Adapting Levels 1 and 2 of Kirkpatrick's Model of Training Evaluation to Examine the Effectiveness of a Tertiary-Level Writing Course

Abstract

This study adapts Levels 1 and 2 of Kirkpatrick's model of training evaluation to evaluate learning outcomes of an English as a second language (ESL) paragraph writing course offered by a major Asian university. The study uses a combination of surveys and writing tests administered at the beginning and end of the course. The survey evaluated changes in students' perception of their skills, attitude, and knowledge (SAK), and the writing tests measured their writing ability. Rasch measurement was applied to examine the psychometric validity of the instruments. The measured abilities were successively subjected to path modeling to evaluate Levels 1 and 2 of the model. The students reported that the module was enjoyable and useful. In addition, their self-perceived level of skills and knowledge developed across time alongside their writing scores but their attitude remained unchanged. Limitations of Kirkpatrick's model as well as lack of solid frameworks for evaluating educational effectiveness in applied linguistics are discussed.

Keywords: attitude; Kirkpatrick's model; knowledge; path model; Rasch measurement; skills; writing

Adapting Levels 1 and 2 of Kirkpatrick's Model of Training Evaluation to Examine the Effectiveness of a Tertiary-Level Writing Course

Introduction

Evaluation of educational impact and development provides valuable feedback to stakeholders including educational institutions, students, and parents (Praslova, 2010). To date, most English as a second language (ESL) studies have been restricted to developmental patterns of ESL learners (e.g., Storch & Tapper, 2009) or the acquisition of certain grammatical rules or vocabulary (e.g., Shintani & Ellis, 2013). Far too little attention has been given to criteria of educational impact specifically in ESL writing.

To assess educational (and training) impact, researchers have attempted to adapt and adopt models from studies in healthcare, marketing, and business enterprises (Arthur, Tubre, Paul, & Edens, 2003a). Kirkpatrick's (1959, 1996) four-level model is one such framework that provides detailed guidelines for evaluating the effectiveness of training. In educational settings, the four levels of the model seek to measure students' reactions toward the program (Level 1), their learning (Level 2), behavior change (Level 3), and final outcome for the institutions (Level 4). Whereas Levels 1 and 2 are relatively straightforward to assess (Praslova, 2010), Levels 3 and 4 have proven to be extremely difficult to examine in educational studies. The reason is that Level 1 and 2 measurements can be implemented while students have not exited the program yet, but Levels 3 and 4 do require post-hoc studies and tracking students' performance in the field. Of greater concern, Level 4 requires the systematic examination of the effect of the educational program on the "growth" and "success" of the educational departments, for assessing which ESL literature has articulated/reported no framework. Therefore, it would be more efficient to study levels individually (Praslova, 2010). In the present study, Levels 1 and 2 are examined within an ESL academic writing course, and suggestions for researching Levels 3 and 4 are provided. The goal of the writing course is to help students develop academic writing skills over 12 weeks of instructions.

According to Bitchener, Young, and Cameron (2005), instructions and continuous feedback can help ESL students develop second language (L2) writing skills and writing style. An example of developmental writing research supporting this claim is a study conducted by Storch and Tapper (2009). In the study, it was reported that a 12-week writing module operating on effective instructions and continuous feedback led to noticeable improvements in L2 writers' accuracy, lexico-grammatical knowledge, and coherence. Such findings are consistent with results of studies conducted by, for example, Bitchener et al. (2005), Storch and Tapper (2000), and Polio, Fleck, and Leder (1998).

In order to maximize the effectiveness of writing courses, it is suggested by Elliot and Klobucar (2013) that the development of writing skills over time is a research focus worth investigating for the purposes of informing teaching methods which are better tailored to pinpoint students' problem areas, and resolving program issues with regard to continuous assessment (CA).

However, ESL writing development does not garner much research focus on the effectiveness of educational programs. In order to explore the effectiveness of writing programs, further research efforts in this field must be made. This study seeks to assess the impact of a tertiary level paragraph writing course, using Kirkpatrick's four-level model of training as the criteria (Kirkpatrick, 1959, 1996). The central questions of the study are as follows:

(1) What are students' reactions to the program?

(2) How do students' perception of their skills, attitude, and knowledge (SAK) develop over the course 12 weeks in the paragraph writing program? How do their writing skills develop?

To answer these questions, the psychometric quality of the items, writing scales, and dimensions they measure must first be ascertained. Accordingly, two psychometric models (Rasch-Andrich Rasch model and many-facet Rasch measurement) were initially applied to analyze the psychometric features of the items, dimensions, and student scores (see the sections *Rasch-Andrich Rating Scale Model (RSM) for RSAK and Norming Session* and *Many-Facet Rasch Measurement (MFRM)* for further information).

Literature Review

Kirkpatrick's Model of Training Evaluation

Kirkpatrick's model is a highly influential framework for evaluating training programs. Although the model has been developed primarily for evaluating training in business, its application to higher education programs has achieved acceptable success (Praslova, 2010). In addition, one has to consider the challenges and difficulties of transferring assessment models across different disciplines and different institutions; however, it is plausible that the model facilitates the assessment of the traits and skills that are instrumental in academic environments and helps provide feedback to students, teachers, and institutions of higher education (Praslova, 2010).

The model comprises four levels: reaction, learning, behavior and results, which are discussed below.

Level 1: Reaction

The basic level of Kirkpatrick's model of training evaluation measures students' reactions to the program. This level can help curriculum designers and teachers understand whether the program was received well by students (Arthur et al., 2003a); it is highly desirable that students perceive the training experience as valuable and effective, and the topics and lessons as relevant and important (Arthur, Bennett, Edens, & Bell, 2003b).

Alliger, Tannenbaum, Bennett, Traver, and Shotland (1997) proposed two concepts/dimensions based on which to measure students' reactions affective dimension, i.e., how much students feel they benefitted from and enjoyed the program, and utility dimension i.e. students' judgements of how much they have gained from the program. These dimensions are measured by self-appraisal surveys in higher education institutions (Arthur et al., 2003a, b). Indeed, students enjoy learning when they are successful and, if they perceive that they are improving, positive perceptions would motivate them (Centra & Gaubatz, 2005). Improvement in their scores also assures students that they can pass the course and motivates them to score better, not just scrape by. Conversely, a lack of discernable improvement could mean either that the teacher's techniques are ineffective or that the students are failing to comprehend or correctly apply what has been taught, which is again demotivating to the students. Under such circumstances, students could blame the teacher or the educational program and find the course unhelpful to their learning. As a result, examining students' attitudes and reactions toward the course should be an integral part of program evaluation (Arthur et al., 2003a, b), as these variables play significant parts in student development.

Relatedly, affective and utility dimensions of reaction render it highly similar to attitude, which is a behavioral component (Level 3). Attitude theories universally agree on three primary dimensions underlying attitude: cognition or individuals' knowledge, affection or individuals' feelings, and behavior or individuals' reactions to objects and/or people (Van Buren & Erskine, 2002). Therefore, assessments of Levels 1 and 2 seem to have a great deal in common.

Level 2: Learning

The second level consists of using psychometric measures (e.g., pre- and post-tests) to evaluate learning outcomes. Measurement at this level taps into students' development in skills, knowledge, and attitude. Multiple methods are applicable at this level, including tests, self-appraisal instruments (inventories and surveys), and interviews (Kirkpatrick, 1996).

Arthur et al. (2003a) argue that the magnitude of learning can be determined by the effect size *ds*—the index of strength of the outcome computed in, for example, *t*-tests. In addition, learning can be examined using growth curve models which provide both magnitude and slope of growth (Duncan, Duncan, & Strychker, 2006). Levels 3 (Behavior) and 4 (Results) are not discussed here since they are not tested in the present study. Interested readers are referred to Devereaux and Yusuf (2003) and Kirkpatrick (1996).

Application of Kirkpatrick's Model for Language Learning

Two primary sets of theories have emerged from the field of language learning: cognitive-based models and task-based models (Chapelle, Jamieson, & Enright, 2008). The former views language learning in terms of the cognitive processes of learners, such as perception and attitude (Aryadoust, Mehran, & Alizadeh, 2016), use of working memory (Baddeley, 2003), and mental lexicon (Atkins & Baddeley, 1998), whereas the latter draws on the performance of language learners in different contexts to predict learning patterns (Chapelle et al., 2008). Kirkpatrick's model lends itself to both styles of studying language learning and theory building and improves on existing methodological approaches. Levels 1 and 2 of Kirkpatrick's model tap into learners' perceptions and attitudes and help measure their cognitive and linguistic development through psychometric and growth modeling (Arthur et al., 2003a).

Methodologically, Levels 1 and 2 of Kirkpatrick's model constitute an improvement over the available approaches to investigating the (measures of) effectiveness of educational programs. According to Seidel and Shavelson (2007), the extant approaches to examining teaching effectiveness comprise (1) survey-based methods that address methodological and statistical limitations of research, and (2) "cognitive psychological" approaches where experiments and quasi-experiments are used to explore the impact of educational programs.

