Contract Learning As an Approach to Individualizing EFL Education in the Context of Assessment for Learning

Hamed Zandi, Shiva Kaivanpanah & Sayyed Mohammad Alavi

To cite this article: Hamed Zandi, Shiva Kaivanpanah & Sayyed Mohammad Alavi (2015) Contract Learning As an Approach to Individualizing EFL Education in the Context of Assessment for Learning, Language Assessment Quarterly, 12:4, 409-429, DOI: 10.1080/15434303.2015.1104315

To link to this article: http://dx.doi.org/10.1080/15434303.2015.1104315

Published online: 01 Dec 2015.

Submit your article to this journal

Article views: 68

View related articles

View Crossmark data
Contract Learning As an Approach to Individualizing EFL Education in the Context of Assessment for Learning

Hamed Zandi

University of Tehran, Tehran, Iran and Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

Shiva Kaivanpanah and Sayyed Mohammad Alavi

University of Tehran, Tehran, Iran

Contract learning as an approach to individualizing education in the context of assessment for learning is relatively underexplored in English as a Foreign Language instruction. The present study used a mixed-methods design to investigate its efficacy to provide feedback to students and improve self-directed learning. Furthermore, it studied students’ impression of the assessment procedure and their agency. After taking a diagnostic test of grammatical knowledge, participants (N = 14 graduate students) received treatment, using learning contracts, and then took a posttest. A paired t test showed a significant improvement in their grammatical knowledge (t = 7.96, df = 13, p = .00), the effect size being r = .91. The analysis of data from an open-ended questionnaire helped explain how the approach had a positive effect on learning of grammar. The results also indicated that the students had favorable attitudes toward contract learning, mentioning that it had a beneficial impact on their studying behavior. The article concludes with a discussion on the validity considerations of using contract learning in an assessment for learning context.

INTRODUCTION

Assessment is generally thought to be intrinsic to successful teaching; however, not all types of assessment can be equally effective. Assessment can help learning if it is able to provide “information that teachers and their students can use as feedback in assessing themselves and one another and in modifying the teaching and learning activities in which they are engaged” (Black, Harrison, Lee, Marshall, & Wiliam, 2004, p. 10). This understanding of the function of assessment is currently at the core of assessment reform in different corners of the world (Inbar-Lourie, 2008). However, there are instances in which the students’ ability levels are heterogeneous so that a whole class feedback session may seem inadequate, or attending to the weaknesses of certain students could distract others from the main focus of class instruction.

Correspondence should be sent to Shiva Kaivanpanah, English Department, Faculty of Foreign Languages and Literatures, University of Tehran, Karger Shomali Street, Between Street No. 15 and 16, Tehran, Iran, P.O. Box 14155-6553. E-mail: shkaivan@ut.ac.ir

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hlaq.
In such circumstances the use of learning contracts—documents prepared by teachers and learners that outline the individualized learning goals—is proposed as a way of individualizing feedback (Davidson, 1986; Moon, 2004).

Because there is a paucity of research on the efficacy of learning contracts in EFL instruction, the current study investigates an approach to the provision of feedback and the improvement of learning in the context of assessment for learning (AFL). It uses a mixed-methods (MM) design to study how successfully assessment can lead to learning by using a diagnostic test to create a weakness and strength profile for the students, providing them with detailed individualized feedback, setting learning goals, and providing the necessary scaffolding to reach these goals. The feedback, the goals, a timetable to reach these goals, and the division of responsibility between the teacher and the student are documented in a contract format. The study underlines the students’ agency in the assessment process and the importance of developing a validity argument for the use of learning contracts. It also offers explanations as to how contract learning can lead to the improvement of grammatical knowledge.

**BACKGROUND TO THE RESEARCH**

The Assessment Reform Group (2002) in the United Kingdom defines AFL as “the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there” (pp. 2–3). This definition highlights feedback; however, it does not explicitly take into account another underlying principle of AFL (i.e., individualization of assessment). In fact, AFL acknowledges that failure in learning can be caused by teachers not recognizing differences in students’ knowledge and ability and consequently subjecting them to the same course of instruction (Wiliam, 2011). Students can have diverse agendas for their lives, they can have different types of motivation, and they can learn at varying rates. Thus, AFL should attempt to accommodate assessment and instruction for every individual. Therefore, diagnosis, feedback, and its individualization are at the core of AFL.

In diagnostic tests the provision of extensive individualized feedback rather than a general or global score is of prime importance (Shohamy, 1992; Spolsky, 1990). A diagnostic test can also be regarded as an example of a “use-oriented test” that “addresses issues related to the rationale for giving tests and the effects that tests have on test-takers, education, and society” (Shohamy, 2001, p. 3). Designing a diagnostic test needs a projection of intended effects on specific test users. Shohamy asserted that test takers should be empowered and considered as important stakeholders in use-oriented tests. Diagnostic information provided in a diagnostic test thus encourages test takers to become active users of test information by analyzing their own performances and guiding their own learning. Accountability can increase by providing detailed assessment criteria and useful diagnostic information in score reporting. This may also help test takers develop positive attitudes toward assessment, acknowledging that taking a test could be not only an activity for external evaluators but also a meaningful activity for themselves (Shohamy, 2001). Despite the great potential of diagnostic tests for positive washback on teaching and learning, only a few studies have investigated this issue.

One of the reasons for the scarcity of diagnostic language tests may be an inadequate understanding of the processes of what has to be learned for one to become a proficient language learner (Alderson, 2005). Nevertheless, there are some voices in mainstream
measurement (e.g., Mislevy, 2009; as cited in Chapelle, 2012) who promote the view that models and constructs do not have to be psychologically real to be useful as long as they are not categorically wrong. This view makes allowances for the design and use of diagnostic tests provided that their validity could be reasonably demonstrated. Similarly, Moss (2003) and Fulcher and Davidson (2007) argue that a line of validity argument for the use of a test in classroom context can be developed provided that the test has a positive effect on learning. A desired effect cannot be achieved simply by administering a diagnostic test and by producing an individualized detailed feedback report; the feedback needs to be conveyed to the students and intervention should be planned accordingly. This article argues that contract learning can be used in the context of AFL for conveying individualized feedback and increasing student agency in use-oriented testing.

