福岡女子大学文学部・国際文理学部紀要
「文藝と思想」第 78 号　2014 年 2 月　(47)～(62)頁

Evidence Based Practice: Evaluating a Reading Textbook

Trevor A. HOLSTER
J. LAKE

Abstract

Evidence based practice in education aims to identify effective pedagogical practices by evaluating all available evidence. This requires explicit specification of desired outcomes, operationalized through curriculum objectives. This paper evaluated outcomes from an academic English reading program, identified areas of weakness, and then gathered evidence to diagnose probable causes of the problems. The key findings from this were that instruction was ineffective for lower ability students and that student engagement and motivation had decreased following their entry to university. This evidential basis allowed a plan for remediation to be implemented. Revised pedagogical objectives were drafted and then evidence was gathered on student needs. This allowed both qualitative and quantitative evidence to be evaluated on the suitability of academic reading textbooks. Finally, outcomes from pilot administration of the revised textbook and objectives were quantitatively evaluated, showing substantive improvement and serving as a case study on the implementation of evidence based practice in curriculum revision.

Introduction and Background

The use of evidence-based practice (EBP), “the implementation of effective treatments in real-world clinical settings” (Saville, 2009) is becoming widely

accepted within the field of education, with Hattie's (2009) synthesis of meta-analyses an exemplar of how research derived evidence can be summarized to provide an evidential basis that is both comprehensive and accessible to non-researchers. The essential feature of Hattie's work is the reporting of effect sizes of educational outcomes so that the relative effectiveness of different pedagogic practices can be compared. This illustrates two crucial requirements for EBP: reporting of effect sizes in a manner allowing comparison between different groups or pedagogical interventions, and explicit definition of objectives. Without these two features, it is not possible to evaluate appropriate evidence and identify treatments that are effective and those that need to be reconsidered. Therefore, before a language program can implement EBP, it is paramount to first understand the nature of pedagogical objectives and how these relate to the structure of a program.

Japanese universities universally offer language courses to students, but institution-wide coordination of language programs is problematic (Inoue, 2006). Fukuoka Women's University (FWU) is an exception in this regard because all first-year and second-year students must complete the Academic English Program (AEP), composed of 15 language classes, organized as four courses: Academic Writing (AW1 to AW4), Academic Reading (AR1 to AR5), Academic Listening (AL1 to AL2), and Communication Strategies (CS1 to CS4). This arrangement of the 15 classes composing the AEP isolates each of the traditional four skills rather than promoting integrative skills classes, making it crucial that each course of study has clearly defined objectives so that teachers and students understand the purpose of each class and its place within the AEP.

Crucially, the AEP was established as a language *program* rather than a collection of isolated classes. A program, in contrast to isolated classes, is made up of different parts that work together for a coherent purpose. For example, a computer program is a series of instructions that are designed to work together for a common purpose. When individual parts of a program do not work together, the program becomes *incoherent*. Two ways that programs can become incoherent are lack of *unity* and lack of *cohesion*. Unity means that the program has a single overall purpose. In a large program, different parts will have different objectives,

but all the parts should contribute to a single overall goal. Cohesion means that the different parts are related to each other in systematic ways. Attempting to combine a collection of isolated classes into a program without a systematic description of the role of each class will result in a lack of both unity and cohesion. The lack of unity results because of a lack of clearly articulated program goals and objectives. The lack of unity in turn precludes any systematic relationship between different classes. Thus, a language program cannot, by definition, have unity and cohesion without clearly specified program goals and objectives. Without a coherent program, EBP is not possible because the nature of the evidence required is inherently disputable, so different analyses may simultaneously define the same practice as both successful and unsuccessful.

The AEP is an academic English program, and the official program goal articulated by FWU is to enable students to function in academic situations in English. However, this goal is far too broadly defined to guide planning of specific objectives at the classroom level. Thus, *goals* are very general and *objectives* are specific, so EBP must be implemented at the level of objectives. Although the general nature of goals means that they are easier to specify, specification of goals alone is insufficient to organize a unified, cohesive program. Unity and cohesion require that all parts of the program are related to the overall program goals, and this is only possible if each class has clearly defined objectives. Key considerations in setting objectives relate to students' motivations, personal goals, and proficiency levels. Thus, successful objectives will be *learner centered*, defined closely with regard to the needs of learners within a specific program and the context of that program.

