Investigating Relationships among Test Takers' Characteristics and Response Formats in a Reading Comprehension Test: A Structural Equation Modeling Approach

Hossein Barati [1], Hamdollah Ravand [2], Vahid Ghasemi [3]

Abstract

The present study attempted to explore the effect of cognitive and metacognitive strategy use and motivation on Iranian test takers' performance on multiple choice (MC) and constructed response (CR) tests of reading comprehension. Three hundred and sixty six students completed a cognitive and metacognitive strategy use questionnaire and a motivation questionnaire along with a MC and a CR form of a reading comprehension test. Independent-samples t-test showed that test takers' performance differed significantly on MC and CR forms of the reading comprehension test. The effect of cognitive and metacognitive strategy use and motivation on test takers' performance on MC and CR test was measured using structural equation modeling (SEM). It was found that cognitive strategy use had a significant direct effect on MC and CR but metacognitive strategy use had a significant indirect effect on MC and CR. The results also showed that motivation affected MC and CR both directly and indirectly. Another finding of the study was that performance on MC and CR tests of reading comprehension was equally affected by motivation and cognitive strategy use.

Keywords: cognitive strategy use, constructed response, metacognitive strategy use, motivation, multiple choice, SEM

1. Introduction

An issue of major concern to all testers is to make sure that test takers' performance is affected most by the ability being measured and least by factors which are not part of the ability we want to measure, otherwise the meaningfulness or validity of score interpretations

1 University of Isfahan, Iran

2 University of Isfahan, Iran, Ravand@vru.ac.ir (corresponding author)

3 University of Isfahan, Iran

will be lessened. Language testing researchers have identified and characterized factors that affect performance on language tests in any assessment situations. Bachman (1990) identified four sets of factors affecting performance in any assessment situation, namely communicative language ability, test method facets, personal attributes, and random factors.

Studies on the effect of test takers' personal characteristics on their test performance abound (e.g., Kunnan 1995; Phakiti 2006 &2008; Purpura 1997 & 1998). There are also a lot of studies on the effect of test method facets on second language test performance (e.g., Shohamy 1984; Shohamy and Inbar 1991; Kobayashi 2002; In'nami and Koizumi 2009). The interaction of test method facets with test takers' personal characteristics, however, has not been given enough attention. Bachman (1990) states that:" … At the same time, these individual attributes may interact with aspects of the methods used in any given test, so that individuals with different backgrounds and personalities may perform differentially on different types of language tests" (p. 114).

The present study thus aimed at investigating the influence of motivation (Mot), cognitive strategy use (CSU), and metacognitive strategy use (MSU) as test takers' characteristics on EFL assessment performance and explored the interaction of these characteristics with constructed response (CR) and multiple choice (MC) response formats, as sources of *construct-irrelevant* variance threatening validity of the reading comprehension tests.

2. Review of literature

2.1. Test method facets

Researchers have mainly studied the effect of the MC and CR test formats on test takers' performance from cognitive and psychometric perspectives. From a cognitive perspective, researchers have discussed whether format affects the meaning of test scores by restricting the nature of the content and processes that can be measured (N. Frederiksen, 1984). Messick (1993) argued that the full range of complex thought represented in CRs cannot be captured by MC items. Martinez (1999) argued that some cognitions associated with complex performances lie outside the range of MC items.

Despite the strong assertions by cognitive theorists, the empirical research has afforded only equivocal evidence that CR tasks necessarily measure skills fundamentally different from the ones tapped by MC questions (Bennett 1993). Katz et.al. (1996)

investigated different strategies subjects adopted to solve items in stem-equivalent CR and MC formats. They hypothesized that to the extent that the processes involved in solving the CR and MC versions of an item are the same, there should be no format effects. They found that in many situations the same reasoning procedures were used regardless of format. Rupp, Ferne and Choi (2006) took a cognitive perspective to studying the way reading assessment through MC items shapes the reading construct. Analysis of verbal reports of the ten participants in the study suggested that the sequence and structure of MC questions appear to provide important cues for test-takers that allow or influence them to select response strategies, which may result in response processes that deviate significantly from those predicted by a model of reading comprehension in non-testing contexts.