Survey-based methods apply, for example, self-appraisals at the time of the study, overlooking the development/growth of students across time (Rowan, Correnti, & Miller, 2002). On the other hand, (quasi-) experimental studies focus primarily on the learning outcomes of educational programs, failing to examine the development of relevant cognitive variables (Seidel & Shavelson, 2007) including changes in students' attitude toward educational programs, their reactions toward the program, and evaluation of their own development, which are addressed in survey-based research. Although (quasi-) experimental studies are highly useful, they are rarely conducted.

Levels 1 and 2 of Kirkpatrick's model are particularly well-suited to merge the two aforementioned research methods. This would result in a comprehensive study design that encompasses both "distal" measures of learning (e.g., attitude, reaction, and self-assessment of knowledge) and "proximal" measures (e.g., writing tests) (Seidel & Shavelson, 2007). In many academic institutes—including the university where the present study was conducted—the effectiveness of educational programs is (partly) measured by examining students' reactions toward the program at the end of the course; the present study adapts this approach where students' pre- and post-course writing proficiency, attitude, reactions, and self-assessed knowledge are measured and compared, thereby making a methodological contribution to the current discussion of effectiveness of pedagogy.

ESL Writing Instruction and Development

This section aims to contextualize the present study within the field of writing instruction. To do so, *pedagogy, feedback*, and *technology* will be discussed (Zhang, Yan, & X. Liu, 2015). Pedagogical research has three main foci: task-based instructions and (quasi-) experiments (de Oliveira & Lan, 2014; Schoonen et al., 2011; Storch; 2009), assessment for learning (AFL; Aryadoust, 2014; Lee & Coniam, 2013; S. Liu & Kunnan, 2016), and teaching academic lexicon and grammar (Coxhead & Byrd, 2007). Regarding feedback, a substantial amount of research shows that providing timely corrective teacher and peer feedback to ESL writers can significantly improve their writing skills (see Q. Liu & Brown, 2015, for a review), although some work has cast doubts on its benefits (see below). Lastly, the incorporation of new technologies in writing pedagogy and assessment has received significant attention in recent years (Li & Kim, 2016; S. Liu & Kunnan, 2016). These topics are further discussed below.

Pedagogical Research

Task-based instructions and (quasi-)experiments. Task-based instructions and (quasi-) experiments constitute the first major type of pedagogical research. In a case study, de Oliveira and Lan (2014) provided assistance to a writing instructor to scaffold her instructions and enhance the assigned writing tasks. The researchers reported that the student under assessment developed the ability to apply scientific lexicon in a fairly effective way. This study was limited in terms of the sample size (*n*=1), but the results might hold in a larger sample of students who possess similar attributes.

Research into (quasi-) experiments and ESL writing development shows that several writing components can be nurtured with well-designed educational programs. While tertiary students' writing skills development has been studied over periods between 10 to 15 weeks (Elliot & Klobucar, 2013; Storch & Tapper, 2009), younger adults' writing skills development has been studied over longer periods of time (Schoonen et al., 2011). In a control study by Storch (2009), the writing skills of L2 learners enrolled in an Australian university were observed over 10 weeks of receiving no instructions (teacher's directives) and feedback, with the objective of studying the effect of "pedagogical intervention." It was observed that the learners' writing skills showed no improvement; this finding is similar to the results of Hinkel's (2003) study of L2 American university students.

The study by Storch (2009) also suggests that the provision of instructions in university writing courses significantly benefits learners' development of L2 writing subskills such as grammar, vocabulary, genre knowledge and coherence. This is further reinforced by Storch and Tapper's (2009) 12-week developmental quasi-experiment conducted on L2 writers' skills such as accuracy, coherence, and lexico-grammatical knowledge, in which it was observed that continual and effective instructions led to significant positive effects on students' skills development. Other supporting examples are given by Bitchener et al. (2005), Polio et al. (1998), and Storch and Tapper (2000) where it is shown that tertiary students' writing abilities show progress over the course of an academic semester.

Without such instructional support in university programs, learners will be unable to reap such benefits. It has also been shown that writing skills develop unevenly, and that educational research is still underevaluating important L2 writing skills. Bae and Lee (2012), for instance, found that Korean students' grammar, coherence, and punctuation skills greatly improved; however, there was negligible progress in task fulfilment, an area in writing research that has received little attention.

Assessment for learning (*AFL*). According to Brown (2004), language teachers spend a large amount of time evaluating their students, and much of this assessment is conducted to promote learning, i.e., AFL. For AFL to be useful, it should be aimed at the right level and aid learners to refine their prior knowledge; it should actively involve the learners; it should consider the pedagogical goals and impart knowledge of the criteria and requirements to the learners; and, where applicable, it should incorporate self- and peer assessments to advance learners' awareness of writing rhetoric (Jones, 2010). Although some researchers have argued that AFL should be an integral part of writing pedagogy (Aryadoust, 2014), there is a dearth of research in this area, which has been attributed to the excessive attention given to validation and reliability research (Lee & Coniam, 2013).

Lee and Coniam (2013) conducted an AFL study of writing that yielded mixed results. In the quantitative section, Lee and Coniam used many-facet Rasch measurement to validate the test results and reported some success, which they attributed to "students' better understanding of the requirements of writing (an outcome of the teachers' attempt at AFL)" (p. 42). However, students' motivation to write showed no statistically significant improvement across time—nevertheless, qualitative data indicated that the students favoured their experience with AFL.

Lee and Coniam's (2013) study is similar to the present study in several important ways: both studies are conducted in Asian contexts; both apply multiple techniques, such as tests and questionnaires, to collect data; and both use quantitative data analysis to examine the research questions. However, some important concerns surrounding AFL, such as motivation and related cognitive traits, were overlooked in Lee and Coniam's study; cognitive traits, such as attitude and reactions, are discussed in the current study. In addition, the definition of motivation and related traits is somewhat unclear in Lee and Coniam's study, warranting a better treatment of these concepts.

Teaching academic lexicon and grammar. Academic writing, like other genres, is characterized by specific lexical and grammatical features (Biber, Conrad, & Cortes, 2004). According to Coxhead and Byrd (2007), vocabulary knowledge in academic writing is different from everyday word use and comprises accurate spelling, expression of meaning, proper grammatical structures, synonyms and antonyms, etymology and word families, and level of formality. The grammatical structures commonly used in academic writing are distinct from everyday English and are characterized by lengthier structures that include passive voice and certain adverbial phrases (Raimes, 2004). Due to these complexities, English native speakers begin to acquire the grammatical and vocabulary knowledge for academic writing only when they receive explicit lessons (mainly at tertiary level institutes) (Laufer, 2005). Once learners begin to learn about the academic genre, it is possible to discern considerable variations in their vocabulary and grammatical knowledge (Coxhead & Byrd, 2007).

To learn academic vocabulary effectively, learners should be given opportunities to read and practice the vocabulary items in both reading and writing and be encouraged to use them (Coxhead & Byrd, 2007). Teachers should teach the meaning and use of academic vocabulary and provide ample examples of how vocabulary items are spelled, punctuated, and weaved together with other vocabulary items to develop clauses and paragraphs (see Lynch & Anderson, 2013). Students are often unaware of, for example, collocations, idioms, concordances, lexical bundles, certain functional words (e.g., *the*), and recurrent vocabulary strings in academic writing (Biber, Conrad, & Reppen, 1998), and teachers should aid them in recognizing and using these structures accurately. Involving students in learning and using these structures in writing, coupled with continuous assessment and monitoring of progress, can maximize students' achievements and prepare them to become fluent writers.

Feedback

Written and oral corrective feedback can be an effective tool to assist learners in improving their writing skills (Biber, Nekrasova, & Horn, 2011), although some concerns have been raised over its role and use by Bruton (2010) and Ferris (2004). According to Van Beuningen (2010), academic writing teachers should not limit their feedback to a narrow range of issues such as lexicogrammatical errors; to be effective, feedback should be provided by experts (e.g., teachers or researchers; Q. Liu & Brown, 2015) and promote students' awareness of the form, content, meaning, clarity, and discourse of academic writing (Xu, 2009) by demanding revisions and multiple drafts.

Several types of feedback have been discussed in the literature such as direct and indirect correction, error coding, general oral/written comments, and "unifocused" and "multifocused" comments (Liu & Brown, 2015). While these feedback types show varying degrees of success in helping students to improve their academic writing skills, it seems that using multiple drafts and tracking the accuracy of students' production plays a more significant part than the specific type of feedback. Indeed, recent research by Knoch, Rouhshad, Oon, and Storch (2015) has shown that undergraduate ESL students made no significant progress in their lexico-grammatical accuracy and complexity or holistic writing scores after spending three years at an Australian university because they received minimal feedback from their professors. In sum, it is important that teachers determine the type of feedback best suited for the context, provide regular feedback, and follow up on learners' progress over time.