Ultimately, motivation is the most important factor in second language learning. This is because developing proficiency takes thousands of hours of study and practice, a process Schumann and Wood (2004) call Sustained Deep Learning (SDL). Because SDL takes years of hard work, the major hurdle for language learners is maintaining motivation. In the SDL model, students will remain motivated when they have positive goal appraisals and environmental engagement. This occurs when students find learning activities novel, pleasant, goal relevant, within their coping ability, and compatible with their self image and social image. When this is achieved, students experience powerful emotional rewards which affect future

preferences and choices in positive ways. These positive experiences are critical to maintain long-term motivation, so the most important factors in choosing classroom activities and textbooks is that students have positive experiences of novelty, pleasantness, relevance, self/social image compatibility, and task difficulty.

Of these motivational factors, the one most easily addressed at the program planning level is task difficulty. If instructional tasks are too easy, students will not be challenged and may become bored, but if tasks are too difficult, they will be unable to cope, leading to stress. Boredom and stress are both demotivating, so objectives must be appropriate for the proficiency level of the students, in turn requiring that the ability of students entering the program be assessed before objectives are specified. The range of student ability in the AEP is quite large. TOEFL scores in 2011 ranged from approximately 350 to slightly above 500. Objectives that are suitable for the highest level students in the AEP are not suitable for the lowest level students. This means that a placement test is needed to separate students into suitable levels, and it also means that different objectives are needed for classes at different levels. Less obviously, if student interests and motivations differ systematically by proficiency level, qualitatively different task types may be required at different levels, meaning that specification of task type is an essential consideration in developing program objectives.

There are many different ways to classify classroom activities. This is because language learning is very complex and involves many different processes. The different ways of classifying activities reflect the different processes that occur. One important classification is between improving *declarative knowledge* and *procedural knowledge*. Declarative knowledge is often called *explicit knowledge* and means that we can explain or describe what we know. For example, being able to explain grammatical differences between English and Japanese is declarative knowledge. Procedural knowledge is often called *implicit knowledge* and relates to being able to do things. For example, the ability to use natural stress and pronunciation without thinking is procedural knowledge. Native speakers of a language internalize procedural knowledge as children, but frequently have very weak declarative knowledge, while second language speakers typically rely much more

on declarative knowledge. It is normal for low level students to develop declarative knowledge first when they learn new language features. As their level improves, they need to spend more time on activities that develop procedural knowledge, so that they can develop fluency with the new language.

Another important classification is between intensive and extensive activities. *Intensive activities* aim to teach small quantities of new language. Introducing new language is a challenge for students, but the small quantity of material keeps it within their coping ability so that they don't become demotivated. Intensive activities often focus on declarative knowledge because teachers need to explain new language features. *Extensive activities* are not intended to introduce new or difficult language. Instead, students must comprehend or produce large quantities of relatively easy language about familiar or interesting topics. This is very important for developing procedural knowledge because procedural knowledge requires thousands of hours of practice. This amount of practice can easily demotivate students, so it's important that students find practice activities novel and relevant. Therefore, it is normal for students to be given choices about extensive activities, for example by being allowed to choose what books to read, what topics to write about, or which partners to work with.

A further distinction is between *teacher centered* and *learner centered* activities. Intensive activities are often teacher centered, meaning that students mostly follow instructions and explanations from the teacher about what to learn and how to learn it. Extensive activities, however, require a learner centered approach, where students make decisions about what to learn and how to learn it. Teacher centered activities and learner centered activities are both important for all levels of student, but lower level students usually need more teacher centered activities and higher level students need more learner centered activities.