Among the prominent proponents of individualized instruction is Parkhurst (1922), who noted that instructions with only a one-way flow of information from teachers to students are less effective than instructions involving an ongoing conversation between teachers and learners. This ongoing conversation is ideally maximized so that it can take place on a one-on-one level. Furthermore, Parkhurst (1922) asserted that “If we use class teaching and individual work in their proper places, the best results will follow” (p. 3). Thus, learning can be optimized if it is perceived as a mutual responsibility for both teachers and learners. However, if teachers and learners follow their own agenda and goals without explicitly negotiating them, they cannot properly divide their respective responsibilities.

With this goal in mind, contract learning (following Berte, 1975; Knowles, 1986; Parkhurst, 1922) provides a pedagogical approach designed to facilitate re-focusing instruction and assessment from the product to the process and makes learning more self-directed. Learning contracts have been used rather extensively in higher education to promote self-study and autonomous learning (e.g., Anderson, Boud, & Sampson, 2013; Gardner & Miller, 1999). Several studies have shown that such contracts can help learners pursue relevant goals (Brewer, Williams, & Sher, 2007), take more responsibility, and become highly motivated (Stephenson & Laycock, 2002).

Learning contracts thus provide a platform for students and teachers to discuss how to meet the curriculum requirements. The gap between their current knowledge and ability and the desired capacity can be identified via a diagnostic test. Then realistic and negotiated learning goals and deadlines can be agreed upon so that students can develop an understanding of their responsibilities and those of their teachers. A student’s responsibility may include self-study, pair-work, group learning, while a teacher’s responsibility may involve monitoring student progress and providing one-on-one tutorial sessions. The agreement between the teacher and student is recorded in a contract that is signed and honored by both parties. Although learning contracts are not a panacea to meet all the requirements of a well-balanced class, they do provide a means of individualizing teaching and measurement to a degree not offered by other teaching and testing methods (Davidson, 1986).

A few studies have examined the use of contract learning in ESL contexts. Davidson (1986) used contract learning in an ESL writing class. He identified 33 categories of sentence level errors committed by 74 students in four ESL writing classes. A reexamination of his findings showed that few errors could be regarded as common among the participant students—59% of the errors were common among only less than 13% of students and only 10% of errors were common among about 77% of students. Thus, teaching grammar to the whole class was not an
efficient way of addressing the issue. This highly individualized nature of errors led Davidson to conclude that contract learning could be used to address the long-standing problem of sentence level errors in ESL writing, without changing the focus of the class from writing to grammar. He reported that of an average of 5.5 learning goals contracted for each participant, an average of three goals were achieved by the end of the study.

Davidson (1984) also suggests that contract learning be considered as self-criterion-referenced assessment, where the progress of an individual in comparison to himself is measured; thus, the measurement is ipsative. Self-criterion-referenced assessment could be used as one of the bases for decision making along with criterion- and norm-referenced measures. This type of multiple referencing helps the accuracy of the judgments about the progress of learners. However, it should be acknowledged that the different assessment information yielded can carry mutually exclusive information, thereby creating a need to present the information from different sources in a score report document that enables decision makers to use the full range of information (Davidson, 2012).

Moon (2004) studied the impact of contract learning on students in an EAP writing course. The findings from this case study resembled those of Davidson (1986); that is, there was a wide range of needs, the majority of which were mutually exclusive. Moon resorted to contract learning to facilitate addressing students’ specific needs based on degree, nature, and urgency of the needs. She also suggested that a thorough introduction to contract learning is necessary to initiate positive attitudes and learner motivation. Furthermore, motivation needs to be sustained by teachers providing ongoing support, especially by teachers interacting with students throughout the contract period so that a trusting relationship can be established. Teachers should thus assume a more directive role especially at the initial stages of the contract learning because they have a better understanding of the procedures and the objectives of the course.

THE CURRENT STUDY

Contract learning accompanied by self-criterion-referenced assessment can be considered a variant of AFL and “use-oriented” (Shohamy, 2001, p. 3) testing, because it can tailor instruction and assessment to individuals through cycles of assessment and feedback, promote self-directed learning by teachers and students discussing responsibilities, and it can integrate the end result of learning with its process. However, because there is a need for more research on these issues, the current study investigates the efficacy of using contract learning and the process via which it affects learning.

The validity framework informing the present study is based on Stobart (2012), who following Kane (2006) demonstrates that current validity theory has three key elements related to how effectively (a) the interpretation and use of the results serve the purpose of assessment, (b) the relevant construct/domain is sampled, and (c) fairness and reliability are addressed.

For the first element the central idea behind validity argument is “to state the proposed interpretation and use explicitly and in some detail, and then to evaluate the plausibility of these proposals” (Kane, 2013, p. 1). To do this, a clear definition of a diagnostic test is in order. A diagnostic test in this study refers to a low-stakes test for identifying strengths and weaknesses in language knowledge and/or the use of that information to enable a detailed analysis of responses, provide (preferably immediate) feedback that this information can be acted upon. This process is
ultimately designed to lead to remediation with further instruction. It is thus proposed that a contract learning approach can promote the individualization of feedback and remedial intervention. This approach uses a self-criterion-referenced framework to systematically measure the within-individual variation in knowledge or ability over the course of study. In this framework whatever prevents assessment results from leading to learning could be a threat to validity.

For the second key element it is important to note how the construct is defined (i.e., teachers and students should have a clear idea of what is to be learned). In the current study extensive test specifications explicitly list the grammatical structures that students need to learn. Furthermore, grammatical knowledge in the specifications is defined as “a set of informational structures” related to grammatical form and meaning that can be accessed in long-term memory (Purpura, 2004, p. 86).

