahoo.com