Japanese high-school English classes often have 40 or more students per class, many students have low engagement and motivation, and teachers are extremely busy. This makes learner centered and extensive activities problematic in typical high-school classrooms, meaning that when students enter university English classes, they are often not familiar with learner centered and extensive activities, and prefer intensive activities where the teacher gives explanations about English and tests focus on

declarative knowledge. Therefore, in order to prepare Japanese university students to study overseas, they need to be introduced to the types of activities that are typical of English language classrooms and understand their purpose.

Although full-time instructors with training in second language teaching are becoming more common in Japanese universities, most universities still rely heavily on part-time teachers without specialized training in second language acquisition, and are thus unfamiliar with the range of activities used in language teaching and the reasons for using them. When language programs employ teachers with such diverse ranges of specializations and backgrounds, unity and cohesion are quickly lost unless textbook selections are monitored to ensure the difficulty level, content specifications, and task types are compatible with the program objectives. In the case of the AEP, by specifying standard textbooks for each course, full-time instructors can prepare teaching materials and explanations of the reasons for task selection for part-time instructors, avoiding unnecessary duplication of work and ensuring that all teachers and classes work towards objectives consistent with the official goals of the AEP.

Analysis and Results

AEP students take TOEFL IP tests at the end of their first, second, and third semesters of study. In the second semester of 2011, the average improvement was 2.7 points. This was a very disappointing result, but some students showed much larger gains while the scores of other students decreased, so more detailed analysis was needed to find the reason for the disappointing results. The Educational Testing Service explain the problem of "regression to the mean" (RTM) (Swinton, 1983), which causes misleading results where low-level students to appear to make large gains and high-level students to appear to make small gains. The implication of this is that EBP cannot use raw placement test scores as grouping variable to compare interventions between groups of different ability, so adjustment for RTM is necessary. A simple way to adjust for RTM is to compare the *distribution* of scores instead of the raw scores of individual students. In Figure 1, the 2011 AEP cohort was divided into eight approximately equal groups

using TOEFL IP scores from July 2011 and then sorted into eight new groups using the scores from January 2012. Group 8 is the lowest AEP level and Group 1 the highest, reflecting the eight class levels in the AEP program. It is essential to understand that Figure 1 does not compare gains for individual students, but the average level of AEP classes if placement was made based on the TOEFL scores from July 2011 or January 2012. It is this comparison of the *distribution* of scores rather than of individual scores that removes the effect of RTM. If AEP instruction was successful, the average TOEFL level of all classes should have improved, but this did not happen. Although the higher level classes improved, the mid level and low level classes showed smaller gains or losses. This indicated that the AEP program in 2011 was ineffective for low-level and mid-level students.

The disappointing results in 2011 prompted an urgent search for evidence based explanations and interventions. All AEP students were surveyed using an attitude to school (ATS) survey (Cheng & Chan, 2003) at the beginning and end of the first

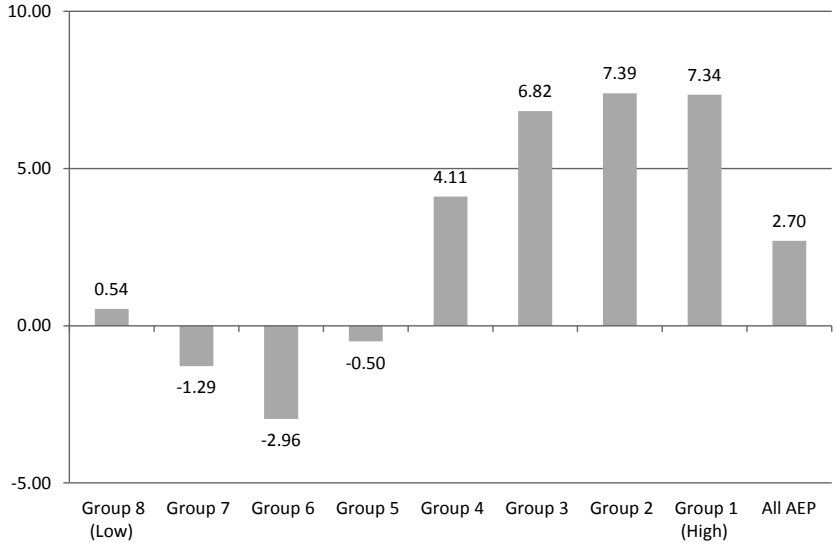

Figure 1 Changes in TOEFL score distribution in 2011. Scores from the July 2011 and January 2012 TOEFL IP tests were used to assign students into 8 levels. Although the four highest levels showed improvement, the lowest four levels did not improve.

