Computers & Education 66 (2013) 54–63

Contents lists available at SciVerse ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Effect of different score reports of Web-based formative test on students' self-regulated learning

Xiaoling Zou [a,*], Xuning Zhang [b]

[a] Research Academia for Linguistics, Cognition & Application, Chongqing University, Chongqing 401331, China
[b] College of Mobile Telecommunications, Chongqing University of Posts and Telecom, Chongqing 401521, China

ARTICLE INFO

Article history:
Received 14 September 2012
Received in revised form
18 February 2013
Accepted 19 February 2013

Keywords:
Web-based formative test
Formative assessment and feedback
Score report
Self-regulated learning

ABSTRACT

1. Introduction

With the rapid development of information technology, Web-based or Internet-based language testing has gradually gained widespread acceptance, thus some internationally acknowledged English testing systems such as TOEFL iBT (Internet-based test), GRE (Graduate Record Examination) and CBIELTS (computer-based IELTS) are widely available all over the world. Especially, when Web-based language testing functions as a vehicle for assessing learning and teaching process, its advantages are getting increasingly prominent compared with the traditional paper-based language testing.

Based on this understanding, the Web-based College English Test of Band-4 (CET-4) has been piloted in various universities of China since December 2008 and is now being introduced on a national scale, although it is still at its pilot stage. To keep pace with the reform of CET-4, some universities, such as Chongqing University, have developed their own Web-based formative language testing system. However, while in the process of implementing the Web-based formative language test, the authors found its effect unsatisfactory. From the perspective of formative assessment, language learners would benefit from obtaining relatively accurate and instructive feedback to understand their learning status at a particular stage, promoting independent and self-regulated learning. However, due to limitations of the old testing system, testees were only able to receive their overall scores. Furthermore, few relevant research findings on score reports of Web-based formative language testing could be used for reference. Until now, few large-scale Web-based language tests have taken the measure of reporting test results back to each single candidate in the form of a detailed score report except TOEFL iBT, whose score reports not only provide candidates with overall scores, but also indicate individual performance of each tested language skill and its importance. They also give feedback and advice on how to enhance a particular language skill. In China, as one of the reform measures of CET-4, score reports have been adopted as a substitute for certificates. Candidates with a score of more than 220 are informed of their total scores, sub-scores, and their percentile ranking for both overall scores and sub-scores by the Examination Committee. However, neither feedback nor advice on language skill enhancement is provided by the CET-4 score report; the candidates have to judge and adjust learning strategies themselves. If more accurate information similar to the score report of TOEFL iBT is adopted in Web-based formative language tests, students may have clearer learning goals and better learning motivation, become more confident and make greater efforts in learning, and their self-regulated learning may be better promoted. Based on this hypothesis, the authors designed and developed a new way of score report with the help of

* Corresponding author. Tel.: +86 2365217412.
E-mail addresses: xiaolingzou@cqu.edu.cn, xiaolingzou@163.com (X. Zou).

0360-1315/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compedu.2013.02.016

the computer, and ran a trial among the students of Chongqing University, hoping the experimental results could support the hypothesis while also providing inspiration for universities beginning to implement Web-based formative language tests.

2. Literature review

2.1. Self-regulated learning and metacognition

Self-regulated learning (SRL) is a relatively new and hot concept in the field of learning strategy, emerging in the 1980s. In general, students are self-regulated when they are meta-cognitively, motivationally, and behaviorally active participants in their own learning process, without relying on teachers, parents, or other educational services (Zimmerman, 1986, 1989). In other words, SRL is the process of planning, monitoring and evaluating learning (Pekrun, Goeta, Titz, & Perry, 2002), or the ability of learners to understand and control their learning process and outcome (Schraw, Crippen & Harley, 2006). Since metacognition is also learners' planning, monitoring and regulation of their own various cognitive activities, some researchers put forward the idea that self-regulated learning includes students' meta-cognitive strategies for planning, monitoring, and modifying their cognition (Brown, Bransford, Campione, & Ferrara, 1983; Corno, 1986; Zimmerman & Martinez-Pons, 1986, 1988). In addition, many studies have also stressed that one of the important aspects of self-regulated learning is that learners use a variety of cognitive and metacognitive strategies to control and regulate their own learning (Boekaerts, 1999; Paris, 2001; Pintrich & De Groot, 1990; Zimmerman, 1989, 2000). Therefore, metacognition and self-regulated learning were combined together in this study.

2.2. Formative assessment & feedback for self-regulated learning

As early as 1999, Tuckman demonstrated the connection between regular feedback on academic performance and improvement in subsequent academic performance. Based on the conceptual model of self-regulated learning and the feedback principles put forward by Nicol and Dick (2006), formative assessment and feedback from teachers, peers, or others belong to external feedback process, which can be positively interpreted, constructed and internalized by learners through the process of internal feedback of self-regulation. Accordingly, Irons (2008) highlighted the need of formative assessment activities and formative feedback in order to promote students' self-regulated learning. Chen, Ho, and Yen (2010) found the computer-assisted test performance of moderate-level learners was significantly improved with the help of meta-cognitive feedback and marking strategy. Lee, Lim, and Grabowski (2010), adopting structural equation modeling, confirmed meta-cognitive feedback facilitated self-regulated learning, effective use of learning strategies, and resulted in the enhancement of academic performance. The role of self-assessment, formative feedback, or metacognitive feedback in promoting self-regulated learning has also been validated through the test developed from the Hot Potatoes software, e-Portfolio and other computer-assisted learning platforms (Alexiou & Paraskeva, 2010; Ibabe & Jauregizar, 2009).