For the third element, reliability has to be conceptualized as the quality (not only consistency but also richness) of information collected about each individual and how it is used in feedback (discussed in the method section). In a nutshell, according to Stobart (2012), the ultimate criterion for the validity of formative assessment is whether it yields its purported consequences (i.e., learning).

An MM design was used to provide a well-supported argument for the efficacy of contract learning. MM research can have both confirmatory and explanatory functions (Teddlie & Tashakkori, 2009). An inquiry on whether contract learning succeeds in improving the students’ grammatical knowledge can have a confirmatory function, whereas a question about the mechanism by which contract learning leads to positive or negative pedagogical consequences has an explanatory function. Moreover, the MM design in this study had a triangulation function (Greene & Caracelli, 1997). In other words, the findings from the quantitative method were used to corroborate the findings from the qualitative strand for research questions 1 and 2. The following research questions were addressed:

1. Is there a significant difference in the students’ grammatical knowledge before and after contract learning?
2. Do the qualitatively oriented questionnaire data substantiate the quantifiable improvement rate?
3. How do the students perceive the diagnostic and feedback phases of AFL?
4. How does contract learning lead to learning?

CONTEXT FOR THE RESEARCH

The current study was conducted at a university in Iran. The English for Academic Purposes course in which the study was carried out, aimed to prepare the students to communicate in spoken and written English. In their teaching team meetings the majority of teachers repeatedly voiced their concerns about the students’ knowledge of certain grammatical categories as implicated in performing the speaking and writing skills. However, experience showed that addressing the grammar needs of the students created two challenges. First, because of heterogeneity of the students, devoting a portion of class time to grammar instruction was not adequate to cover their diverse needs. Second, devoting more class
time to grammar instruction not only distracted students from the main objectives of the curriculum but also left some dissatisfied.

**METHODOLOGY**

The data for both strands of MM were collected at the same time (i.e., parallel data collection). Figure 1 illustrates the procedures adopted to develop and use the contract learning.

**Participants**

An announcement about the study was emailed to more than 200 graduate students majoring in different fields—19 male and female non-English majors volunteered (about 5% of the population). They were interested in improving their English because, although English is not the medium of instruction, many professors required students to use textbooks, write papers, and give talks in English. The students’ self-reported level of grammatical proficiency ranged from beginner to intermediate. At the time of the study, except for one participant, no one was taking an English course, and their exposure to the English materials outside the study was limited. A few participants (N = 5) dropped out at different stages of the research. Thus, the data from 14 of these participants who took the pre- and posttests were analyzed in the t test, and questionnaire data from 13 participants were used in the qualitative phase of the study.

**Instrumentation**

Three instruments used in the present study included a diagnostic test of grammatical knowledge, a contract form, and a questionnaire.
A Diagnostic Test of Grammatical Knowledge

The test included a number of grammatical categories from a list that the instructors at the university had prepared over several meetings. The list comprised the units in *Oxford Grammar Practice Intermediate* (Eastwood, 2009) that were deemed essential. The specifications for the test were prepared by following the model suggested in Davidson and Lynch (2002). They included detailed information about how three items types (multiple-choice, editing, and translation) should be written. These three item types were included in the test on the grounds that although multiple-choice questions are commonly used to test language knowledge, in the recently proposed systematic approach to item writing (Shin, 2012), and elsewhere in the literature (e.g., Brown & Hudson, 1998), using a range of item types was thought to be more desirable because “all items have their particular strengths and weaknesses and tend to engage different skills. By using a variety of different task types, the test was far more likely to provide a balanced assessment” (Buck, 2001, p. 153). Therefore, to increase diagnostic precision, each grammatical category was tested three times via different formats (e.g., there was a translation, an editing, and a multiple-choice item to test conditional type II), and failure on one or more of the trio would indicate some weakness in knowledge of that category.

The specifications provided several sample items and indicated how the test was supposed to be prepared and administered. Two experts in language testing and four TESOL graduate students reviewed the specifications. All of the reviewers had experience teaching grammar, and a few of them were native speakers of English. The reviewers’ comments were recorded, transcribed, analyzed, and the necessary changes were made to the specifications. A test, comprising 89 items, was generated on the basis of the specifications and piloted twice with two groups of participants (*N* = 40, *N* = 33), similar to the ones in the main study. The following are examples of items on the test:

**Translation from L1 to L2**

L1: [پیش بیفته بک کتاب امانت رفته‌م.]
L2: /I borrowed a book last week. /

**Editing the sentence by changing a word or phrase.**

(*We have arrived in Zanjan at 10 p.m. last night when it was raining.)*

**Multiple-Choice**

He _______ baseball with his friends before going abroad last Sunday.

a. would play b. is playing c. has played d. played

Classical test theory was used to analyze the item statistics. Based on the results, items with very low difficulty or discrimination were revised and piloted once more. Finally, the test was administered to 158 participants. The items were reanalyzed and calibrated by using one-parameter item response theory. Calibration of the items was necessary for constructing a computer-generated diagnostic map discussed below. In addition, the reliability of the test, which was used as pretest and posttest, was calculated by using classical test theory (*r*<sub>Cronbach’s alpha</sub> = .95).
Contract Form

A contract form was prepared to provide feedback and organize the learning process. It included a summary profile of the weaknesses and strengths of the student, a description of the responsibility of the teacher and the student; goals and deadlines for the first feedback session, dates of the follow up meetings, and a study guideline. After the pilot study (described below), the contract (see Appendix I) became more detailed and the guidelines were made more explicit.

The Questionnaire

An open-ended questionnaire was administered to the participants \( N = 13 \). To prepare the questionnaire, one of the researchers conducted an open-ended interview with each pilot participant regarding the test itself, the feedback procedures, and contract learning. After each interview, the researcher noted the questions and answers in a research log, reflected upon them, refined the questions, and interviewed the next participant. After examining the interview log and the participants’ comments, recurring themes were categorized as follows: likes, dislikes, and comments on item types, two types of feedback (immediate and delayed), self-study material, one-on-one meetings, contract form, and experience of contract learning as a whole, as well as self-assessment of their progress. These themes were reflected in an open-ended questionnaire termed *I Think* for the main study to improve the format of the contract forms. Participants in the main study were then required to record and explain their opinions and ruminations, including likes, dislikes, and suggestions for above-mentioned themes.