semester of 2012. Figure 2 compares the mean response to each of the nine survey items in Week 1 and Week 14. It must be noted that this survey addresses attitudes in general, not just to the AEP program, so the problems identified must be considered to reflect on FWU generally, not just AEP classes. Attitudes at the beginning of the semester were positive overall. As FWU classes had just begun when the first survey was conducted, this must have reflected a combination of high-school experience and anticipation of university classes. The positive overall results suggest that students had high anticipation of FWU classes. However, attitudes became substantively less positive by the end of the semester. This may be unavoidable because students' excitement about entering university cannot be sustained after they experience the reality of university study, but two survey items raise concerns. Item 2, students' "sense of achievement", and Item 6, students' "participation in school life" were rated low at the beginning of the semester and fell even further by the end. These two items are of concern because a positive experience at university should improve them, or at least hold them

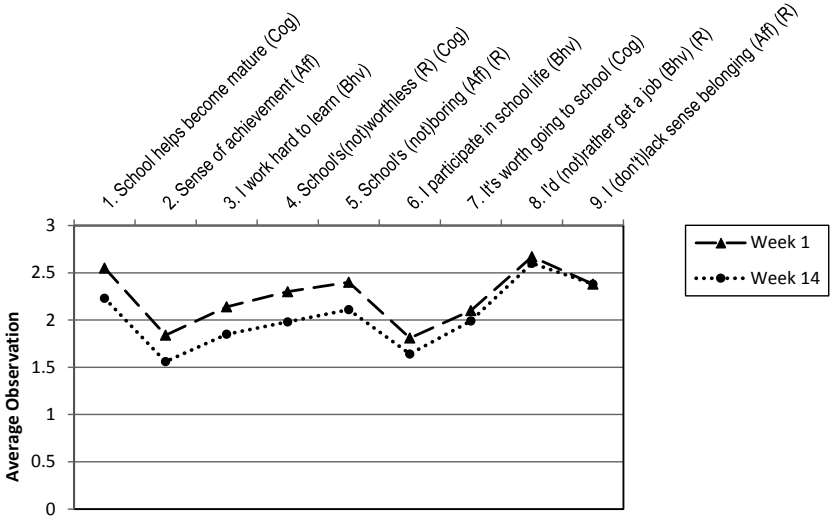

Figure 2 Changes in students' attitude to school during the first semester of 2012. A higher position on the vertical scale indicates a more positive attitude to school. Large decreases were seen in six of the nine items. Most noticeably, reported sense of achievement (Item 2) and participation in school life (Item 6) were very low at the end of the semester.

constant. Additionally, Item 5, "School's (not) boring", showed a large decrease. This evidence supported the view that FWU classes were not engaging students in interesting and challenging activities and that major changes in the curriculum were required. The first stage of this process was evaluation of the suitability of AEP textbooks. Five of the 15 AEP classes teach academic reading, so reading textbooks were considered first as this seemed likely to produce the biggest improvement.

Vocabulary tests were administered in the second semester of 2011. Figure 3 shows estimated vocabulary knowledge, which provides a guide to the semantic knowledge of students, allowing comparison to the semantic difficulty of texts (Stenner, Burdick, Sanford, & Burdick, 2007). Written academic text requires extensive knowledge of morphological derivatives (Biber, 2006), so word families were considered to be the most useful measure of semantic knowledge for the AEP. Word families comprise a headword, or dictionary form, plus morphological derivatives. For example, "run" is a headword, with "ran", "running", "runs", and "runner" other derived forms belonging to the same family. AEP students mostly fall between 2000 and 2500 word families. This is insufficient to read academic English, so vocabulary development in the Academic Word List (AWL) (Coxhead, 2000) and 3000 to 5000 word families was considered a priority.