In a sense, these empirical studies focus mainly on online feedback and assessment with the aid of information platform, which synchronizes feedback with the learning process. But with regard to testing, such continuous and immediate feedback is difficult to accomplish, because in most universities, examinations are usually arranged at the end of the course, there are an insufficient number of tests for continuous formative assessment and feedback. In addition, it is difficult for learners to get instant feedback on their testing process and performance, for the usual communication between peers and teachers, as well as self-assessment, is impossible due to the testing situation, large number of students and workload. However, timely and accurate assessment and feedback are possible if they are offered formative tests occurring in the middle of the course or while in the learning process with the help of network and information technology. Therefore, for the purpose of improving self-regulated learning, the Web-based formative test system was developed.

2.3. Score reporting of language test

Recently, criticism related to the timeliness of score reporting came from some researchers (Huff & Goodman, 2007; Trout & Hyde, 2006), who regarded the delay between tests and reporting of the results as a significant drawback. Besides, there has been increasing demand for reporting sub-scores of individual testees (Goodman & Hambleton, 2004; Haladyna & Kramer, 2005; Ling, 2009). Test takers want to know about their strengths and weaknesses in different areas for future improvement; teachers and deans want to understand test performance on various sub-areas to make necessary improvements to teaching and curriculum (Ling, 2009), despite the debates on whether to report sub-scores and under what conditions (Ferrara & DeMauro, 2006).

However, studies relevant to score reporting are limited to the evaluation of reporting features. Although some guidelines for score reporting were proposed (Aschbacher & Herman, 1991; Goodman & Hambleton, 2004), concrete examples of score reports implementing these guidelines are few (Goodman & Hambleton, 2004), due to divergent testing programs and needs for feedback. So as an effective means of providing test feedback, this study developed the new score report of the Web-based formative test, whose effect on students' self-regulated learning was to be verified through experiment and interview among the students of Chongqing University.

3. Methodology

3.1. Subjects

237 non-English major undergraduates from Chongqing University, ranging in age from 17 to 21, with an average age of 18.7, were chosen as subjects of the experiment, among whom 118 students with 81 males and 37 females were majoring in mechanical engineering and 119 students with 76 men and 43women in electrical engineering (See Table 1). Besides, all of the students had studied English for at least 6 years before entering the university and their English proficiency was evaluated as Band 3 in the placement test when they were enrolled in the university. Since the formative Web-based test is carried out respectively among the four grades from Band 1 to Band 4, it is better to limit the data analysis process to the same grade. Moreover, all the subjects had the experience of autonomous learning in the self-access

Table 1
Demographic description of sample students.

Major	Sample number	Gender		Average age	Years of English learning before entering university	English level in placement test
		Male	Female			
Mechanical Engineering	118	81	37	18.6	6	Band 3
Electrical Engineering	119	76	43	18.8	6	Band 3

center (SAC), since all students of our university were required to learn independently at least 1 h a week in the SAC, and had participated in at least two Web-based formative English tests before the experiment, so they were familiar with related operations on the computer. Although all of the subjects were required to complete the test carefully with good timing and to finish the questionnaire carefully without missing any item and to response according to their true situation, not all of them would take part in the formal experiment since the subjects who could not complete all the items in due time or hand in valid questionnaires would be excluded from the sample. Such process of sampling was to decrease the inaccuracy of the score analysis process, to minimize the unnecessarily negative effect of the situational familiarity on test performance and to ensure the validity and reliability of the Web-based test as well as the questionnaire survey.

3.2. Instruments

3.2.1. Pre-test/post-test questionnaire and interview survey

The questionnaire adopted in this research is part of Regulation of Cognition from the MAI (Meta-cognitive Awareness Inventory) designed by Schraw and Dennison (1994), for measuring students' central metacognitive strategy use within self-regulation. The questionnaire is comprised of five meta-cognitive regulation strategies: planning (P), information management strategy (IMS), monitoring (M), debugging strategy (DS) and evaluation (E), with 35 items in total. The subjects were required to rate these items on a five-point Likert scale ranging from strongly disagree to strongly agree (see Appendix A).

Pre-test/post-test questionnaire surveys were conducted respectively before the experiment and eight weeks after the experiment. The focus of the pre-test questionnaire was to figure out student use of self-regulation strategies before the experiment. The post-test questionnaire used eight weeks after the experiment was mainly to investigate the effects of both the traditional and new score reports on EFL students' self-regulation strategy use. The one-to-one interviews mainly concentrated on the exact effect of the score report, designed partially in light of Nicol and M. Dick's (2006) seven principles, conducted after the post-test questionnaire was finished through the question-and-answer approach for the purpose of discovering their changes in attitude and behavior, as well as finding out possible problems with the new score report so as to make complementary analysis to other aspects of self-regulated learning.