Pilot Study of the Contract Learning Procedure

The procedure used to provide feedback based on the diagnostic test and the pedagogical intervention via contract learning was piloted with four male and female participants. Their test papers were scored and a diagnostic map (see Appendix II) for each of them was generated with ConstructMap software (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2006). The map provided the following information: (a) the ability level of the participants in terms of logits, (b) items they had answered correctly, (c) items they had answered incorrectly, and (d) items with difficulty levels below, at, and above the ability level of the test takers.

The items on the map also were coded so that each referred to a specific grammatical structure. For example, the ability to negate a sentence in present simple as measured by a translation item, editing item, and MC was each represented by a different code in one of the sections of the map.

To make the interpretation of the map more accessible, a spreadsheet was used (see Appendix III) in which the codes were categorized on the basis of grammatical forms that were begin measured (e.g., past simple, present simple, past progressive, just to name a few). During a one-on-one meeting with the researchers, every participant received the map and a coded spreadsheet that helped identify the major weaknesses (i.e., the items below or at the ability level of the participants—the items that were expected to be easy for them—which were answered incorrectly). The researcher then explained the logic of the map and walked the participants through their profile. After that, the researcher and the participants discussed the examinee’s strengths and weaknesses.
Then, the participants’ attention was drawn to their weaknesses, and a number of grammatical structures that they needed to study were included in the contracts. However, the participants were allowed to select a number of structures themselves and were asked how much time they were willing to spend on improving their weak points. Thus, on the basis of the priorities suggested by the researcher, along with the participants’ wants and projected time commitment, the participants were asked to study a number of units in the Oxford Practice Grammar (Eastwood, 2009), complete the exercises, and check their answers against the answer key. The areas of weaknesses of the participants, their priority, and the units they were supposed to study were summarized in the contract form. It stipulated that the participants agreed to study the selected units and complete the exercises, and the researcher agreed to monitor their progress and provide mini-lessons upon request. Some participants preferred to have two or three deadlines for studying the units (i.e., a number of units by a certain date and the rest by another date) to reduce the psychological pressure on the participants and help them manage their time.

Although the feedback procedure was broadly similar for all participants, the contracts were unique to each individual for the types of errors they had to eliminate, the units they had to study, and the deadlines they had to meet. The less-proficient students were encouraged to study at a slower pace and use the opportunity to participate frequently in one-on-one meetings.

At the end of the contract period the participants received the same test again; however, because of the differences in their negotiated contracts, there was a difference in the test items each individual was expected to answer correctly. Therefore, on the individualized posttest, only the items that were relevant for the specific participant were given because the agency of the students in determining the assessment criteria is important in contract learning. The time between the first and second administration was more than 2 weeks. The test papers were scored in the presence of each learner and feedback on incorrect answers was provided. Afterward, an index of learner improvement was calculated, indicating that out of all the contracted errors what percentage was learned.

Data Collection and Analysis

A similar procedure was followed for the main study. Participants could talk about their language problems in the one-on-one meetings that usually lasted for about 30 minutes. Participants visited the researcher three to five times to receive initial feedback, one-on-one counseling, and final progress feedback. During the final feedback session, which involved correcting the test paper, the reasons for incorrect answers to the items that they were supposed to have learned were elicited and the learner’s progress was discussed. The participants were asked to record why they had failed to answer certain test items in their I Think questionnaire.

The pre- and posttests were scored on the basis of the criteria specified in the test specifications. For the qualitative data, each individual’s response to each item on the questionnaire, irrespective of its length, constituted a unit of information in the analysis. In other words, the data were broken down according to the responses to each question, and the responses were compared, contrasted, and explored to find patterns and themes. After the initial reading of the data, some themes and topics emerged. A categorical strategy was used to break down and
The Effect of Contract Learning on the Development of Grammatical Knowledge

To answer the first and second research questions on the participants’ development of grammatical knowledge, the quantitative analyses of the pre- and posttest scores were conducted alongside qualitative analyses of participants’ self-assessment based on their answers to one of the questions on I Think requiring them to explain whether they were satisfied with their progress. Table 1 shows the participants’ code, their grammatical knowledge score in terms of logits, the number of grammatical categories (e.g., past progressive, conditional type I) contracted for each person, and the number of errors made by participants in the pre- and posttest. It should be noted that only the errors corresponding to the grammatical categories specified in each participant’s contract were calculated.

To empirically verify the participants’ progress, the improvement rate was calculated individually. This was in accordance with the spirit of self-criterion-referenced assessment, where each person is compared to him or herself and not to others, because they have a different profile of abilities and background experiences. For example, participant #318 had made 31 errors related to six grammatical categories. On the posttest, #318 repeated six of 31 errors; she had also failed on an item that she had previously answered correctly on the pretest. In short, on the posttest, she had made seven errors.

Then, the progress rate was calculated by subtracting the number of pretest errors from the posttest ones and dividing it by the number of pretest errors. The resulting value was expressed in terms of percentages (i.e., #318 has not repeated 77% of the pretest errors). In other words, first the ability level in logits—based on the pretest—was calculated and a diagnostic map was generated. Then, the results on the pretest were individualized so that the score for each participant was calculated on the basis of only the failed items related to the contracted categories. Similarly, the posttests were individualized for each participant so that they answered the items related to only the categories of grammar they were supposed to have studied. It was found that, of the categories of errors identified in the individual contracts, 59% of errors were common among all of the participants. This meant that had there been a whole class remedial
course for these students, 41% of the material taught would have been irrelevant to each individual’s immediate needs. Therefore, because the needs of the participants were diverse, the decision to use contract learning instead of whole class lessons seemed justified.