Technology

Research on the effect of technology on learners' development in academic writing has yielded mixed results. The inconclusive results are attributed to the type of technology being used and its purpose. One emerging research stream examines the accuracy and precision of feedback provided by electronic raters, such as *WriteToLearn* and *Criterion®*, in assessment contexts. The majority of these studies cast doubt on the accuracy of electronic raters (Lavolette, Polio, & Kahng, 2015), showing that they have difficulty identifying errors made by non-native speakers of English (S. Liu & Kunnan, 2016; Dikli & Bleyle, 2014) and that the accuracy of the feedback provided varies significantly across error types (Lavolette et al., 2015). In other studies, researchers found that the feedback generated by electronic raters tends to be generic, redundant, and fairly lengthy (Dikli & Bleyle, 2014; Hoang & Kunnan, in press).

Research into the application of non-assessment technologies has reported more promising results. For example, it has been shown that Web 2.0 technologies can promote collaboration and peer discussion, resulting in higher quality writing in terms of content and organization (Strobl, 2015). Similarly, Lan, Sung, Cheng, and Chang (2015) found supporting evidence for the role of computers in collaborative and pre-writing activities in young ESL learners' writing skills, which is consistent with the results of Lee's (2013) study of collaborative writing and concept mapping. In addition, real-time discussion/chat forums (e.g., Google Chat, Yahoo Messenger) and mobile applications (e.g., WhatsApp, Skype, Viber) facilitate “connectivity” and “availability” and promote student-teacher communication (Bouhnik & Deshen, 2014; Trenkov, 2014). By carefully choosing the proper technologies for teaching academic writing, teachers and students alike can benefit from the advances in this field.

The above survey of the literature has identified some of the major components and procedures involved in the development of ESL students’ academic writing skills. The survey suggests that the inclusion of task-based writing lessons, grammar and lexicon lessons, AFL, continuous teacher feedback, and proper use of technology can significantly facilitate students’ writing development. To situate the present study in relation to this body of literature, the identified components are discussed in the context of the paragraph writing module where the present study was performed.

The Paragraph Writing Module

The module was carried out in two-hour sessions, twice a week, over 12 weeks, and covered several components including producing effective outlines, research, and oral presentations. The module is compulsory for students with low academic writing skills (writing ability is tested using an integrated writing exam when students enter the university, and low ability students are placed in this module) and it aims to enhance their writing skills, alongside grammar and vocabulary knowledge, through various activities and assignments. By the end of the module, students are expected to be able to write coherent paragraphs about academic subjects, read and understand pertinent texts, and have expanded their grammatical and vocabulary knowledge. In the present study, students’ development was measured by comparing their skills when they entered the course and when they completed it, allowing for a pre-test/post-test design (see *Pre- and Post-Course Writing Tests* for further information). Due to spatial constraints, only the five main components of the module are discussed below:

(a) *Explicit writing lessons.* Tutors start each of these lessons with an explanation of the content to be covered, then introduce a model for illustration purposes and allow students to practice the learned content by creating links between reading and writing. Students are given guidance and support until they are able to write independently.

For instance, in a lesson covering the accurate use of English grammar, and coherence (e.g., connectors and anaphora), the tutor starts off by explaining these topics and emphasizing concepts such as unity and organization, as well as the structures of topic, support and conclusion statements. The tutor then provides students with multiple texts of different lengths and discusses the underlying structures of these texts with the class. Students are then asked to participate by scanning the texts to identify and edit incoherent or poorly organized passages.

(b) *Grammar and vocabulary lessons*. Mostly explicit (and sometimes implicit) lessons on grammar and vocabulary designed to improve students' ability to organize and clearly articulate their opinions, are also held. Progress in grammar and vocabulary, the two main elements of writing, has been shown to greatly improve students' overall writing skills (Coxhead, 2012). In these lessons, students are provided with several sentences and paragraphs containing target academic vocabulary.

(c) *Continuous assessment (CA).* The CA approach is taken as it allows tutors to regularly and efficiently assess students' writing skills and hence maintains a conducive classroom environment for learning. Such assessment also helps tutors in catering their teaching methods and materials to students (Le Grange *&* Reddy, 1998). Similar to formative assessment, CA is expected to foster tutor-student interactions and effectively balance both assessment and learning objectives.

(d) *Consultation and teacher feedback.* Students often have various emotive and motivational reasons behind not participating in dialogues with their peers and tutors (Pajares

& Valiante, 1997). To mitigate this, tutors arrange multiple 20-30 minute consultation slots with each student. While assessing for content, organization and language, tutors also provide both written and oral feedback with regard to any stylistic and linguistic issues for each of several drafts of the students' writings.

(e) *Online follow-up practices and discussions*. Some tutors create online chat rooms for students to discuss defined topics. These online discussions are intended to promote a sense of rapport between tutors and students, and guarantee students that their efforts to improve are acknowledged and valued (Deden & Carter, 1996). Tutors also use other web-based services such as email and online messaging (IVLE chat) to keep students updated, and *Turnitin* to collect students' submissions of their individually written paragraph drafts.

Method

Participants

The participants involved in this study were 85 first year university learners (both males and females) enrolled in a basic academic English course at a major Asian university taught by four tutors. The cohort comprised 11 classes and students were aged between 18 to 22 years old and came from various countries, including China, Malaysia, Indonesia, India, Myanmar, Singapore, and Vietnam. The students' majors included business, computer and electrical engineering, geography, social science, and real estate.

Students enrolled into the course had been required to do so based on their results for a qualifying English test. This qualifying English test is a source-based writing placement test made mandatory for undergraduate students admitted into NUS who are not competent in academic English language. Based on their performance in the test, students will either be placed in the basic English module discussed in this study which focuses primarily on paragraph writing skills, or in an English for Academic Purposes (EAP), or exempted from having to take supplementary academic English programs.

Appraising Success of the Writing Program

In the present study, the success of the program was measured in terms of students' growth over time as well as change in their perceptions of the program (Kirkpatrick, 1996). Like commonly applied pre-test/post-test research designs and relevant statistical techniques such as *t*-tests, research into growth measurement investigates learners' performance at certain time points and estimates the progress made (Duncan et al., 2006). The advantage of the growth measurement design used in this study is that it improves on the statistical precision and enables researchers to examine the magnitude of development across time. In addition, growth measurement does not necessarily require a control group to ascertain the growth of the learners (Duncan et al., 2006; Schoonen et al., 2011). For example, most previous growth measurement studies in language learning did not use control groups (e.g., Polio et al., 1998; Schoonen et al., 2011; Storch, 2009); two previous studies in the same context as this work also did not use a control group (AUTHOR, XXXX). Thus, the present study did not include a control group because recruiting a control group is unfeasible in this context. The present study is a longitudinal study with two time points (pre-test and post-test) (Duncan et al., 2006) where learners' growth in grammar and vocabulary, coherence and cohesion, and task fulfilment is regarded as evidence for the success of a language learning program (Kirkpatrick, 1996). Change in learners' perceptions of their knowledge, ability, and attitude toward the course is also treated as evidence for the success of the program (Kirkpatrick, 1996).

Instruments

Reaction, Skills, Attitude, and Knowledge (RSAK) Survey

This questionnaire is modelled after Levels 1 and 2 of the Kirkpatrick model and comprises 37 items measuring four constructs: reaction (four items), students' perception of their skills (15 items), attitude (five items), and knowledge (13 items). The survey questions were developed to tap into the kinds of knowledge and skills taught in the writing program, including linguistic accuracy, organization, and relevance. For example, the item “I know the difference between simple, compound, and complex sentences” engages self-assessed linguistic accuracy, whereas “I know how to organize Cause and Effect paragraphs” measures self-rated knowledge of organization. The clarity of items was checked by two experts with extensive teaching and research experience. Each construct was subjected to Rasch-Andrich RSM separately and item and person fit coefficients and reliability statistics were examined. Data analysis was performed using FACETS computer package, Version 3.71 (Linacre, 2013a) (See Appendix 1).

Writing Tests: Pre- and Post-Course Writing Tests

As earlier noted, Pre- and Post-Course assessments were not originally included in the curriculum and these were added to monitor students’ development in the present study. Four major prompts relevant to the course objectives were developed by the researcher and the course coordinator. The prompts were then submitted to five writing experts with extensive teaching and research background in the field of academic writing for evaluation. An eight-item questionnaire was developed based on Kroll and Reid’s (1994) guidelines for developing unambiguous, clear, and unbiased prompts (see Appendix 2). The experts then used this questionnaire as the basis for rating the prompts and provided written feedback on each prompt. Based on expert judgments, ultimately, two prompts—one expository and one comparison and contrast prompt—were found to be suitable for the Pre- and Post-Course assessments (see Table 1). The solicitation of expert judgments also provides content-related evidence of validity for the prompts (Messick, 1989).

Raters marked the students’ paragraphs by applying a three-component analytical rating scale. The three components are language (i.e., grammar, vocabulary, and mechanics) which comprises 60% of the total mark, organization (i.e., cohesion and coherence) which comprises 20%, and content (i.e., relevance) which comprises the remaining 20%. Students are awarded marks ranging from 1 to 100, which would then be converted to alphabetical grades, F through A+.