3.2.2. Web-based formative College English tests

The Web-based formative College English tests were designed and implemented by the curriculum standards of College English. According to Widdowson (1996), from the perspective of language skills, the behavior of speaking and writing is active, known as productive skills; listening and reading are passive, called the receptive skills. In Wen Qiufang's view (2008), translation is also one of the productive skills. Therefore, on designing the structure of the Web-based test, the receptive language skill sections were configured with 2 long conversations of listening comprehension (LCC), 4 passages of listening comprehension (LC) and 3 passages of reading comprehension (RC); in addition, 2 short passages of English-to-Chinese translation (E-CT) and 2 passages of compound dictation (CD) were added in the productive language skill sections. The test questions of the listening and reading sections were objective items, scored automatically by the computer system; and translation and compound dictation sections were set with subjective items, which were graded online manually by the authors.

3.2.3. Score report

The score report of TOEFL iBT was borrowed as a template for the design and development of the new form of score report of the Web-based formative test, which delivered not only overall scores but also sub-scores to each subject, and his/her percentile position as well as corresponding feedback on self-regulation strategy use. The feedback and the advice on self-regulation strategies were designed and revised according to the scores of the previous Web-based test, along with data from the pre-test questionnaire. This new form of score report (see Appendix B) was handed out to the experimental group after the next Web-based formative test.

3.2.4. Structural equation modeling

The structural equation modeling (SEM), a theory-testing statistical technique, was adopted as one of the major tools in the confirmation of the structural validity of the questionnaire data as well as the Web-based test.

The reasons for adopting the confirmatory factor analysis (CFA) in SEM mainly lie in its advantage over the exploratory factor analysis (EFA) in the following aspects: (1) CFA is able to calculate the measurement errors of each individual item extracted from the variance of these items. Actually in social sciences, such measurement errors often come from similar origin, which means these errors have a sort of covariate relationship. But EFA presupposes that they are uncorrelated to each other. So in a sense, the accuracy of the parameters in CFA is higher than that in EFA. (2) With CFA, researchers are able to pre-determine certain common factor or factors that an item belongs to, just according to related theory or empirical rules. Therefore, an item can at the same time be distributed to different common factors, with one of its factor loadings set as the fixed, or even the factor loadings of several items can be set as equal. While in EFA, the common factors are worked out in a data-driven process, and the factor loadings can only fall on one common factor. (3) In an EFA process, the relationship between common factors must either be oblique or orthogonal, that is to say, common factors are either completely correlated or totally uncorrelated to each other. In contrast, in a CFA process, in accordance with related theory or empirical rules, certain common factors are possible to be pre-determined to be either correlated or uncorrelated to one another, or even to be equal in correlation.

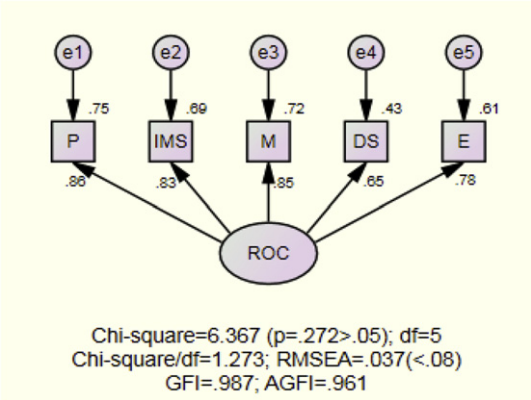

Fig. 1. Confirmatory factor analysis of the ROC questionnaire (Model 1).

Handy software named AMOS was employed when drafting the theoretically and/or empirically based models. With AMOS, both observed and latent variables can be plotted conveniently with restrictions on the parameters specified automatically by the program. Therefore, in order to make sure of the precision of the experimental data, the authors validate the questionnaire data and the Web-based test from an overall theoretically and/or empirically based perspectives of CFA rather than on a data-driven base of EFA.

3.3. Treatment

Prior to the formal experiment, the pre-test questionnaire was used for the determination of the control and the experimental group. All of the 237 subjects participated in the experiment in the SAC. The pre-test questionnaire was handed out to the subjects, who were required to complete all the questionnaire items in 15 min according to their real learning experience.

After all the subjects finished the questionnaire, they were asked to take part in the Web-based formative test. The Web-based test system automatically divided the candidates into 5 different groups in accordance with the final digit of their student number, each group was respectively configured with Test A, B, C, D and E and required to complete all the items within the given time. After the test, only the traditional overall-score report was given to the subjects.

After all the data of the pre-test questionnaire were collected, those who handed in invalid questionnaire or who failed to complete all the questionnaire items or whose responses were abnormal were excluded. According to the total scores of each meta-cognitive strategy use in the valid pre-test questionnaire, the remaining 200 subjects were first sorted in descending order and then divided into two parallel groups in even-odd order, with 100 subjects in each group, respectively specified as the control group and the experimental group. As a result, the subjects in both the control and the experimental group were at the same level in terms of self-regulation strategy use, which paved the way for observing the differences made by the newly designed score report between these two groups in self-regulation strategy use.