However, determining the reading ability of persons and reading difficulty of

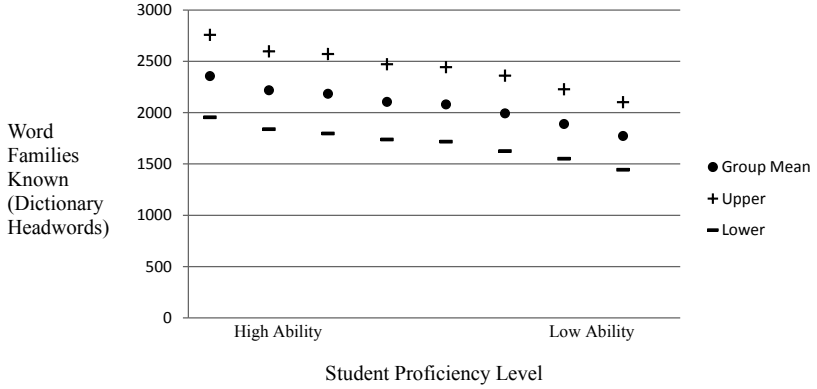

Figure 3 Vocabulary levels by class group for 2011 AEP students. Estimated knowledge of word families is shown for each level of student.

texts requires consideration of both semantic and syntactic factors (Stenner, et al., 2007), so the Lexile Framework for Reading was chosen as the most appropriate measure of reading ability and text difficulty, following Wright and Stenner (1999). The range of scores of AEP students on the TOEFL reading section in 2011 is shown in Table 1, as a mean score and the mean score plus and minus two standard deviations. Following equating guidelines published by the Educational Testing Service (ETS, 2005) and Metametrics (2013), the estimated Lexile level of AEP students is typically between 900 and 1250, with a mean of about 1030. Following Wright and Stenner's (1999, p. 35) chart of grade level equivalents of Lexile measures, AEP students fall into an approximate range from 6th to 12th grade, with an average of about 9th grade. As Metametrics (2013) report, university textbooks typically range from Lexile 1250 to Lexile 1450, meaning that unsimplified university texts are unequivocally too difficult for use in AEP classes. Instead, high-school level readings on familiar topics are needed with the objective of raising lower level students to the 1000 Lexile level and average students to the 1250 Lexile range by program exit.

Having established an approximate measure of students' reading level, provisional objectives for reading classes were drafted, as shown in the Appendix. These were not intended to be definitive, but rather to guide curriculum in an iterative manner, with objectives and curriculum specifications evolving concurrently as more evidence becomes available about the effect of interventions. However, despite the constantly changing nature of objectives during the curriculum development phase, it is crucial to the success of EBP to explicitly define objectives in order that the hypothesized effect of interventions can be compared to the observed effects.

The analysis of textbooks was the most time consuming part of the project. The first consideration in this was the quality of the samples of English, both written and spoken. Although textbooks also contain practice tasks, teachers can

Table 1 FWU Reading Ability

	Mean Score	SD	Mean + 2SD	Mean - 2SD
TOEFL iBT	42.26	5.59	53.45	31.08
Lexile	1030	n.a.	1260	900

supplement these relatively easily, but audio recordings of dialogues and reading passages cannot easily be changed or supplemented, so the quality of these is the primary consideration, with instructional tasks a secondary consideration. The *Reading Explorer* (Douglas & McIntyre, 2009) series of reading textbooks was suggested by a Japanese teacher working in the AEP. Native speaker teachers impressionistically reviewed the book and agreed with Graham's (2012, p. 311) view that

> this book does a superb job of teaching content-based reading skills and presenting and reinforcing vocabulary, its greatest asset in the classroom is that it helps students take an interest in the world and increases their appreciation and understanding of other students' cultures.