Three experienced raters who had taught various academic writing modules were contracted to rate the Pre- and Post-Course assessment scripts in order to distribute the large marking load. However, during the Post-Course assessment phase, one of the raters was unable to continue her duties and was therefore substituted with another rater. The marks were submitted to many-facet Rasch measurement (MFRM) for psychometric quality control. MFRM has been used extensively in validation research to examine the generalizability and construct validity of writing tests and consistency of raters' performances (see Eckes, 2011). Combined with expert judgments, MFRM provides evidence supporting the validity of the uses and interpretations of the scores in the present study.

Data Analysis

Rasch-Andrich Rating Scale Model (RSM) for RSAK

Every student completed the RSAK survey which measures their reactions to the program and perceptions of their SAK. As previously discussed, the RSAK survey evaluates four components (Knowledge, Skill, Attitude, and Reaction), thus resulting in four *dimensions*. It is important to examine the psychometric quality of each dimension before putting it to any use (Bond & Fox, 2015)—i.e., we must determine whether the items measuring each dimension are confounded by irrelevant factors and whether the items enjoy the right endorsibility level and are not too easy or too difficult to endorse. To do so, the RSAK dimensions were separately subjected to RSM and item measures (i.e., how highly or minimally items were endorsed), person measures, infit and outfit mean square (MNSQ) indices, and scoring category difficulties were estimated. These indices basically perform a quality control analysis and, once we are certain that the quality of the item information and person (student) information is reliably high, we can employ student measures to answer the research questions of this study.

Infit MNSQ is an inlier-sensitive index of fit which can capture erratic patterns lying near the estimated ability of the students; outfit MNSQ is an outlier-sensitive fit statistic which can capture perturbations on the extreme points of the scale (i.e., maximum and minimum). For a sample size between 30-250, Bond and Fox (2015) suggest that MNSQ values should range between 0.5-1.5 to indicate good fit; misfits can often be attributed to errors in rating (Engelhard, 2012). Furthermore, scoring category difficulties should monotonically increase; that is, each higher category should be from 1.4 to 5 logits more difficult to choose than its immediate lower category (Bond & Fox, 2015).

It is important to note that psychometric validation of measurement instruments can also be conducted using exploratory and confirmatory factor analysis (EFA/CFA). In the present study, RSM was chosen over EFA/CFA because of the limitations of EFA/CFA and the relative merits of RSM. EFA is applied when there is no pre-specified measurement model (Smith, 1996), unlike the present study where four dimensions are each tested by a discrete instrument. In addition, CFA is more restrictive, requiring a minimum item-to-person ratio of 1 to 20 for stable parameter estimations (Tanaka, 1987), which would be a total of 740 students (20 students × 37 items) in the present study; in contrast, Rasch measurement and RSM can be tested using a sample of just 30 people (Linacre, 1994). RSM also provides item- and person-level diagnostic information (fit statistics), which helps researchers identify the potential causes of anomaly in data, whereas EFA and CFA merely provide global fit statistics.

Norming Session and Many-Facet Rasch Measurement (MFRM)

A norming session was arranged, during which the rating scale was discussed and raters were given the opportunity to put the rating scale to practice by marking four scripts individually. The raters were free to voice their concerns and difficulties encountered with regard to the marking procedure (e.g. the interpretation of levels, cut-off scores, etc.), which were discussed with the coordinator and researcher.

MFRM was used to carry out reliability and validity check and adjust for the effects of facets influencing students' scores such as raters' severity/leniency levels (Engelhard, 2012). Like RSM, which was earlier discussed, MFRM examines the psychometric quality of assessments. The primary difference between these models is that MFRM lends itself to the analysis of data where raters' subjective judgements are a likely source of measurement error, whereas RSM is not suitable for examining such data. An anchoring technique was used in performing MFRM. Anchoring rater and item measures greatly improves the precision of the marks awarded and accounts for the influence of raters on scores (see Linacre, 2013b). The technique was implemented in two stages as follows (a full treatment of this technique falls outside of the scope of this paper; interested readers are referred to Bond and Fox, 2015).

First, 72 common scripts (i.e., scripts to be marked by all raters) that had been randomly selected from the 11 classes (Stage 1) were marked by the raters. The awarded marks were analyzed with MFRM to generate an "anchor file" containing rater severity and scoring category measures. The rater's severity measures in the second round of rating and MFRM were constrained to these anchor estimates (mathematical notation of the model is presented in Appendix 3).

In Stage 2, the raters were tasked with marking between 56 and 116 scripts. As before, the awarded marks were analyzed with MFRM, with the tutors' severity measures anchored to their measures obtained in Stage 1 (Linacre, 20113b). At the end of this round of MFRM analysis, fit statistics are obtained and the afore-mentioned adjustments can be made, allowing fair scores to be computed. Texts exhibiting poor fit (MNSQ values falling outside the range of 0.5-1.5) were read, discussed and re-marked by the researcher and course coordinator.

The Pre- and Post-Course scripts were marked using the aforementioned marking and anchoring methods. The finalized students' fair Pre- and Post-Course scores were subsequently used in the path modelling.

Rasch reliability and separation indices for students and raters were also calculated to find “statistically distinguishable levels of performance” for every facet in the analysis (Linacre, 2013b, p. 293). High student reliability, low rater reliability (an indicator of rater homogeneity) and high separation indices are preferred to ensure high precision measurements (Linacre, 2013b).

Path Modeling

Student measures estimated by MFRM and RSM were used in a series of *t*-tests comparing Pre- and Post-Course data and a path model to examine the effectiveness of the module for the students. Path model analysis was performed on AMOS computer package. Path modeling is an extension of regression modeling where the effect of multiple independent (exogenous) variables on multiple dependent (endogenous) variables can be estimated. Figure 1 consists of students’ Pre- and Post-Course SAK as well as fair scores. Small circles represent errors terms estimated only for endogenous variables. Arrows running between exogenous and endogenous variables represent their relationship which is quantified by regression coefficients. Low regression coefficients indicate lack of consistency between Pre- and Post-Course data and thus change in data over time. In addition, the four dimensions of RSAK in Figure 1 are intimately related according to the theoretical framework reviewed earlier. The correlation paths (two-headed arrows) capture this interconnectedness (Duncan et al., 2006). The same interconnectedness also exists among the Post-SAK measures, but this relationship is statistically captured by correlating the residuals in the exogenous variables rather than the variables themselves (rectangles) (Duncan et al., 2006). Therefore, the presence of the two-headed arrows in the Pre -SAK variables and Post-SAK residuals points to the presence of the same phenomenon—interconnectedness.

To estimate the magnitude of change, *t*-tests were conducted (Duncan et al., 2006). Because some students’ answers on SAK were missing, means and intercepts of the data were estimated—this is a requirement in estimating path models with missing data (Schumacker & Lomax, 2010).

Results

Answering the Research Questions

The first research question is concerned with students' perceptions and was answered via RSM analysis. The second research question investigates students' development in SAK and writing skills; to address this question, students' perceptions of their SAK at Pre- and Post-Course time points were independently subjected to RMS analysis and their linearized measures were recorded. Next, students' Pre- and Post-Course writing performances were validated using MFRM and their Fair scores were recorded. Finally, their SAK and Fair scores were used in the aforementioned path model. In the following sections, I present the results of the preliminary psychometric analysis of the Knowledge, Attitude, and Skill dimensions (*Preliminary Psychometric Analysis*) followed by the analysis pertinent to the research questions of the study.

Preliminary Psychometric Analysis

Table 2 presents the psychometric features of the Knowledge, Attitude, and Skill dimensions (due to space constraints, students' statistics are not presented). As can be seen, there is a fairly wide range of item endorsibility across the three dimensions, indicating that the items have suitable levels of endorsibility—i.e., they are neither too easy nor too difficult to endorse. This is further testified by the item and person reliability and separation indices, which show (approximately) two to three distinct strata of item reliability and person ability across the three dimensions. In addition, the infit MNSQ and outfit MNSQ values fall within the range of 0.5 to 1.5, suggesting a lack of construct-irrelevant factors and acceptable reliability. The only exception is item 15, whose outfit MNSQ values are only slightly outside of the acceptable range. This analysis shows that the data can be reliably used to answer Research Question 2. The psychometric analysis of the Reaction dimension is discussed in the following section.

Research Question 1

Rasch model item and person reliability estimates were .93 (separation=3.67) and .58 (separation=1.17), respectively. The relatively low person reliability is likely due to the number of items, as reliability is sample-dependent (Linacre, 2013b). Items and students all fitted the model reasonably well (infit/outfit MNSQ: 0.67–1.33) (see Table 2), indicating the lack of perturbations and erratic patterns in the data.