To test whether the improvement was statistically significant, a paired $t$ test was calculated between the results of the individualized pretests and posttests. The assumption of normal distribution was met for both pre- and posttest scores ($K$olmogorov-Smirnov $Z_{\text{pretest}} = .69$, $P_{\text{pretest}} = .72$; $Z_{\text{posttest}} = .97$, $P_{\text{posttest}} = .30$). The paired $t$ test revealed a 65% improvement rate in the scores between the pre- and posttest conditions, which was statistically significant ($t = 7.96$, $df = 13$, $p = .00$), the effect size being $r = .91$, which is a very large effect size indicating the findings are not only meaningful but also important. In fact, the large effect size in the present study suggests that the improvement rate is considerably high.

In addition, the questionnaire data were analyzed to understand participants’ perceptions of the diagnostic test, their self-assessment, and the consequences of using the learning contracts. Moss (1998) stressed the importance of paying attention to consequences of assessment and suggested that researchers use qualitative approaches to start a dialogue with the participants to understand their interpretation of the processes, results and consequences of assessment, and their subsequent reactions. Furthermore, Moss (2003), citing Thompson (1990), argued that valid interpretations should be justified, not merely imposed. If a learner disagrees with the

<table>
<thead>
<tr>
<th>Code of each participant</th>
<th>Ability level in logits based on the pretest</th>
<th>Number of contracted grammatical categories</th>
<th>Number of failed items on the individualized pretest</th>
<th>Number of failed items on the individualized posttest</th>
<th>Improvement rate = (correctly answered items on the individualized posttest out of the total number of items on the individualized posttest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>319</td>
<td>2.13</td>
<td>5</td>
<td>15</td>
<td>4</td>
<td>11/15 = 73%*</td>
</tr>
<tr>
<td>322</td>
<td>1.65</td>
<td>5</td>
<td>20</td>
<td>7</td>
<td>13/20 = 65%</td>
</tr>
<tr>
<td>326</td>
<td>1.57</td>
<td>4</td>
<td>20</td>
<td>8</td>
<td>12/20 = 60%</td>
</tr>
<tr>
<td>329</td>
<td>1.56</td>
<td>6</td>
<td>22</td>
<td>5</td>
<td>17/22 = 77%</td>
</tr>
<tr>
<td>324</td>
<td>1.03</td>
<td>6</td>
<td>23</td>
<td>13</td>
<td>10/23 = 43%</td>
</tr>
<tr>
<td>318</td>
<td>0.83</td>
<td>6</td>
<td>31</td>
<td>7</td>
<td>24/31 = 77%</td>
</tr>
<tr>
<td>327</td>
<td>0.56</td>
<td>7</td>
<td>34</td>
<td>7</td>
<td>27/34 = 79%</td>
</tr>
<tr>
<td>323</td>
<td>0.31</td>
<td>7</td>
<td>dropped out</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>321</td>
<td>$-0.18$</td>
<td>5</td>
<td>35</td>
<td>12</td>
<td>23/35 = 65%</td>
</tr>
<tr>
<td>320</td>
<td>$-0.36$</td>
<td>9</td>
<td>50</td>
<td>11</td>
<td>39/50 = 78%</td>
</tr>
<tr>
<td>328</td>
<td>$-0.36$</td>
<td>6</td>
<td>35</td>
<td>7</td>
<td>27/35 = 77%</td>
</tr>
<tr>
<td>330</td>
<td>$-1.58$</td>
<td>6</td>
<td>49</td>
<td>29</td>
<td>20/49 = 40%</td>
</tr>
<tr>
<td>332</td>
<td>$-2.21$</td>
<td>5</td>
<td>dropped out</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>333</td>
<td>0.69</td>
<td>3</td>
<td>14</td>
<td>6</td>
<td>6/14 = 42%</td>
</tr>
<tr>
<td>334</td>
<td>0.00</td>
<td>5</td>
<td>dropped out</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>335</td>
<td>$-2.43$</td>
<td>3</td>
<td>23</td>
<td>8</td>
<td>15/23 = 65%</td>
</tr>
<tr>
<td>Average</td>
<td>0.24</td>
<td>5.53</td>
<td>28.43</td>
<td>9.85</td>
<td>65%</td>
</tr>
</tbody>
</table>

Note. *(15-4)/15 = 73%. It means this participant decided to work on some categories in his contract. The number of items related to the categories that he had failed was calculated. There were 15 failed items. After the treatment he took the individualized posttest. However, he failed only 4 of those 15 items. Therefore, this time, he answered 11 of 15 items correctly, hence the 73% improvement rate.
teacher about the result of the evaluation, they should be able to discuss the judgment; the ensuing debate can then help clarify and modify each party’s expectations and criteria for progress, thereby involving the learner in the assessment and validation process (Fulcher & Davidson, 2007). To illustrate, participant #319 was dissatisfied with her progress. It turned out that she was comparatively more advanced and the test did not successfully diagnose her weaknesses; #319 wrote: “I learned only two grammatical points and I was expecting to learn more; I think it was because the test was not comprehensive.” Except for participant #319, the rest expressed satisfaction with their progress mentioning reasons such as:

#318: Yes. Many of the grammatical points that I had problems with were solved and in the second exam many of the problems of the first exam were solved.