In processing the test scores, to ensure the reliability and the validity of the data for subsequent steps of research, score equating was carried out between sub-scores of each question type, with the internal consistency of these equated scores inspected, and structural equation modeling was employed to analyze the construct validity of the test.

On the basis of fundamental reliability and validity of the test and the questionnaire, the relationship between the scores of the questionnaire and the Web-based test was verified via multiple regression analysis, where questionnaire items of each cognitive regulation strategy that had a significant effect on the sub-scores of the Web-based test were therefore labeled. At length, with the exact relationship between the cognitive regulation strategies and the Web-based test scores in the multiple regression models as a reference, the labeled questionnaire items served as a benchmark for further revision, from which feedback and advice was developed. In the new score report, different students could receive different feedback and advice corresponding to their test results and learning experience.

Another Web-based formative test was carried out in the next semester. All valid subjects took part in this test abiding by the rules of the previous test. After this test, the traditional overall-score report was handed out to the subjects in the control group, and the subjects in the experimental group were delivered the newly developed score report. Eight weeks later, the subjects in both groups were asked to re-complete the questionnaire.

Paired sample *t*-tests of the pre-test/post-test questionnaire data were conducted respectively in the control and the experimental group to analyze the effect of different ways of score reporting on cognitive regulation strategy use. The post-experiment difference in cognitive regulation strategy use was analyzed between the control and the experimental group through an independent sample *t*-test.

Table 2
Regression weights of the variables in Model 1.

			Estimate	Standardized estimate	S.E.	C.R.	*P*
IMS	←	ROC	1.177	0.828	0.082	14.385	***
M	←	ROC	0.939	0.850	0.063	14.996	***
DS	←	ROC	0.525	0.654	0.051	10.208	***
P	←	ROC	1.000	0.864			
E	←	ROC	0.744	0.778	0.057	13.077	***

***The regression weight is significantly different from zero at the 0.001 level (two-tailed).

Table 3
Goodness-of-fit indices of the two models.

Model	χ^2	$p > 0.05$	df	χ^2/df	GFI	AGFI	NFI	TLI	CFI	RMSEA
				<2.00	>0.90	>0.90	>0.90	>0.90	>0.90	<0.08
1	6.367	0.272	5	1.273	0.987	0.961	0.989	0.995	0.998	0.037
2	7.191	0.126	4	1.798	0.986	0.947	0.978	0.974	0.990	0.063

Before conducting the interview, all subjects within each group had been separately sub-divided into high, middle and low achievers according to their overall scores of the last Web-based test. Three high achievers, four middle achievers and three low achievers were randomly picked out respectively from each group, with a total of 10 students within each one. This was to make sure that students of different levels could be sampled out and reasonably represented the entire subjects. In the interview, the subjects were asked to talk with the authors individually and answer the authors' questions one by one as follows:

1) After receiving the score report, have you made self-judgment of or reflection on your testing process and everyday learning or exchanged your learning experience or performance with your peers or classmates?
2) Were you motivated or discouraged when you understood your performance in the test? Did you make better learning plans or set better goals, or make effort for the improvement of your English according to the score report?
3) Was the feedback from the score report clear enough for you to take further actions?

4. Data analysis results

4.1. Pre-test questionnaire data

Through statistical analysis with SPSS 20.0, the reliability value (Cronbach's α) of the questionnaire is 0.940, with the total score of the five meta-cognitive strategies having the reliability value (Cronbach's α) of 0.889, both of which are higher than the minimum standard 0.7. Meanwhile, AMOS 20.0 was used to carry out the confirmatory factor analysis (CFA) of the questionnaire data to confirm the construct validity of the questionnaire. As is shown in Fig. 1, the five meta-cognitive strategies respectively serve as the measured indicators of the model, and "Regulation of Cognition" (ROC) as the latent variable.

As is displayed in Table 2, all regression weights (factor loadings) of the five measured indicators are significant at the 0.001 level. And Table 3 indicates that all goodness-of-fit indices of Model 1 (χ^2/df, p, RMSEA, GFI, AGFI, NFI, TLI and CFI), in terms of the external quality of the model, are up to the standard, which means the external quality of the model is fine. Construct reliability and the average variance extracted also meet the requirement (see Table 4), and there are no negative values in the estimates of variances (see Appendix C), which indicates that the internal quality of the model is fine.

4.2. The previous Web-based formative test

On account of the inconsistent reliability of each set of test, the commonly adopted equation for test score equating under unequal reliability was employed when gathering and analyzing the Web-based test scores, to process score equating between the sub-scores of all 5 sets of tests:

$$L_x(y) = \mu_x + R_x\sigma_x\left(y - \mu_y\right)/\left(R_y\sigma_y\right)$$

In the equation above, μ_x and μ_y stand respectively for the mean values of Test X and Test Y; σ_x and σ_y represent respectively the standard deviation values of Test X and Test Y; R_x and R_y respectively are the square root of the reliability values of Test X and Test Y. Via this equation, a candidate's score in Test Y is able to be transformed into the equivalent score of Test X, i.e. $L_x(y)$. The statistical analysis with the help of SPSS 20.0 indicates that both the reliability values before and after the score equating of Test A–E, as well as the reliability values of the sub-scores after the score equating, are higher than 0.700 (see Table 5), which met the fundamental requirement for testing reliability.