To supplement these qualitative impressions, *Reading Explorer* was subjected to a more detailed quantitative analysis, beginning with lexical content. Lexical content was analyzed using the *AntWordProfiler* software package (Anthony, 2011) and separated into Level 1 (BNC 1k), Level 2 (BNC 2k), and Level 3 (BNC 3k, 4k, 5k, and AWL) vocabulary, respectively indicating vocabulary that is essential for any use of English, vocabulary necessary for minimal survival, and vocabulary necessary for participation in academic life. Figure 4 shows that *Reading Explorer 1, Reading Explorer 2*, and *Reading Explorer 3*, will expose students to approximately 1250 words from the AWL and 3k to 5k word families indicating that the semantic level of these books was well matched to the needs of AEP students.

The syntactic match between AEP students and *Reading Explorer* was investigated using the Lexile level of samples of text from each unit of *Reading Explorer 1, Reading Explorer 2*, and *Reading Explorer 3*, as shown in Figure 5. From this it can be seen that *Reading Explorer1* has a range of text difficulty suitable for low and average ability AEP students, while *Reading Explorer 2* and *Reading Explorer 3* introduce language at a level comparable to introductory university textbooks. Therefore it was recommended that first semester reading classes (AR1 and AR2) use *Reading Explorer1* as an introduction to academic reading, while second semester (AR3 and AR4) and third semester (AR5) classes respectively use

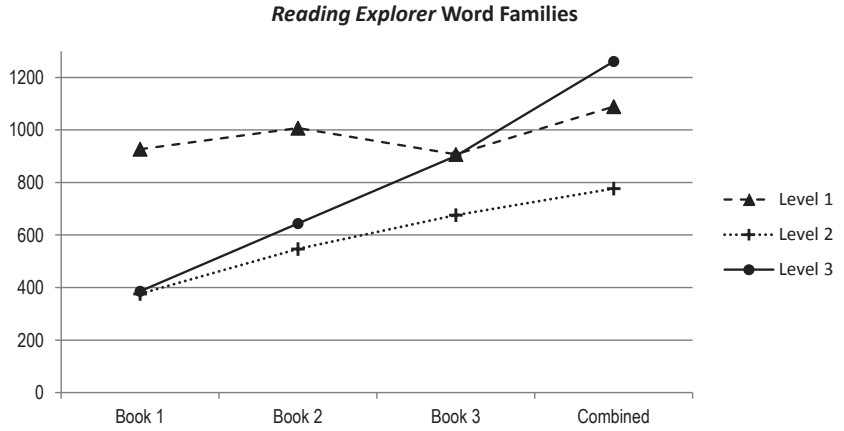

Figure 4 **The number of word families for *Reading Explorer 1, Reading Explorer 2*, and *Reading Explorer 3*. After completing all three books, students would have been exposed to approximately 1250 words from the Academic Word List and 3000 to 5000 level word families.**

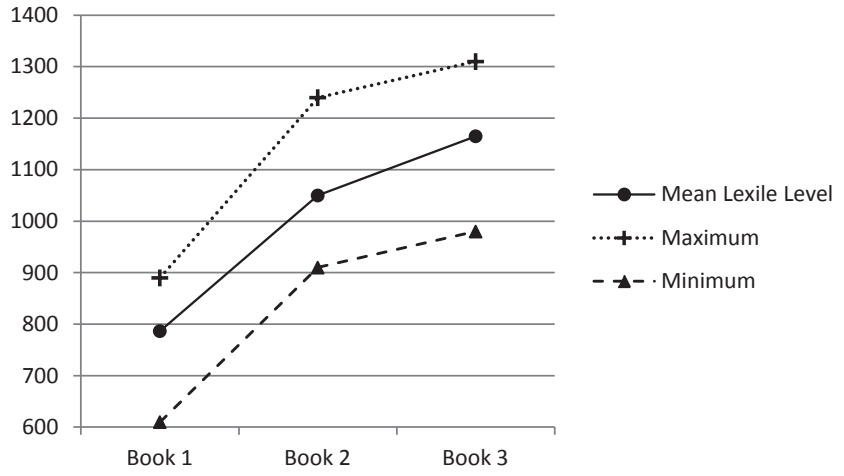

Figure 5 **The *Lexile* level of *Reading Explorer 1, Reading Explorer 2,* and *Reading Explorer 3*. Students who completed all three books would have been exposed to texts ranging from high-school level to introductory university level textbooks.**

Reading Explorer 2 and *Reading Explorer 3* for intensive practice with academic reading.