Table 3 presents the item fits and endorsements. On average, 80% of students indicated their satisfaction with the module by giving positive evaluations. Items 1-3 are positively worded, but Item 4 is negatively phrased and its statistics should be read reversely. Only one student strongly disagreed with enjoyment, usefulness, and utility of the module (items 1-3), and two expressed their dislike toward attending the classes. By contrast, the majority either agreed or strongly agreed on the joy, usefulness, and utility of taking the course. The most highly-endorsed item is #1, suggesting that students rated the joy of attending the classes most highly. The most lowly-endorsed item is #3, suggesting that students felt they could have spent their time more usefully in class.

Research Question 2

To determine whether students' perceptions of their SAK developed alongside their writing skills over the course of 12 weeks (Research Question 2), the psychometric quality of the instruments was first examined. Subsequently, the estimates were fed into AMOS computer package for examining their possible relationships. I report the results of RSM and MFRM analyses below.

RSM Analysis of the SAK

Table 4 presents Pre- and Post-Course SAK results. The first column gives the name of the SAK dimension alongside the time point. Columns two to seven present infit and outfit MNSQ coefficients followed by item and person reliability and separation columns. Overall, infit/outfit MNSQ data suggests all dimensions fit the model reasonably well, suggesting lack of anomalous data patterns and irrelevant factors affecting students' performance. At least two separation levels for person and items emerged in each dimension, indicating that the item/person measures differentiated at least two levels of person ability and item difficulty (endorsibility).

In addition, there are four columns giving average endorsements (from 1 or strongly disagree to 4 or strongly agree) for each scoring category and their measures in log-odd units (logits). Students' positive evaluations are indicated by the endorsement of categories 3 and 4, and negative evaluations by the endorsement of categories 1 and 2. As there is some missing Post-Course data, the displayed percentages should be used for comparison.

For the Knowledge dimension of SAK, negative evaluations of were endorsed in 16% of the Pre-Course responses, whereas this plummeted to 7% in Post-Course evaluations; furthermore, the proportion of positive evaluations of knowledge increased from 84% in Pre-Course to 93% in Post-Course.

For the Skills dimension of SAK, negative evaluations dropped from 51% Pre-Course to 15% Post-Course, and positive evaluations rose from 49% Pre-Course to 85% Post-Course. Similarly, in the Attitude dimension of SAK, negative evaluations decreased from 47% Pre-Course to 6% Post-Course, while positive evaluations increased from 53% Pre-Course to 93% Post-Course. This similar trend emerged across all dimensions, suggesting that some students entered the module with low self-appraised knowledge and skills but exited the course with greater confidence about their skills and knowledge. Although their attitudes remain unchanged. The statistical significance of this finding will be examined under *Path Model Analysis.*

Finally, every scoring category has an estimated endorsibility measure in average endorsement columns. For example, the endorsibility of category 1 in Pre- and Post-Course knowledge is -4.19 and -5.24, respectively, indicating that it was relatively easier for the students to lowly rate their Pre-Course than Post-Course knowledge.

MFRM Analysis

The reliability and separation indices for both students and raters and their accompanying fit statistics are compiled in Table 5. Students' fair scores were adjusted to account for differential rater severity. As shown in Table 4, at Pre- and Post-Course time points, students exhibited high reliability indices (.95 and .91) and separation statistics (4.19 and 3.10). The separation statistics suggest there are four and three distinguishable ability levels in Pre-Course and Post-Course data respectively, i.e., the writing tests have consistently distinguished four and three ability levels at Pre- and Post-Course stages. The decrease in number of student ability levels from Pre- to Post-Course suggests that the instructions (teacher's directives) received in the course caused students' abilities to become more homogenous.

In addition, average infit and outfit MNSQ statistics indicate that overall, the data fits the model adequately. A few misfitting cases were identified which were re-examined and re-marked by the module coordinator and researcher to enhance the precision of the scores.

In contrast, as shown in Table 6, raters' reliability statistics are zero, suggesting homogeneous rater severity/leniency levels and high inter-rater reliability. The infit and outfit MNSQ values range between 0.80 and 1.25, suggesting high consistency in terms of marking patterns.

Path Model Analysis

Figure 2 demonstrates the path model of students' SAK and writing skills' development across time. The change in mean scores from Pre- to Post-Course is noticeable: whereas the linearized mean scores of Pre-Course skills and knowledge (estimated by RSM) were -0.040 (*SD*=1.22) and 0.121 (*SD*=1.14), respectively, they increased to 1.28 (*SD*=1.63) and 1.67 (*SD*=2.04) in Post-Course. Similarly, Fair scores' mean index increased from 17.66 (*SD*=2.80) to 18.41 (*SD*=1.50). However, attitude went down from 1.68 (*SD*=1.38) to 1.49 (*SD*=2.01).

The path model fits the data well, as indicated by its fit statistics: χ^2=30; *df*=14; χ^2/*df*=1.522; CFI=0.962; NFI=907; and RMSEA=0.079. Three out of four regression paths from Pre- to Post-Course SAK and Fair score variables are statistically significant: Pre- and Post-Course knowledge coefficient=.14 (p < 0.01); skills coefficient =.29 (p < 0.01); and Fair score coefficient=.41 (p > 0.01). This indicates that, for example, when Pre-Course Fair scores increase by one standard deviation, Post-Course Fair scores go up by 0.41 standard deviations. Therefore, 41% of variance is shared between Pre- and Post-Course Fair scores, whereas 59% of variance of Post-Course remains unexplained by Pre-Course Fair scores. Similarly, Pre- and Post-Course knowledge and skills share only 14% and 29%, leaving a markedly high amount of variance unexplained by Pre-Course data. Attitude regression path is statistically non-significant (p > 0.05).

Four *t*-tests were performed to examine the statistical differences between the mean scores from Pre- to Post-Course. As Table 7 shows, all comparisons except attitude yielded statistically significant p values (p <0.05), indicating that the observed changes are not attributable to chance.

Discussion

This study set out with the aim of investigating the impact of an academic writing module, adapting Kirkpatrick's model of training evaluation. Due to the limitations of the study, only Levels 1 and 2 were examined. Initially, the psychometric quality of the instruments was closely examined using MFRM and RSM, to ascertain the reliability and trustworthiness of the data before using them to answer the research questions. This analysis is highly important, since the analysis of data collected via poorly developed instruments would yield no psychometrically valid results (Bond & Fox, 2015; Engelhard, 2012). The instruments were found to be psychometrically reliable and were therefore used to answer the research questions.

Research Question 1

The first research question addressed students' reactions to the program. The majority of the students reacted positively to the program; they liked attending the classes and believed that the module was joyful, useful, and effective. A possible explanation for this might be the design of the module and its components. While the module focuses primarily on writing skills, it does require students to carry out research on the topics they are expected to write about and present the results to their classmates. These sessions alongside other "fun" educational activities such as discussions, using multimodal equipment such as computers, and the constant access to high speed Internet in class would all make the educational environment more appealing to the students.

However, students also felt that the time of the class could have been spent more usefully. The questionnaire did not contain items soliciting students' suggestions on how to spend the time more practically, but based on students' responses in the semesterly feedback exercise conducted by the university, one suggestion is that some of the in-class grammar lessons and assessments could be conducted online and more time could be allocated to in-class discussions and exchanging teacher and peer feedback. As the university has moved toward multimodal and Internet-based teaching methods, students would expect to be able to perform more activities that recognize learners' autonomy. This also highlights the important role of needs analysis which some researchers applying Kirkpatrick's model have stressed (Wexley & Latham, 2002).

Kirkpatrick’s model provides no guidelines on developing instruments for assessing reactions; it is highly recommended that future research incorporates both open-ended and Likert scale items which target the causes of students’ (dis)satisfaction with the module and their recommendation on how to render the module more joyful and useful.

Research Question 2

The second research question addressed the development of students’ SAK and writing skills over the course of 12 weeks. Students’ Pre- and Post-Course perceptions of their knowledge and skills were assessed using a survey validated through RSM. Two primary points need to be considered in interpreting the results of the survey. First, the survey is a self-appraisal instrument which has no correlation with Pre-Course Fair scores, but correlated weakly with Post-Course Fair scores. This result may be explained by the fact that as students received training during the course, they gradually became aware of their skills and knowledge, so their final Fair scores had a better correlation with their perceptions of their skills and knowledge.

Second, as Kirkpatrick (1996) stated, knowledge and skills are different concepts. Knowledge (in this study) refers to students’ information about language, and skills are the result of applying knowledge. Measuring both constructs is necessary in impact research, as students need to have both. Knowledge of language and writing mechanisms helps when students attend lectures where these concepts are discussed. Knowledge is further necessary when students have in-class discussions and meet their tutors for consultation and receiving feedback; it becomes the “common alphabet” of communication in these contexts. Skills, on the other hand, are the ability to apply knowledge and produce new pieces of writing. Students’ self-appraisal of these two concepts indicates how well they are aware of their weaknesses and strengths.