The statistical method of SEM is employed. Confirmatory analysis of the construct validity of the Web-based test is conducted with the help of the software AMOSTM 20.0. In accordance with the theory of receptive and productive language skills, a model of path analysis with latent variables was established, with the equated sub-scores as the observed variables and "receptive skills" and "productive skills" as the two latent variables (see Fig. 2).

As can be seen from Table 6, all regression weights reach the significance level of 0.001. The residual err1 is also significant at the 0.05 level ($p = 0.011$), all estimates of variances are positive (see Appendix C) and the goodness-of-fit indices of Model 2 come up to the model fit standard (see Table 3). As is shown in Fig. 2, the latent variable "receptive skills" explains 71% of the variance of the latent variable "productive skills", which conforms to relevant theoretical assumptions. The construct reliability (ρ_c) of each measured model (the model of

Table 4
Construct reliability (ρ_c) and the average variance extracted (ρ_v) of each measured model.

Measured model	$\rho_c > 0.6$	$\rho_v > 0.5$
Regulation of Cognition (ROC)	0.897	0.638
Receptive skills	0.764	0.520
Productive skills	0.694	0.533

Table 5
Reliability (R^2, Cronbach's α) of Test A–Test E before and after the score equating.

Test	No. of participants	Pre-R^2	Post-R^2	Post-R^2 of sub-scores
A	53	0.886	0.821	0.784
B	35	0.792		
C	37	0.762		
D	39	0.788		
E	36	0.776		

receptive skills and the model of productive skills), along with the average variance extracted (ρ_V), also achieve the requirements (see Table 4). Therefore, both the external and the internal quality of Model 2 are fine.

And with regard to content validity, the panelists of the test verifying group assessed that the content of the Web-based test was in line with the requirements of the curriculum standards of College English, which suggests that the Web-based test can, to certain extent, examine students' required language skills.

4.3. Relationship between cognitive regulation strategies and Web-based test scores

Based on the basic reliability and validity of the questionnaire data as well as of the Web-based test, the relationship between cognitive strategies and Web-based test scores is further conducted through multiple regression analysis. As is shown in Table 7, the questionnaire items in each construct of the ROC inventory that exert a significant effect on Web-based test sub-scores are labeled. Except for items of DS (Debugging Strategy), which exert no significant predictive power on the scores of E-CT, all the other labeled questionnaire items can partly explain the variance of the corresponding sub-scores ($p < 0.05$).

4.4. New score report development

When drafting and developing detailed feedback and advice, the labeled questionnaire items serve as a benchmark for further revision. According to Nicol and Dick's (2006) seven principles for formative feedback that effectively support self-regulated learning, the sub-scores are classified into Grade A, B and C, based on percentile positions: scores falling into the percentile of 75%–100% are marked Grade A as the higher group, 25.5%–74.5% marked Grade B as the middle group and 0.5%–25% marked Grade C as the lower group. Each set of feedback and advice is attached with the appropriate level of effort consistent with the corresponding grade (1 = you should work hard; 2 = you should work harder; 3 = you should work very hard). For instance, if a candidate's score in reading comprehension falls in the percentile of 20%, then he/she is notified with relevant feedback and advice, that the level of his/her reading ability is C, and he/she should work very hard to use every strategy to enhance reading ability. The new score report and detailed feedback and advice can be found in Appendix B.

4.5. Contrast between the traditional and the new score reports

As is shown in Table 8, the independent sample *t*-test indicates that, before the experiment, the cognitive regulation strategy use between the control and the experimental group was not significantly different ($p > 0.05$).

After the experiment, paired sample *t*-tests of pre-/post-experiment cognitive regulation strategy use were respectively conducted within the control group as well as the experimental group. Table 8 shows that, within the control group that received the overall score

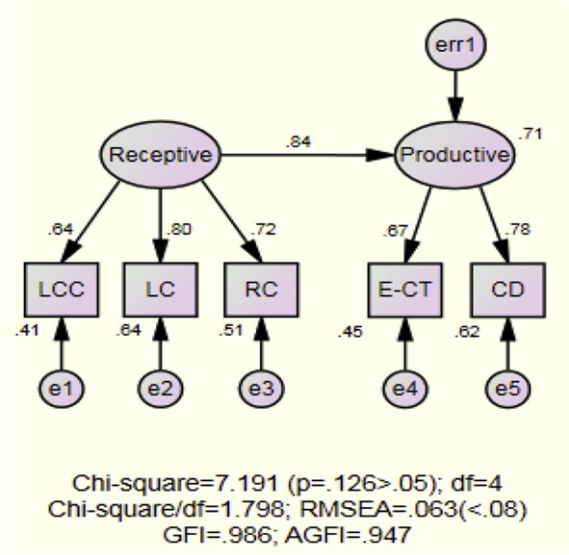

Fig. 2. Path analysis with latent variables of receptive skills and productive skills (Model 2).