Reading Explorer was introduced as a recommended textbook in 2012, although not adopted by all teachers. Comparison of TOEFL results between the second semester of 2011 and 2012 is shown in Figure 6. (It should be noted that, although Figure 1 and Figure 6 both show 2011 students, Figure 1 compares the score distribution of the 2011 cohort, while Figure 6 compares reported score gains between the 2011 and 2012 cohorts. Thus the low, mid, and high groups shown in Figure 6 are not directly comparable to the groups shown in Figure 1.) In this case, RTM was adjusted for by estimating true pre-test scores from

$$X_t = X + r(\bar{X} - X) \qquad (1)$$

where X_t is the true pre-test score, X is the observed pre-test score, $\bar{X}$ is the sample mean of observed pre-test scores, and r is the Pearson moment correlation between the pre-test and post-test scores.

Figure 6 shows that low-level students made modest raw gains in 2011, but this was entirely an artifact of RTM and that only high-level students showed true gains after adjustment for RTM. The 2012 results showed marked improvement, with substantive gains observed at all levels after adjustment for RTM. In addition

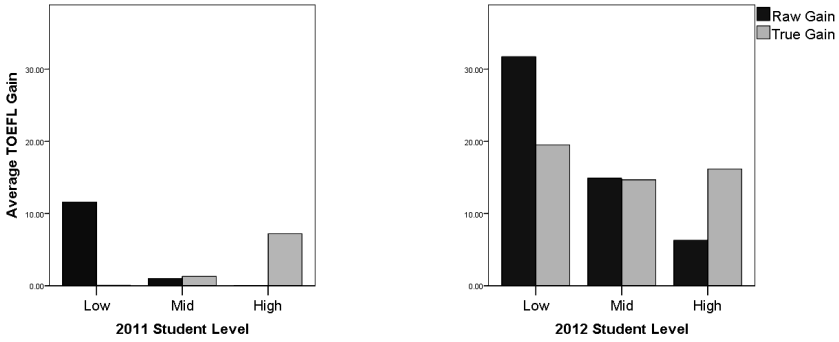

Figure 6 TOEFL score gains for 2011 and 2012.
The mean TOEFL gains for the second semesters of 2011 and 2012 are shown for low, mid, and high-level AEP students as both raw gains and true gains adjusted for regression to the mean. In 2011, higher level students showed small improvements but lower level students did not. In 2012, after curriculum revision, all levels of student showed substantive true gains.

to the enormous improvement seen in low-level students, high-level students also showed improved gains in 2012. An independent samples *t*-test between the 2011 pre-test scores ($M = 439.23$, $S.D. = 34.05$, $n = 230$) and the 2012 pre-test scores ($M = 436.08$, $S.D. = 36.32$, $n = 240$) did not find statistical or substantively significant differences ($t(468) = 0.971$, $p = .33$.) This evidence supports the view that matching textbooks to student ability levels improved learning across all ability levels rather than improving instruction for low-level students at the expense of high-level students.

Conclusions and Summary

This study serves as a case study of implementing EBP in curriculum planning. Multiple sources of information were combined in the analysis of 2011 results, providing an evidential basis of weaknesses in AEP instruction. After adjustment for regression, test scores indicated that instruction for low and mid-level students was ineffective and an attitude to school survey indicated that student engagement declined markedly. Comparison of students' vocabulary level and TOEFL scores indicated that their English comprehension was insufficient to read typical university textbooks, making simplified texts necessary to match students' coping ability. This led to writing of provisional reading objectives to guide textbook selection, followed by quantitative analysis of the lexical and syntactic burden of candidate textbooks. The analysis of *Reading Explorer* showed it to be well matched to student ability level so it was adopted provisionally as the recommended reading textbook. The final step in the EBP process was analysis of outcomes, with substantive improvement in TOEFL scores providing supporting evidence that the revised course objectives and textbook were successful. In particular, low and mid-level students showed dramatic improvements between 2011 and 2012, with high-level students also showing substantive improvement. Although much work remains in improving the coherence of the AEP curriculum, the process and results presented here illustrate how EBP can guide curriculum development and allow evaluation of the success or otherwise of pedagogical interventions.