Another approach to assessing knowledge and skills is giving tests to students. While this can render assessment more objective, a limitation of knowledge tests (e.g., grammar and vocabulary tests) in the present writing course would be that they might not seem to be highly "joyful" and relevant to students' studies. Giving writing tests, however, is a feasible and reliable approach (Engelhard, 2012) which can be bolstered by skills and knowledge surveys.

One unanticipated finding was the lack of change in students' attitude during the course. It is difficult to explain this result, but a possible explanation may be that the items had to be slightly modified to fit the Post-Course survey. For example, a Pre-Course item would read as "I believe I *will* not gain in my English skills in this course.", whereas it had to be changed to "I believe I did not gain in my English skills in this course." This type of modification could affect the constructs in the two time points, though it is the only possible option available in this study. Future studies on the reliability and equality of such items are therefore recommended.

Finally, the path model analysis revealed that through the module, students' writing skills improved greatly. The overall results of the current study concur with the past findings by Storch and Tapper (2009) and Andrew & Romova (2011) that writing programs held over one or two academic semesters incorporating grammatical and vocabulary lessons, writing practices, and continual feedback may improve L2 writers' skills.

Conclusions and Limitations of the Study

This study is one of the fewest that have implemented "on-site" training. In their meta-analysis of the research informed by Kirkpatrick's model, Arthur et al. (2013b) found only one "on-site" study out of 379 studies investigated, arguing that on-site studies lack validated evaluation systems. It is important to initially develop reliable evaluation methods in effectiveness research to achieve the best results.

The individual significance (weighting) of each component of the module requires future research. The relative success of the writing program under evaluation suggests that a combination of the aforementioned facilitative factors would very likely promote learning of academic writing principles. This lends support to the existing research showing that instruction (e.g., de Oliveira & Lan, 2014), AFL (Lee & Coniam, 2013), teaching academic lexicon and grammar (Coxhead & Byrd, 2007), the provision of feedback (Q. Liu & Brown, 2015), and technologies that promote connectivity and perpetual teacher-student interactions (Trenkov, 2014) facilitate the development of learners' academic writing skills. However, the role and significance of each individual factor remains unclear. In addition, whether other cognitive and non-cognitive factors contribute to students' development remains an open question. Future research can address these gaps by (1) developing a clear theory of writing development where the contributing factors are determined and (2) measuring/investigating the impact of each factor.

Survey-based methods are improved by using Levels 1 and 2 (see Seidel & Shavelson, 2007). However, a limitation of the Kirkpatrick's model of training evaluation is that it provides no guidelines to identify the most effective techniques and/or methodologies applied. Additionally, the framework does not require a strict experimental design (Arthur et al., 2003b), which can be a subject of contention. To scholars who judge the value of training only within the experimental design framework, this is a limitation. Nevertheless, numerous writing researchers mentioned in the present study have not applied experimental designs but have reported promising findings (e.g., Polio et al., 1998; Schoonen et al., 2011; Storch, 2009; Storch & Tapper, 2000, 2009). In other words, while quasi-experiments might be affected by factors such as history, maturation, and interaction of selection and treatment, previous research into students' writing development shows that the likelihood of Pre- and Post-Course assessments exerting a meaningful impact on the reliability of assessments is fairly low (e.g., Polio et al., 1998). In academic environments where the effectiveness of certain modules is examined, it is implausible to adhere to the restrictive requirements of experimental designs. Consequences of assigning students to control and experiment groups would dissuade institutions of higher education to adapt such approaches. For example, if a group of students receives more effective training than others, the disadvantaged group who failed to efficiently make use of their time (e.g., one semester) and training would be unsatisfied. Therefore, the institutions will have to provide the effective training method again to this group, which is both expensive and time-consuming.

Like Lee and Coniam's (2013) study in which students' motivation to write showed no statistically significant improvement, students' attitudes in the present study did not improve over time. I suggest that a qualitative data collection technique (e.g., ethnographic observations and interviews) be added to future research. The inclusion of qualitative data would offer at least two benefits: first, it can uncover information that cannot be captured by questionnaires and surveys; and, second, the uncovered information can be used to improve the precision and reliability of the questionnaires.

In the present study, expert judgements were solicited during the development process of the writing prompts and the questionnaires, providing evidence of content validity (Messick, 1989). Interviewing the students and the teachers involved in the study would provide further evidence supporting the validity of the outcome of the study. Future researchers can apply a mixed method approach that benefits from both qualitative and quantitative data analysis techniques.

A point about Levels 3 and 4 is in place. Despite the important implications of the Kirkpatrick's model, difficulties would arise when we adopt the model in short-term studies specifically where first year university learners are continuously assessed. The model would suggest that when they exit the language program, students should be tracked and their performance in the university subject modules which demand effective use of language should be examined. To do so, the language departments should establish coordination with other departments so as to track students' use of their obtained knowledge and skills as well as their experiences with various discourses. Implementing such plans is valuable yet often impossible to execute given the incurred costs and more importantly, the lack of a useful framework in applied linguistics for tracking students. It is necessary to establish cost-efficient methods and theoretically sound frameworks to explore Levels 3 and 4 of the Kirkpatrick's model.

References

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of relations among training criteria. *Personnel Psychology, 50*, 341–358.

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Edens, P. S. (2003a). Teaching effectiveness: the relationship between reaction and learning criteria. *Educational Psychology, 23*, 275–285.

Arthur, W., Jr., Bennett, W. J., Edens, P. S., & Bell, S. T. (2003b). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*, 234–245.

Aryadoust, V. (2014). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology: An International Journal of Experimental Educational Psychology*. DOI: 10.1080/01443410.2014.950946

Aryadoust, V., Mehran, P., & Alizadeh, P. (2016). Validating a computer-assisted language learning attitude instrument used in Iranian EFL context: An evidence-based approach. *Computer Assisted Language Learning Journal, 29*(3), 561-595.

Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics, 19,* 537–552.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36,* 189–208.

Bae, J., & Lee, Y.-S. (2012). Evaluating the development of children's writing ability in an EFL context. *Language Assessment Quarterly, 9*(4), 348-374.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis*. TOEFL iBT Research Report No TOEFLiBT-14.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL students. *Journal of Second Language Writing, 12*(3), 191-205.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bouhnik, D., & Deshen, M. (2014). WhatsApp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education: Research, 13*, 217-231.

Brown, H. (2004). *Language assessment: Principles and classroom practices*. White Plains, NJ: Pearson Education Inc.

Bruton, A. (2010). Another reply to Truscott on error correction: Improved situated designs over statistics. *System, 38,* 491–498.

Centra, J. A., & Gaubatz, N. B. (2000). *Student perceptions of learning and instructional effectiveness in college courses.* Research Report No. 9. The Student Instructional Report II. Princeton, NJ: Educational Testing Service.

Chapelle, C. A., Jamieson, J., & Enright, M. K. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

Coxhead, A. (2012). Academic vocabulary, writing and English for academic purposes: Perspectives from second language learners. *RELC Journal, 43*(1), 137-145.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16,* 129–147.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16,* 129–147.

De Oliveira, L. C., & Lan, S.-W. (2014). Writing science in an upper elementary classroom: A genre-based approach to teaching English language learners. *Journal of Second Language Writing, 25*, 23–39.

Deden, A., & Carter, V. K. (1996). Using technology to enhance students' skills. In E. Jones (ed.), *Preparing competent college graduates: Setting new and higher expectations for student learning. New directions for higher education*, *96*, (pp. 81–92). San Francisco: Jossey-Bass.

Devereaux, P. J., & Yusuf, S. (2003). The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of Internal Medicine*, *254*(2), 105-113.

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1-17.

Duncan, T. E., Duncan, S. C., & Strychker, L. A. (2006). *An introduction to latent variable growth curve modeling.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt, Germany: Peter Lang.

Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* New York: Routledge.

Ferris, D. R. (2004). The ‘‘Grammar correction’’ debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime?). *Journal of Second Language Writing, 13*, 49–62.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37,* 275–301.

Hoang, G., & Kunnan, A. J. (in press). Automated writing instructional tool for English language learners: A case study of MyAccess. *Language Assessment Quarterly.*

Jones, J. (2010). The role of assessment for learning in the management of primary to secondary transition: Implications for language teachers. *Language Learning Journal, 35*(2), 175–191.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training and Development, 13,* 3–9.

Kirkpatrick, D. L. (1996). Invited reaction: Reaction to Holton article. *Human Resource Development Quarterly, 7,* 23–25.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28*, 39–52.

Laufer, B. (2005). Instructed second language vocabulary learning: The fault in the 'default hypothesis'. In A. Housen & M. Pierrard (Eds.), *Investigations in instructed second language acquisition* (pp. 311–329). Berlin: Mouton de Gruyter.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*(2), 50–68.

Le Grange, L., & Reddy, C. (1998). *Continuous Assessment: an introduction and guidelines to implementation.* Cape Town, South Africa: Juta & Co.

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*, 34–50.

Lee, Y. (2013). Collaborative concept mapping as a pre-writing strategy for L2 learning: A Korean application. *International Journal of Information and Education Technology, 3*(2), 254–258.