Table 6
Regression weights of the variables in Model 2.

			Estimate	Standardized estimate	S.E.	C.R.	*P*
Productive	←	Receptive	0.788	0.841	0.099	7.984	***
RC	←	Receptive	1.000	0.715			
LC	←	Receptive	1.467	0.801	0.161	9.132	***
LCC	←	Receptive	0.568	0.639	0.073	7.826	***
CD	←	Productive	1.000	0.785			
E-CT	←	Productive	0.407	0.670	0.053	7.691	***

***The regression weight is significantly different from zero at the 0.001 level (two-tailed).

report, the mean values of monitoring and evaluation strategy were significantly higher than before the experiment ($p_m = 0.046 < 0.05$; $p_e = 0.015 < 0.05$); while within the experimental group given the new score report, the mean values of all cognitive regulation strategies were significantly higher than before the experiment ($p < 0.001$). Furthermore, the results of the independent sample *t*-test of the post-experiment indicate the mean values of all cognitive regulation strategies in the experimental group are significantly higher ($p < 0.001$), compared with those in the control group.

4.6. Interview data

Interview was conducted as a further exploration, and particularly, as a complement to the quantitative data above. Table 9 summarizes the answers to each interview question respectively from the students of the control and the experimental group.

Through the interview, the authors discovered that students in the control group were easily affected negatively by the single overall score in the traditional score report. Without exact data on testing performance, students' self-judgment or reflection of the testing and the learning process was hard to be effective, which led to difficulties in amending their learning goals or plans. In sharing with others their learning experience, such a score might not be clear and useful enough for making improvements, so the only thing in their mind was to "make effort to get a higher score next time". Many of the students, especially those who were in the middle and lower subdivisions, were possibly confused or even depressed when finding out their performances unsatisfactory. Generally, all interviewed students in the control group believed it necessary to reform the way of reporting test results.

Meanwhile, in the experimental group, students were delivered relatively clearer feedback information and advice which could help them judge or reflect on their testing process, learning state and strategy use, based on which more explicit learning goals or plans could be made. When communicating with others, they wanted to talk more about their learning experience, especially strategy use, and more students tended to make effort to get higher scores by improving the learning process. All of the students thought that they were more motivated and confident than ever with the new score report helping them to become more active in the learning process. However, problems with the new score report were also discovered. For instance, sometimes some students found their self-judgment or reflection was not in conformity with what was offered in the score report, especially for those whose scores were relatively better; they could not find a proper way to balance their own ways of learning and the advised learning strategies when trying to regulate their learning. Some other

Table 7
The multiple regression model of items of each cognitive regulation strategy and sub-scores of each item type.

Item type	ROC strategies	Labeled items	R^2	Adjusted R^2	$R^2\Delta$	Standardized coefficient β	Sig.
RC	P	2	0.039	0.034	0.039	0.197	0.005
	IMS	16	0.045	0.040	0.045	0.212	0.003
	M	21/22	0.088	0.078	0.025	0.178/0.174	0.021
	DS	26	0.035	0.030	0.035	0.186	0.008
	E	35	0.020	0.015	0.020	0.142	0.045
LC	P	6/2	0.080	0.071	0.020	0.205/0.147	0.039
	IMS	12/9	0.107	0.098	0.040	0.218/0.205	0.003
	M	22/19	0.106	0.097	0.022	0.221/0.164	0.029
	DS	26	0.063	0.059	0.063	0.252	0.000
	E	35	0.042	0.037	0.042	0.205	0.004
LCC	P	2	0.041	0.036	0.041	0.202	0.004
	IMS	9/17	0.073	0.064	0.021	0.196/0.148	0.037
	M	20/24	0.097	0.088	0.028	0.208/0.177	0.014
	DS	28	0.056	0.051	0.056	0.236	0.001
	E	35	0.045	0.040	0.045	0.212	0.003
CD	P	7	0.058	0.053	0.058	0.24	0.001
	IMS	16/14/9	0.130	0.117	0.020	0.166/0.167/0.150	0.036
	M	21/19	0.107	0.098	0.032	0.202/0.193	0.008
	DS	26	0.037	0.032	0.037	0.192	0.007
	E	35	0.028	0.023	0.028	0.166	0.019
E-CT	P	6/2	0.057	0.047	0.020	0.154/0.146	0.044
	IMS	14	0.056	0.051	0.056	0.237	0.001
	M	19	0.059	0.054	0.059	0.243	0.001
	DS	–	–	–	–	–	–
	E	35	0.043	0.038	0.043	0.207	0.003

Table 8
Pre-/post-experiment cognitive regulation strategy use between the control group and the experimental group.