Participants’ Impression of the Diagnostic and Feedback Phases of AFL

The findings indicated that participants improved their grammar, and 12 of 13 seemed satisfied with their achievement. Moreover, while agreement among readers may be a welcome outcome, disagreement is also a validity resource (Moss & Schutz, 2001). According to Broadfoot (2005), in AFL the link between means and ends is not always explicitly demonstrated, and assumptions on the effects of assessment on learners are not scrutinized:

When it comes to facilitating or inhibiting learning, the context in which the learning is taking place, the degree of collaboration between teacher and student and between students themselves, the degree of confidence possessed by students, the opportunity for effective communication around learning, and the degree of coercion associated with high-stakes summative testing are all critically important dimensions. (p. 132)

Although considering the emotional impact of assessment in the design, conduct, and validation of the test relates to the consequential validity of the assessment (Messick, 1989), an examination of the affective domain of assessment is largely neglected in research and practice. According to Broadfoot (2005), good assessment for learning goes beyond “the importance of focused and timely feedback together with opportunities for self-assessment . . . it also introduces an explicit concern with learners’ feelings; with the affective domain” (p. 134). Furthermore, having a positive attitude toward the test enhances learners’ confidence and engagement. Therefore, in the current study, one of the main concerns was the test takers’ negative reaction to the test and to the contract learning approach. To examine the students’ perceptions regarding the assessment process, their answers to questions about the diagnostic test and the ensuing feedback were analyzed. A synthesis of their comments showed that 12 of 13 participants seemed to have a favorable impression of the three item formats on the test. In their comments about the whole test, all but one participant indicated that she considered the diagnostic test as a “good” test, citing reasons such as motivation for studying and diagnosis of weak points and strengths. The following provides examples of some of the comments:

#326: I liked the test because it was a motivation to study the grammar that seems to be simple but actually is complex.
Because I had to answer the questions in limited time and I liked it that there were a variety of questions that help us realize our weakness and strength and that we answered the same questions in the second exam was very good because I can evaluate my performance better.

One of the main features of contract learning is the opportunity it provides for one-on-one meetings with the teacher. In fact, 10 of 11 participants, who had responded to the question about feedback, had a positive opinion about the meetings. However, one participant with mixed views stated: “an organized and precise plan, dedication of enough time to the students, careful examination of nuances, proper conduct... [But] I did not like presentation of general grammatical points and not focusing on the details; it is better to mention every point with its exceptions. I think the main problem is with the exception.” Because other participants did not mention this objection, it might be regarded as a matter of personal impression and an isolated incident, rather than the general practice in the feedback sessions.

Most of the participants (9 of 11) were satisfied with the feedback procedure during the meetings, citing reasons such as its effectiveness, detail, and individualized nature. Two participants suggested that the feedback sessions should become even more extended and interactive, thereby suggesting that different individuals have various preferences for amount, timing, and type of feedback. In fact, the results from other parts of the questionnaire also support this finding.

To explore the reason(s) students had failed the items related to the grammatical categories they were supposed to have learnt, the posttest papers were corrected in the presence of the participants, and they were asked to explain why they had made a particular error. This is congruent with the literature on provision of feedback, where teachers are not (or should not be) interested in just any student answers, they are interested in hearing and seeing the reasons and justifications involved in deciding on particular answers. Student reasoning made accessible to teachers in this way can then be used for analysis, which may lead to further support and guidance. (Leung, 2004, p. 31)

Accordingly, the participants were asked to note the main reasons for making errors on the posttest. Carelessness, lack of attention, and sometimes not having learned a grammatical structure properly were the most cited reasons:

#318: 1. Carelessness 2. I was expecting the grammar being tested in the questions to be more difficult and did not attend to easier grammatical structures such as the past form or past participle of the verbs.

Because there were two occasions for feedback on the test, one general and delayed after the pretest and one immediate and detailed after the posttest, the participants were asked to compare the effectiveness of the two. Six of eight participants who had responded to that question considered the second one more beneficial:

#324: In the first feedback session, I did not realize the nature of mistakes because I only knew which items I had failed. But in the second feedback, I realized that many of my mistakes are related to third person singular. If I knew this before the second exam, I would probably pay more attention and my mistakes could be far less. Maybe it would have been better if the details were reviewed during the feedback after the first test. I prefer the second type of feedback because I realized the problems that I did not expect them.
However, a few of the participants believed both types of feedback had a role to play and suggested both types of feedback are beneficial in different ways:

#327: Since in the second feedback the errors were fewer, the feedback was more comprehensive. The first feedback was general because it was too time consuming to go over all the errors.

#329: The first feedback played a role in helping me know where in grammar I have problems. If it was not for the first feedback and only feedback was about the details, I could not learn the grammar very well. The second feedback helped me know where I have been careless but the role of the first feedback and the book was more important.

In sum, it seems that most of the participants have expressed favorable views about the procedures adopted to produce a customized diagnostic profile, the method of conveying feedback, meetings, and the self-study material. Furthermore, the findings indicate that different participants preferred or needed different types of feedback, which can imply that students may benefit from multiple types of feedback.

How Does Contract Learning Lead to Learning?

This study does not claim to have identified all the factors that contribute to the learning of grammar and does not suggest a causal mechanism at work in contract learning. Nevertheless, the findings and arguments provided here—about the role of feedback, noticing, and motivation—signal a number of areas for further inquiry.

Sadler (1989) argued that for learning to take place, learners need to be aware of the standards they should aim for, compare their current level with the desired one, and take appropriate measures to close the gap. Feedback by teachers and learners’ self-monitoring play the main role in this process, and the goal of education should be to assist them to become progressively more independent from teacher feedback and more reliant on self-monitoring.

It is commonly agreed that in language learning, the transition from reliance on teacher feedback to self-monitoring is a long-term and complex process for most students. For students who have already developed a sound system of grammar, feedback may be provided implicitly and in a way that encourages them to self-monitor and ultimately become independent from the teacher, hence the suggestion of Sadler (1989). However, at the initial stages of learning, withdrawal of feedback can be detrimental, because the teacher is one of the few sources of input in an EFL context. The participants of this study were at stages in which they needed feedback for the grammatical structures they had answered incorrectly on the test. Furthermore, according to Schmidt’s (1990, 2001) noticing hypothesis, input intended as feedback (e.g., on form) needs to be noticed before it can be converted into intake. The progress rate of the participants (Table 1) and questionnaire data indicate that the diagnostic test and feedback seem to have played a role in raising awareness.

Moreover, it seems that mere noticing of errors may not be the only necessary condition for one to learn a grammatical structure in EFL settings. One would need to take deliberate steps toward learning the correct forms, because opportunities for L2 input are very limited outside of school. Thus, in the current study the participants self-studied a grammar book and discussed their problems in the meetings.
For the significance of the contract learning approach in the process of learning, most of the respondents (12 of 13) expressed favorable opinions referring to the following reasons:

#325: I voluntarily accepted this contract and this made me committed to continue with more motivation and do my studies.