Linacre, J. M. (2013a). *A user's guide to FACETS Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2013b). Facets [Rasch measurement computer program]. Chicago, IL: Winsteps.com.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing, 30,* 66–81.

Liu, S., & Kunnan, A. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case Study of WriteToLearn. *CALICO Journal, 33*(1), 71-91.

Lynch, T., & Anderson, K. (2013). *Grammar for academic writing*. Edinburgh: English Language Teaching Centre, University of Edinburgh.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education /. Macmillan.

Pajares, F., & Valiante, G. (1997). Influence of writing self-efficacy beliefs on the writing performance of upper elementary students. *Journal of Educational Research, 90,* 353-360.

Polio, C., Fleck, C., & Leder, N. (1998). ''If I only had more time:'' ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7,* 43-68.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in Higher Education. *Educational Assessment, Evaluation, and Accountability, 22,* 215–225.

Raimes, A. (2004). *Grammar troublespots: A guide for student writers.* Cambridge, UK: Cambridge University Press.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Schoonen, R., Gelderen, van A., Stoel, R.D., Hulstijn, J., Glopper, de K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31-79.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499.

Shintani, N., & Ellis, R. (2013). The comparative effect of metalinguistic explanation and direct written corrective feedback on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing, 23*(2), 286-306.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3,* 25-40.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103–118.

Storch, N., & Tapper, J. (2000). Discipline specific academic writing: what content teachers comment on*. Higher Education Research and Development, 19,* 337-356.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes, 8*(3), 207–223.

Strobl, C. (2015). Affordances of Web 2.0 Technologies for Collaborative Advanced Writing in a Foreign Language. *The Computer Assisted Language Instruction Consortium (CALICO), 31*(1), 1-18.

Tanaka, J. S. (1987). “How big is big enough?”: Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134-146.

Trenkov, L. (2014). *Managing teacher-student interaction via WhatsApp platform.* Proceedings of EDULEARN14, 6596–6600.

Van Beuningen, C. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies, 10,* 1–27.

Van Buren, M. E., & Erskine, W. (2002). *The 2002 ASTD state of the industry report.* Alexandria, VA: American Society of Training and Development.

Wexley, K. N., & Latham, G. P. (2002). *Developing and training human resources in organizations* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Xu, C. (2009). Overgeneralization from a narrow focus: A response to Ellis et al. (2008) and Bitchener (2008). *Journal of Second Language Writing, 18*, 270–275.

Zhang, Z., Yan, X., & Liu, X. (2015). The development of EFL writing instruction and research in China: An update from the International Conference on English Language Teaching. *Journal of Second Language Writing, 30*, 14–18.

Appendices

Appendix 1

Dear Student,

· Please fill in the following questionnaire carefully and thoughtfully!
· There are no right or wrong answers so please try to be as accurate as possible in your responses. Your evaluation will assist us in making this module more productive and meaningful. Thank you for your cooperation!

1=strongly disagree. 2=disagree. 3=agree. 4=strongly agree.

Knowledge				
1. I understand how a paragraph in an academic text is organized.	1	2	3	4
2. I know the rules of English grammar very well.	1	2	3	4
3. I know the difference between simple, compound and complex sentences.	1	2	3	4
4. I know the meaning of coherence in a paragraph.	1	2	3	4
5. I know the meaning of unity in a paragraph.	1	2	3	4
6. I know the difference between a dependent and an independent clause.	1	2	3	4
7. I know the different parts of speech.	1	2	3	4
8. I don’t know the rules of English grammar well enough.	1	2	3	4
9. I know what a topic sentence is.	1	2	3	4
10.I know how to develop a paragraph logically.	1	2	3	4
11.I know how to organize Comparison and Contrast paragraphs.	1	2	3	4
12.I know how to organize Cause and Effect paragraphs.	1	2	3	4
13.I know what transition words are.	1	2	3	4
Skill				
14. I can write a well-organized and clear paragraph.	1	2	3	4
15. I can edit and improve the organization of my essays/assignments	1	2	3	4
16. I am able to identify and correct grammar errors in my written work.	1	2	3	4
17. I am able to clearly express my ideas and points of view in an academic setting	1	2	3	4
18. I can engage others in a meaningful discussion in English.	1	2	3	4
19. I can write an effective topic sentence.	1	2	3	4
20. I have no difficulty with choice of words in writing.	1	2	3	4
21. I can write effective specific supporting information for the topic sentence.	1	2	3	4
22. I know how to continue learning English on my own.	1	2	3	4
23. I can effectively connect each idea to the rest of the ideas in the paragraph.	1	2	3	4
24. I can use verb tenses accurately in writing.	1	2	3	4
25. I can use subject-verb agreement accurately in writing.	1	2	3	4
26. I can use prepositions accurately in writing.	1	2	3	4
27. I can use capital letters appropriately in writing.	1	2	3	4
28. I can write a coherent paragraph.	1	2	3	4
Attitude				
29. I’m not sure what I learnt from this course.	1	2	3	4
30. I put in effort to work on improving my English in this course.	1	2	3	4
31. I picked up practical English skills in this course.	1	2	3	4
32. I believe the level of my English skills is sufficient for my needs.	1	2	3	4
33. I believe I did not gain in my English skills in this course.	1	2	3	4
Reaction				
34. I enjoyed attending this English class.	1	2	3	4
35. I believe I learnt useful things in this course.	1	2	3	4
36. I made the best use of my time in this course.	1	2	3	4
37. I disliked attending English classes.				

Appendix 2

The Questionnaire Eliciting Expert Judgments on the Bias, Clarity, and Lack of Ambiguity of the Prompts.

Item		Strongly disagree	Disagree	Agree	Strongly agree
1	The task is sufficiently challenging to discriminate between high-and low-ability students.				
2	The ideas in the tasks are within the experience of the students.				
3	The task is culturally ambiguous.				
4	The task leads students to construe the topic differently than intended.				
5	The task allows for some degree of freedom to show their background knowledge.				
6	The task is understandable to low-ability readers.				
7	Students can address the task in the time frame.				
8	The task specifies the rhetorical properties of the response (e.g., comparison & contrast).				
Further Comments?					

Appendix 3

MFRM Formula

The MFRM is expressed as follows:

$$\log \frac{p_{nijk}}{p_{nijk-1}} = B_n - D_i - C_j - E_h$$

, where

P_{nijk} is the probability of student *n* being awarded on item *i* a rating of *k* by evaluator *j*;

$P_{nijk\text{-}1}$ is the probability of student *n* being awarded on item *i* a rating of *k*-1 by evaluator *j*;

B_n is the ability (proficiency) of student *n*;

D_i is the difficulty level of item (scoring criterion) *i*;

C_j is the severity of evaluator *j*;

E_h is the difficulty level of the threshold from category *k*-1 to category *k* of the scale unique to item *i*.

Figures

Figure 1. Path model estimating change in data.

Figure 2. Path model of the relationship between Pre- and Post-Course SAK and Fair scores.

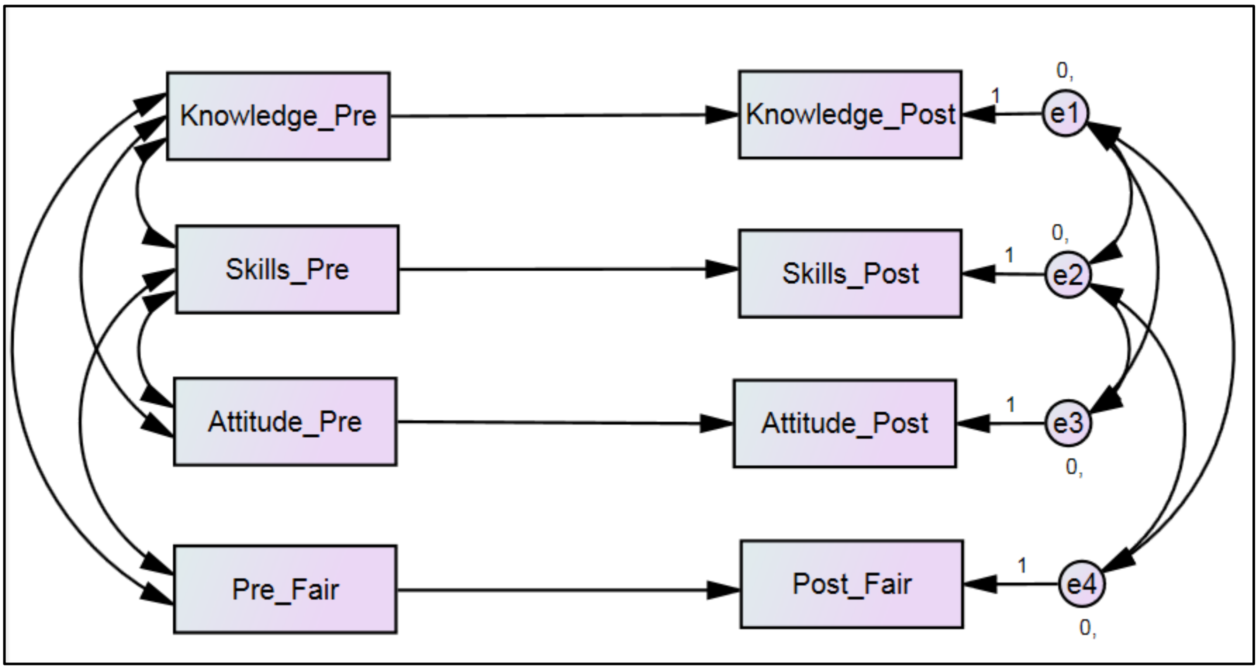

Figure 1. Path model estimating change in data.