ROC strategy	Group	Pre-experiment			Post-experiment			Paired *t*-test
		Means	S.D.	Sig.	Means	SD	Sig.	Sig.
P	Experimental	23.833	4.793	0.594	28.090	3.861	0.000	0.000
	Control	23.462	5.037		24.330	3.602		0.079
IMS	Experimental	32.960	5.937	0.710	39.300	5.476	0.000	0.000
	Control	33.278	6.142		34.500	5.169		0.061
M	Experimental	23.134	4.841	0.874	27.570	3.988	0.000	0.000
	Control	23.239	4.540		24.130	3.858		0.046
DS	Experimental	17.900	3.280	0.942	20.810	2.604	0.000	0.000
	Control	17.865	3.537		18.370	2.145		0.109
E	Experimental	19.683	3.803	0.856	23.670	3.723	0.000	0.000
	Control	19.787	4.302		20.710	3.195		0.015

students, especially those who needed to make more effort, still could not get an accurate idea of the specific techniques used both in the test and everyday learning, and they suggested that the score report should provide more explicit and concrete strategic advice.

5. Discussion

From the external and internal quality of Model 1 and Model 2, it has been confirmed that both the pre-test questionnaire and the Web-based test are reliable and valid, which lay solid foundations for the analysis of the relationship between cognitive regulation strategies and Web-based test scores. The confirmation of both models together paves the way for designing and developing feedback and advice in the new score report.

From the multiple regression analysis (Table 7), the labeled questionnaire items can explain the variance of certain question types by from 1.5% to 11.7% (minimum adjusted $R^2 = 0.015$, maximum adjusted $R^2 = 0.117$), which conforms to relevant empirical discoveries and indicates that although the strategic factors such as the self-regulation strategies cause a positive change in the performance of certain types of language test, they are not the decisive ones. Given that the core factor that determines the student performance is the linguistic knowledge, the multiple regression analysis only focuses on the overall correlation between cognitive regulation strategies and each question type of the Web-based test, instead of on specific levels of the scores. Therefore, in the new score report, students of each level of performance could see all strategic feedback information that has been revised according to the labeled questionnaire items and the specific question types, who then just need to decide the level effort of implementing these strategies.

From the paired sample *t*-test within the control and the experimental group, it is obvious that both of the two score-reporting approaches are able to promote, to some extent, students' use of self-regulation strategies. However, the independent sample *t*-test between the control and the experimental group clearly reveals that the new score report more evidently promotes students' self-regulation strategy use in all aspects. In light of the previous literature, similar positive effect can be seen that the more specific formative feedback information can to certain degree better boost student metacognition and self-regulation strategy use.

Accordingly, in terms of quantitative evidence, part of the research hypothesis has been supported that the new way of score reporting can better promote self-regulation strategy use than the traditional one.

However, such a conclusion cannot fully support the research hypothesis. Thus, the qualitative one-to-one interview served as a complement to the quantitative evidence. Through the interview, aspects other than self-regulation strategy use are explored and verified that the feedback information in the new score report more positively stimulates students' goal-setting, motivation, confidence and effort-making.

Both the quantitative and the qualitative analytical contrast together provide support to the research hypothesis, which makes the experiment complete and reasonable.

Table 9
Summary of the interview.

Topics	Control group	Experimental group
Self-judgment/reflection & peer interaction	Judging and reflecting on testing and learning process merely based on overall scores; Peer interaction limited to telling the overall scores rather than sharing and discussing the learning experience.	Judging and reflecting on the testing process, the learning state and strategy use based on the detailed feedback information; tended to discuss with peers about how to best use the strategies in the learning process.
Goal-setting/motivation/confidence/effort-making	Learning goals could not be made clearly or modified reasonably; tended to be confused or depressed; made blind efforts merely for getting higher scores next time.	Easier to make more explicit learning goals and plans; continued to learn with stronger motivation and confidence; made more effort and focused more on the learning process with clearer direction.
Clarity of feedback information	Not clear enough to be motivated to take effective measures to get better scores; believed it very necessary to take measurements to report clearer test results.	The sub-scores and corresponding percentile positions could clearly mirror the learning state; the advice helped make a relatively better understanding of how to use various strategies in testing and everyday study.

Inevitably, problems occur in the process of designing and developing the new score report. Although it was certain that the Web-based test was reliable and valid, some items could be better in reliability and validity, e.g. the question types of LCC and E-CT. The research hypothesis is based on the fact that students are familiar with the computer environment and the Web-based test, but due to the limitations of testing design and other unknown uncontrollable factors, these two question types bore relatively lower regression weights compared with other question types in Model 2. Therefore, the feedback information based on these two question types was insufficient, which could have impacted the quality of the new score report. One of the effective ways to improve the quality of feedback information is to increase the reliability and the validity of the Web-based test, especially on sub-scales.

6. Conclusions and recommendations

The exact effect of the different forms of score reports on EFL students' self-regulated learning, from both the quantitative and qualitative aspects, has provided support to the research hypothesis. From the quantitative aspect, it has gained support that the new score report better promotes self-regulated learning in all aspects of regulation strategy use; the qualitative data shows that with the help of the new score report, EFL students have clearer learning goals, better learning motivation, more confidence, as well as greater effort. The new score report can boost EFL students' self-evaluation and reflection on their testing process and everyday learning, encourage communication between peers and offer a more useful and reasonable reference to their self-regulated learning.