#326: Relative to the time I spent for this project, I learned good points. Since a small set of structures were contracted for me and the expectations of the contract was not much, everything was satisfactory. The planning was new to me. Appropriate timing with recommendations that were not challenging could help me move very gradually towards the learning goals.

#327: It seems that this method is essential. [That it has] a categorization of errors and provides some guidelines to avoid confusion when studying [is good]. Everything was carried out with precision and organization and regular follow-ups have increased the efficiency.

#329: I liked it. The lessons to be studied were specified.

As these examples show, factors such as the examinees’ awareness of their weaknesses and their willingness to improve their grammar, their involvement in the process of setting realistic learning goals, setting deadlines, and specification of the expectations, in addition to the support and guidelines available in the form of one-on-one meetings seem to have contributed to their learning of grammar. The findings echo those of Turner’s (2012) review on classroom-based assessment, where she found that in most studies factors such as teacher and student interaction, feedback, uptake, and reflection support learning.

However, it is not clear which factor(s) had more weight in the process. One may argue that possibly the one-on-one tuition has been more essential than the self-regulation brought about by the learning contracts. It should be noted that most of the participants used the opportunity for one-on-one tuition only once. Besides, they used the time of their meetings to ask for more explanations about the material they had self-studied. This implies that they had already tried to learn most of the material themselves but needed more assistance for a few structures. Four participants (i.e., #318, #320, #330, & #334), however, had requested more than one meeting. As Table 1 suggests, those were the learners who generally had lower grammatical ability and had more contracted grammatical categories to study; thus, it was natural for them to request more meetings. It is also likely that an integration of a variety of factors, which shape the contract learning approach, has created an effect that has been greater than the sum of its parts.

**LIMITATIONS OF THE STUDY**

Participants accepted to enroll for the study because they were interested in improving their English grammar; there were no other incentives such as classroom score or credit. It is likely that the improvement rates might have increased if the students had enrolled in a course where their contract fulfillment counted toward their final score.

The present study was not conducted in a class setting because it was important to eliminate any sources of input, save the contract learning. This prevents us from making strong claims about the effectiveness of contract learning in a real classroom context, where interaction of many factors may facilitate or prohibit its outcome.
The sample for contract learning was drawn from adult students at a prestigious university with its own high-end competitive academic culture. Although the sample was not random, it was fairly representative of the population (e.g., in terms of language ability level, gender, field of study, and years at school). However, the results cannot be generalizable beyond this specific population to teenagers at high schools, where students may not feel the same level of responsibility to fulfill the terms of a contract or to an academic culture where students may not possess a similar amount of inclination to study and improve.

Thus, one of the major limitations of the present study is that convenience sampling is adopted, which can limit statistical generalization of findings. Nonetheless, the findings are still notable because they have implications that go beyond this study. In other words, although the statistical generalization is limited, the “analytic generalization” (Yin, 2009, p. 21) is not because (a) findings about usefulness of contract learning in Davidson’s (1986) study are replicated in a different setting; (b) an explanation is offered as to why and how diagnostic testing, contract learning, ipsative measurement, and emotional reaction of the students to assessment can play a role in AFL; (c) a validation argument for the AFL procedure is offered. Therefore, similar lessons could be applied to analogous contexts where AFL is practiced and involve heterogeneous students.

CONCLUSIONS

The motivation for this study was to explore the possibilities offered by contract learning in an AFL context with the objective of improving the students’ grammar without distracting them from the main focus of an EAP course. Claims about efficacy cannot be drawn without addressing one of the key concerns when introducing a new pedagogical or assessment approach into a classroom, namely, validity.

By taking into account Stobart’s (2012) three key elements in validity theory, the findings suggest that (a) the interpretation and use of the test results seem to have served the purpose of assessment (i.e., learning); (b) there were indications that the relevant domain was sampled (as evidenced by the consensus of the teachers about the needs of the students studying in that university, which was reflected in the test specifications), and students have presumably developed a clear understanding of their learning objectives based on negotiations with the teacher/researcher during learning contracts; and (c) an attempt was made to address fairness and reliability, as evidenced by rich information collected about each individual and how it was reflected in the feedback provided to the participants.

The evidence from the present study documents how assessment and learning are bridged, suggesting that the diagnostic test followed by contract learning had an impact on the development of grammatical knowledge as measured ipsatively, a finding that resonated well with the students’ own reflections on their experience and contributed to creating positive emotional and motivational consequences. The study might not have been able to identify all the factors that explain how this approach leads to learning; nevertheless, it suggests that systematic feedback based on diagnostic assistance can help (lower)intermediate students notice their errors. In addition, guidelines on how to improve weaknesses, and teacher’s role in supporting the students and helping them take responsibility for their learning can form a learning success cycle, which tends to boost motivation and positive attitudes toward learning.
Although contract learning yielded encouraging results, it may not be practical in all contexts (e.g., where teachers are culturally regarded as the sole authority, where the teachers cannot dedicate enough quality time to develop diagnostic tests and provide feedback, and where teachers lack relevant assessment knowledge). To have a more accurate estimate of the effectiveness of this approach, further research is needed to examine contract learning in different classroom contexts. Moreover, whether there is a threshold age influencing how students assume the responsibility needed for contract learning and examination of teachers’ perceptions of this approach is open to further inquiry.

Finally, although the present study indicates that a contract learning approach to individualizing feedback and self-criterion-referenced assessment for measuring progress of the students can be a promising interface between assessment and learning, exploring the full potentials of this approach requires more attention from both teaching and testing communities.

ACKNOWLEDGEMENT

We would like to express our gratitude to Prof. Fred Davison for sharing with us his experience about contract learning. We are also grateful to the anonymous reviewers of this article for their insightful comments. We would like to thank the editors of Language Assessment Quarterly for their valuable feedback and constant encouragement. Their suggestions were of immense help.