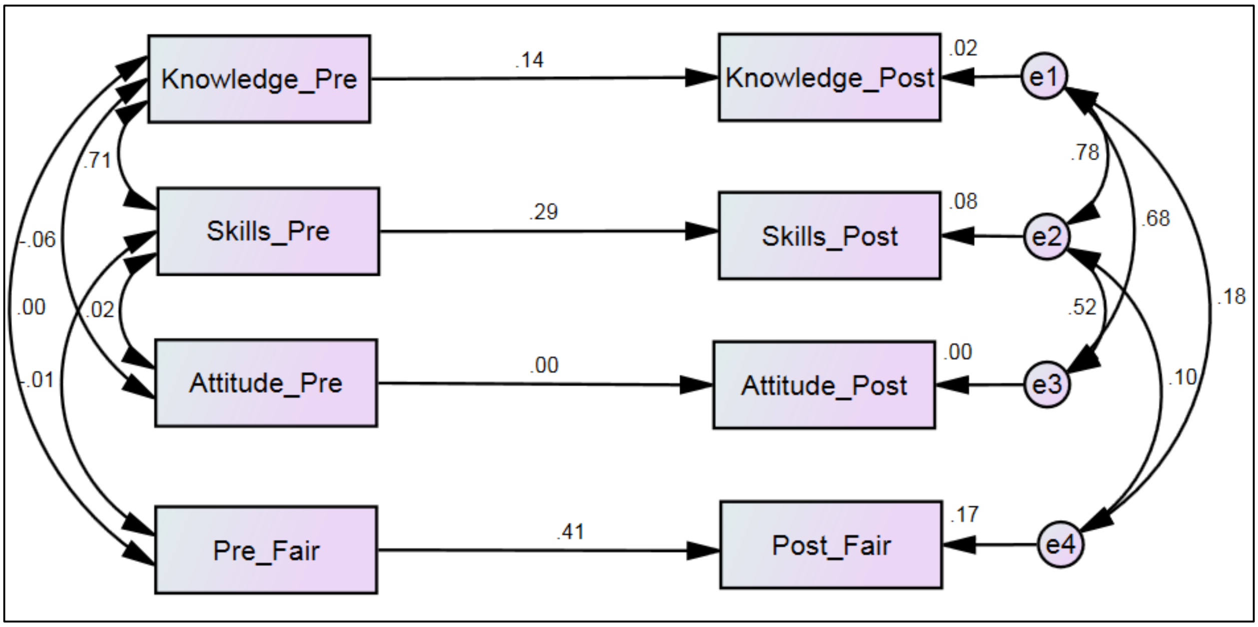

Figure 2. Path model of the relationship between Pre- and Post-Course SAK and Fair scores.

Tables

Table 1
Pre- and Post-Course Prompts

Assessment	*Prompt*
Pre- and Post-Course Prompt 1	Explore one or more reasons why teenagers are hooked on computer games.
Pre- and Post-Course Prompt 2	Compare and contrast classroom learning with and without the aid of computers.

Table 2
Psychometric Features of the Knowledge, Skills, and Attitude Dimensions

Knowledge

Items	Endorsibility measure	Infit MNSQ	Outfit MNSQ
1	-0.86	0.83	0.73
2	1.99	1.32	1.65
3	0.61	1.45	1.43
4	0.03	0.53	0.37
5	0.03	0.68	0.54
6	-1.29	1.15	1.21
7	1.92	1.32	1.54
9	-2.23	0.97	1.14
11	0.89	0.80	0.95
12	0.18	0.74	0.64
13	-0.12	0.74	0.68

Item reliability(separation) =.84 (2.27)
Person reliability(separation) = .91 (3.10)

Skills

Items	Endorsibility measure	Infit MNSQ	Outfit MNSQ
14	-1.15	1.08	1.15
15	-2.58	1.41	1.49
16	0.58	0.87	0.73
17	-0.36	1.11	0.98
18	0.32	1.15	1.09
19	0.58	1.12	1.15
20	0.06	0.98	0.89
21	-0.08	0.55	0.49

22	0.06	0.44	0.34
23	-0.22	1.47	1.32
24	-0.49	0.90	0.97
25	1.07	0.93	0.92
26	0.83	1.12	1.16
27	0.71	0.94	0.91
28	-0.50	0.56	0.42
Item reliability(separation) =.84 (2.27)			
Person reliability(separation) = .91 (3.10)			
Attitude			
Items	Endorsibility measure	Infit MNSQ	Outfit MNSQ
29	-0.55	1.50	1.54
30	0.81	0.68	0.66
31	0.21	1.15	1.08
32	-0.55	0.49	0.42
33	0.62	1.06	0.97
Item reliability(separation) =.84 (2.27)			
Person reliability(separation) = .91 (3.10)			

Table 3

Four Items Measuring Students' Reactions to the Course

Item	Category endorsement					RSM results	
	1	2	3	4	Missing	Infit MNSQ	Outfit MNSQ
1. I enjoyed attending this English class.	1(1.2%)	13(15.3%)	58 (68.2%)	6 (7.1%)	7 (8.2%)	1.01	1.33
2. I believe I learnt useful things in this course.	0	1(1.2%)	54 (63.5%)	22(25.9%)	8 (9.4%)	0.73	0.71
3. I made the best use of my time in this course.	0	3(3.5%)	57(67.1%)	18(21.2%)	7 (8.2%)	0.76	0.67
4. I disliked attending English classes.	14(16.5%)	49(57.6%)	13(15.3%)	2(2.4%)	7 (8.2%)	1.24	1.28

Table 4

Application of RMS to Pre-Course and Post-Course SAK

	Infit MNSQ	Outfit MNSQ	Item reliability	Item separation	Person reliability	Person separation	Average endorsement1 /measure	Average endorsement2/ measure	Average endorsement3 /measure	Average endorsement4 /measure
Knowledge_Pre	1.00	0.95	.97	5.84	.73	1.77	17(2%)/-4.19	97(14%)/-1.87	429(61%)/1.52	158(23%)/4.85
Knowledge_Post	0.96	1.00	.91	3.10	.84	2.27	1(0.00%)/-5.24	41(7%)/-2.56	407(71%)/2.06	126(22%)/6.22
Skills_Pre	0.99	0.99	.95	4.44	.86	2.50	69(5%)/-4.93	603(46%)/-1.90	591(45%)/1.91	59(4%)/4.90
Skills_Post	9.92	0.96	.82	2.16	.82	2.13	1(0.00%)/-6.16	100(15%)/-2.74	505(75%)/2.53	65(10%)/6.57
Attitude_Pre	0.99	0.99	.96	4.97	.79	2.00	44(4%)/-4.92	433(43%)/-1.92	482(48%)/1.91	54(5%)/ 4.95
Attitude_Post	1.01	0.99	.79	2.00	.77	2.00	4(1%)/-3.43	21(5%)/-1.75	220(57%)/1.07	139(36%)/4.59

Table 5

Rasch Model Reliability Statistics of Students' Scores across Time

	Pre-Course	Post-Course
Rasch model reliability	.95	.91
Rasch model separation	4.19	3.10
Overall infit MNSQ	1.02	1.09
Overall outfit MNSQ	0.95	1.04

Table 6

Raters' Severity Measure and Fit across Time

Observed average	Fair score	Severity measure	*SE*	Infit MNSQ	Outfit MNSQ	Raters
Pre-Course						
21.36	21.17	-0.62	0.04	1.25	1.10	1
16.66	16.57	0.53	0.04	0.98	1.05	2
18.57	18.37	0.09	0.04	0.87	0.80	3
Post-Course						
18.80	19.02	-0.01	0.05	1.10	0.88	1
18.80	19.02	-0.01	0.05	1.17	1.21	2
18.90	18.93	0.03	0.05	0.99	0.92	4
	Reliability = 0.00		Separation = 0.00			

Table 7

T-tests between Pre- and Post-Course SAK and Fair Scores

Test	Mean	*SD*	*t* value	*df*	*p* value
Pre_Fair - Post_Fair	-0.743	2.619	-2.287	64	0.026
Knowledge_Pre - Knowledge_Post	-1.557	2.198	-6.529	84	0.000
Attitude_Pre - Attitude_Post	0.196	2.385	0.761	84	0.449
Skills_Pre - Skills_Post	-1.259	1.843	-6.299	84	0.000