References

Anthony, L. (2011). AntWordProfiler. Retrieved from http://www.antlab.sci.waseda.ac.jp/antword-profiler_index.html 15 August, 2011.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Cheng, S.-T., & Chan, A. C. M. (2003). The development of a brief beasure of school attitude. *Educational and Psychological Measurement, 63*(6), 1060-1070. doi: 10.1177/0013164403251334

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

Douglas, N., & McIntyre, P. (2009). *Reading explorer*. Boston: Heinle.

ETS. (2005). TOEFL internet-based test: Score comparison tables: Educational Testing Service.

Graham, M. (2012). Book and materials reviews: Reading Explorer. *TESOL Journal, 3*(2), 310-311. doi: 10.1002/tesj.20

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Inoue, N. (2006). What's going on inside the pine tower of babel: Foreign language curriculum reform in a Japanese university. *Languages and Cultures Series, 16*, 87-115.

Metametrics. (2013). Improve your reading ability with Lexile measures Retrieved 15 August, 2013, from http://www.lexile.com/toefl/

Saville, B. K. (2009). Using evidence-based teaching methods to improve education Retrieved 9 August, 2013, from https://tle.wisc.edu/node/1045

Schumann, J. H., & Wood, L. A. (2004). The neurobiology of motivation. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert & L. A. Wood (Eds.), *The neurobiology of learning*. (pp. 23-42). London: Lawrence Erlbaum Associates.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). The Lexile Framework for Reading technical report. Durham, NC: MetaMetrics Inc.

Swinton, S. S. (1983). A manual for assessing language growth in instructional settings. Princeton: Educational Testing Service.

Wright, B. D., & Stenner, A. J. (1999). One fish, two fish: Rasch measures reading best. *Popular Measurement, 1*, 34-38.

Appendix

Provisional Academic Reading Objectives

Academic Reading 1

The objective of this course is to improve students' reading fluency through extensive reading activities and development of sight-word vocabulary. Extensive reading homework of graded readers will be monitored by internet based reports. In-class activities will focus on improving

reading speed and improving sight-word recognition of foundational vocabulary. At the end of the course, all students are expected to demonstrate a minimum reading speed of 100 words per minute with foundational vocabulary.

Academic Reading 2

The objective of this course is to improve students' reading comprehension of simple academic texts through development of academic vocabulary and derivational affixes. Classroom tasks will focus on practicing skills and strategies that underlie reading for comprehension. High-frequency vocabulary will be reviewed and instruction will focus on approximately 250 key academic word families, and derivational affixes commonly associated with them. At the end of this course, students are expected to have minimum vocabulary sizes of 2500 word families, including all foundational vocabulary, 80% of high-frequency vocabulary, and 50% of academic word list vocabulary.

Academic Reading 3

This course will focus on reading fluency and sight-word vocabulary, following on from AR1. Extensive reading homework of graded readers will be monitored by internet based reports. In-class activities will focus on improving reading speed and improving sight-word recognition of high-frequency vocabulary. At the end of the course, all students are expected to demonstrate a minimum reading speed of 150 words per minute with foundational vocabulary and 100 words per minute with high-frequency vocabulary.

Academic Reading 4

This course will aim to improve students' reading comprehension of academic texts through development of academic vocabulary and derivational affixes, following on from AR2. Classroom tasks will focus on practicing skills and strategies that underlie reading for comprehension. Instruction will focus on approximately 250 key academic word families, and derivational affixes commonly associated with them. At the end of this course, students are expected to have minimum vocabulary sizes of 3000 word families, including all foundational and high-frequency vocabulary and 80% of academic word list vocabulary.

Academic Reading 5

This course will focus on reading for academic research. Students will practice skimming to identify texts relevant to their academic interests, and then scanning for relevant details, followed by note-taking and summarizing.