Still, the changes brought by different forms of score report need more complete and accurate evidence. If the duration of the experiment had been longer, students' self-regulated learning could have possibly been changed in more observable ways; more reliable and elaborate evidence could have possibly been found. Therefore, future researchers are recommended to conduct the experiment for a longer period of time and with a larger sampling scale from different levels of English aptitude, varied professional backgrounds or distinctive goal orientations and so on.

Secondly, the feedback information of the new score report was only based on the analysis of the relationship between test performance of English skills (except speaking) and self-regulation strategy use, which may have been, to some extent, too general. Future research might as well focus on individual language skills and try to explore more specific relationships and changes.

Finally this research drew its conclusion mainly from quantitative analysis, which could not make possible deeper exploration of EFL students' detailed learning processes. Accordingly, larger scales of qualitative research, especially the research aimed at individual cases could be conducted in future research.

All in all, the research related to the score reporting of English tests, especially of Web-based tests, is still not used very extensively in China. Different learning and teaching features of different nations should inevitably be taken into consideration in order to extend the research result to a wider scale. More empirical and theoretical studies are needed to make the application of important feedback more efficient to the improvement of EFL learners and the improvement of English teaching.

Notes on contributors

Xiaoling Zou is a Professor of College of Foreign Languages, Chongqing University, a Vice director of the Association of College English Teaching and Research of Chongqing, a Director of Institute of Foreign Language Education Technology in Research Academia for Linguistics Cognition & Application, Chongqing University. Her main research interests are computer-assisted language learning, English instruction and bilingual dictionary research.

Xuning Zhang is an assistant of College of Mobile Telecommunications, Chongqing University of Posts and Telecom. His research interest is computer-assisted language learning.

Acknowledgments

This article was based on project "Practical research of Web-based College English tests" supported by the Foundation of Education Committee of Chongqing and Chongqing University in China, under Grant No. 101303. Authors acknowledge the project partners.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2013.02.016.

References

Alexiou, A., & Paraskeva, F. (2010). Enhancing self-regulated learning skills through the implementation of an e-portfolio tool. *Procedia Social and Behavioral Sciences, 2*, 3048–3054.

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting (CSE Technical Report 326)*. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Boekaerts, M. (1999). Self-regulated learning: where we are today. *International Journal of Educational Research, 31*, 445–457.

Brown, A. L., Bransford, J. D., Campione, J. C., & Ferrara, R. A. (1983). Learning, remembering and understanding. In E. Markman (Series Ed.) & J. Flavell (Vol. Ed.), *Cognitive development: Vol. 3. Handbook of child psychology* (pp. 77–166). New York: Wiley.

Chen, L.-J., Ho, R.-G., & Yen, Y.-C. (2010). Marking strategies in metacognition-valuated computer-based testing. *Educational Technology & Society, 13*(1), 246–259.

Corno, L. (1986). The metacognitive control components of self-regulated learning. *Contemporary Educational Psychology, 11*, 333–346.

Ferrara, S., & DeMauro, G. (2006). Standardized assessment of individual achievement in K-12. *Educational Measurement, 4*, 579–621.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220.

Haladyna, T. M., & Kramer, G. (2005). *Poly-scoring of multiple-choice item responses in a high-stakes test*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge, UK: Cambridge University Press.

Ibabe, I., & Jauregizar, J. (2009). Online self-assessment with feedback and metacognitive knowledge. *Higher Education, 59*(2), 243–258.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback*. NY: Routledge.
Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development, 58*(6), 629–648.
Ling, G. M. (2009). *Why the Major Field (Business) Test does not report sub-scores of individual test-takers – Reliability and construct validity evidence*. Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). San Diego, CA.
Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studied in Higher Education, 31*(2), 199–218.
Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*(2), 89–101.
Pekrun, R., Goeta, T., Titz, W., & Perry, R. W. (2002). Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research. *Educational Psychologist, 37*, 91–105.
Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40.
Schraw, G.., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1), 111–139.
Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475.
Trout, D. L., & Hyde, E. (2006). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
Tuckman, B. (1999). *A tripartite model of motivation for achievement: Attitude, drive strategy*. Paper presented at American Psychological Association Conference, Boston, MA.
Wen, Q. F. (2008). On the output-driven hypothesis and reform of English-skill courses for English majors. *Foreign Language World, 2*, 2–9.
Widdowson, H. G. (1996). *Teaching language as communication*. UK: Oxford University Press.
Zimmerman, B. J. (1986). Development of self-regulated learning: which are the key subprocesses? *Contemporary Educational Psychology, 16*, 307–313.
Zimmerman, B. J. (1989). A social cognitive view self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339.
Zimmerman, B. J. (2000). Attaining self-regulation: a social cognitive perspective. In M. Borkowsk, P. Pintrich, & M. Zeidner (Eds.), *Self-regulation: Theory, research, and application* (pp. 13–19). Orlando, FL: Academic.
Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal, 23*, 614–628.
Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology, 80*, 284–290.