REFERENCES


Davidson, F. (1984, March). *Teaching and testing ESL composition through contract learning.* Paper presented at the annual meeting of Teachers of English to Speakers of other languages, 18th, Houston, TX.


**APPENDIX I: A SAMPLE OF THE CONTRACT FROM USED FOR THE STUDY**

*Learning Contract Form*

I have received a report about my performance on the Diagnostic Grammar Test. According to the report I have problems in the following areas:

- Past simple (form, order, negation, verb formation) #
- Present simple (form, order, negation, adverb of frequency) #
- Past Progressive (form, order, negation) #
- Present perfect (form, order, negation) #
- Future simple (form, order, negation) #
- Relative clauses (S/T,P; Do/T,P; Gen/T,P) #
- Zero conditional #
- Conditional Type I #
- Conditional type II #

I would agree to study [# of items from the above list] items until [date] and [# of items from the above list] items until [date]. I also agree to meet [instructor’s name] twice from [date] to [date] to discuss my progress.

Signature

*Guidelines*

You can refer to Oxford Practice Grammar book or its PDF copy for self-study. You may do the exercises and check your answers by referring to the answer key at the appendix. If you face a particular problem or have a question you can bring it to the one on one meetings.

<table>
<thead>
<tr>
<th>1. Past simple</th>
<th>Unit: 8 (p. 18-9) PDF: Unit 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Present simple</td>
<td>Unit: 5 (p.10-11) PDF: Unit 4</td>
</tr>
<tr>
<td>3. Past Progressive</td>
<td>Units: 9 &amp; 10 (p. 20-5) PDF: Unit 9,10 &amp; test 2</td>
</tr>
<tr>
<td>4. Present perfect</td>
<td>Unit: 11 &amp; 12 (p. 26-9) PDF: Units 11 &amp; 12</td>
</tr>
<tr>
<td>5. Future simple</td>
<td>Units: 22 &amp; 23 (p. 56-9) PDF: Units 22 &amp; 23</td>
</tr>
<tr>
<td>7. Zero conditional</td>
<td>Unit: 144 (p. 346-7) PDF: Unit 144</td>
</tr>
<tr>
<td>8. Conditional type I</td>
<td>Unit: 144 (p. 346-7) PDF: Unit 144</td>
</tr>
<tr>
<td>9. Conditional type II</td>
<td>Unit: 145 (p. 348-9) PDF: Unit 145</td>
</tr>
</tbody>
</table>
APPENDIX II: A SAMPLE OF THE DIAGNOSTIC MAP PRODUCED BY CONSTRUCTMAP USED IN THE STUDY FOR PARTICIPANT #318)

Note. All the items on the test were coded on the basis of their specifications. The first three digits indicate the grammatical category; Ed, M, and T indicate the item format (editing, multiple choice, and translation, respectively); the digit(s) after (Ed, M, or T) indicate the number of the item on the particular part of the test; point 1(.1) which appears at the end of all items indicates the items measure the same construct. For example, 111T1.1 refers to the first item on the translation part of the test that measures past simple. The specifications of items and how they relate to grammatical categories are indicated in Appendix III.
APPENDIX III: AN EXCERPT OF THE SPREADSHEET FILE BASED ON THE APPENDIX II USED FOR PROVIDING THE INITIAL FEEDBACK TO PARTICIPANT #318

<table>
<thead>
<tr>
<th>Item codes</th>
<th>Past simple</th>
<th>Item codes</th>
<th>Present simple</th>
<th>Item codes</th>
<th>Past progressive</th>
</tr>
</thead>
<tbody>
<tr>
<td>111T1</td>
<td>111T1</td>
<td>213T7a</td>
<td>211T4a</td>
<td>311T2</td>
<td>311T22</td>
</tr>
<tr>
<td>112T6</td>
<td>1211T6</td>
<td>214T7b</td>
<td>221E626</td>
<td>311T22</td>
<td>321E614</td>
</tr>
<tr>
<td>114T11</td>
<td>131M6</td>
<td>211T4a</td>
<td>231M3</td>
<td>314T9</td>
<td>331M32</td>
</tr>
<tr>
<td>121E66</td>
<td>212T4b</td>
<td></td>
<td></td>
<td>321E614</td>
<td></td>
</tr>
<tr>
<td>123E68</td>
<td>112T6 Negation</td>
<td>221E626</td>
<td>212T4b Adverb of frequency</td>
<td>322E617</td>
<td>314T9 Past progressive and past simple</td>
</tr>
<tr>
<td>122E64</td>
<td>122E64</td>
<td>222E627</td>
<td>222E67</td>
<td>323E619</td>
<td></td>
</tr>
<tr>
<td>124E611</td>
<td>132M8</td>
<td>222E67</td>
<td>232M14</td>
<td>322E617</td>
<td>Negation</td>
</tr>
<tr>
<td>131M6</td>
<td>225E61</td>
<td>231M3</td>
<td>213T7a Negation</td>
<td>333M10</td>
<td>Order of elements</td>
</tr>
<tr>
<td>132M8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>133M1</td>
<td>123E6B Order of elements</td>
<td>233M20</td>
<td>223E627</td>
<td>323E619</td>
<td></td>
</tr>
<tr>
<td>134M18</td>
<td>133M1</td>
<td>232M14</td>
<td>233M20</td>
<td>333M10</td>
<td></td>
</tr>
<tr>
<td>114T11</td>
<td>124E615</td>
<td>214T7b</td>
<td>Order of elements</td>
<td>225E61</td>
<td>A kind of negation with &quot;no&quot;</td>
</tr>
<tr>
<td>134M18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The highlighted items are those indicated by the map in Appendix II to be below or at the ability level of the participant but the participant has answered them incorrectly (the darker the highlight the more immediate the need). A quick look at the spreadsheet file indicates that this particular learner’s immediate need is improvement of her knowledge of past simple, the priority being with irregular verbs. A quick look at the spreadsheet file indicates that this particular learner’s immediate need is improvement of her knowledge of past simple, the priority being with irregular verbs.