From a psychometric perspective, researchers have investigated the effect of three types of response formats on test performance in the literature: (a) stem equivalent MC and CR, (b) stem non-equivalent but content equivalent MC and CR, and (c) MC and CR which are qualitatively different; they are written to tap a different aspect of the content domain or cognitive ability. Rodriguez (2003) in a meta-analysis of the studies investigating the trait equivalence of MC and CR, found that when stems are equivalent across the two formats, the correlations tend to be significantly higher than when the stems are not equivalent. A psychometric investigation that employed cognitive markers to explain format differences, was conducted by Hancock (1994). Hancock tested the hypothesis that MC items are unable to assess higher-order thinking skills as well as CR items do. He constructed tests for two measurement classes as half MC and half CR; the tests contained equal numbers of items in each format written for the knowledge, comprehension, application, and analysis levels of Bloom's (Bloom, Englehart, Furst, Hill & Krathawl, 1956) taxonomy. A pattern of generally high disattenuated correlations between MC and CR measures within each taxonomic level was found, indicating that the two formats measure similar constructs at different levels of complexity.

Shohamy (1984) examined the effect of MC and CR testing methods on the measurement of reading comprehension. She found that the MC version had a significantly higher mean score than CR. She argued that a different skill may be invoked to fulfill the MC task. The difference can be attributed to the type of strategies involved in doing MC and CR items. Doing MC items involves comprehension and selection, while CR requires comprehension and production, production probably being a higher-level and thus more difficult task than MC (Shohamy, 1984).

Kobayashi (2002) investigated the impact of different response formats and genre types on the performance of Japanese adult ESL readers. He found that text organization, which is related to text type or genre, did not lead to strong performance differences for test formats that measured less integrative comprehension such as cloze tests or for learners of limited ESL proficiency. On the contrary, stronger performance differences due to organizational differences in texts were observed for testing formats that measure more integrative forms of comprehension, especially for learners with higher levels of ESL proficiency. This interaction effect between learner competency and testing format highlights, again, that the construct of reading comprehension that is assessed and the processes that learners engage in will change as a result of the testing format and text types used.

In'nami and Koizumi (2009) conducted a meta-analysis on the effects of MC and CR formats on L1 reading, L2 reading, and L2 listening test performance. The results using the mixed effects model of meta-analysis indicated that MC formats were easier than CR formats in L1 reading and L2 listening, with the degree of format effect ranging from small to large in L1 reading and medium to large in L2 listening. Overall, format effects in L2 reading were not found, although MC formats were found to be easier than CR formats when any one of the following four conditions was met: the studies involved between-subjects designs, random assignment, stem-equivalent items, or learners with a high L2 proficiency level. Format effects favoring MC formats across the three domains were consistently observed when studies employed between-subjects designs, random assignment, or stem-equivalent items.

Rauch and Hartig (2010) studied the dimensionality of a reading comprehension assessment with stem-non-equivalent MC and CR items. They found taht reading proficiency measured simultaneously with MC and CR items could be described more adequately with a two-dimensional IRT model than with a unidimensional model. In the two-dimensional model, proficiency aspects which all items had in common built a general latent dimension and proficiency aspects specific to CR items built an additional, nested latent dimension. The authors further demonstrated that a substantial amount of variance in reading test performance was due to differences in ability that were only measured by CR items.

Before wrapping up this part, it is worthy of note that studies comparing trait equivalence of MC and CR are mainly of two types (In'nami and Koizumi 2009): (a) studies comparing the relative difficulty of the two formats and (b) correlational studies. The studies of the first type, which have investigated the differences between MC and CR from a psychometric perspective (eg,. Shohamy, 1984; Kobayashi, 2002), have taken the

comparability of the test forms for granted. According to Traub (1993) and Messick (1993) a comparison of the relative difficulty of the items in their different formats would only be meaningful if the two formats could be shown to be measuring the same or substantially the same construct. As for the second type of studies, the results obtained by Currie and Chiramanee (2010) called into question the view that correlation coefficients at or close to unity indicate that measurements are congeneric (measure the same thing).

Currie and Chiramanee found high correlations between the MC and CR test consisting of stem-equivalent items. However, direct comparison of the responses to the items in the two tests showed that 73.9% of the answers given in the MC tests were not the answers that the participants offered to the stem-equivalent CR items. Certainly the suggestion that the two item formats were measuring the same construct was not supported by the data from the direct comparison of the participant's responses. Two conclusions might be drawn from the direct comparison of the responses to the items in the two tests. It could be argued that the two item formats are both valid methods of measuring different though largely separate language -related constructs. But in the view of Currie and Chiramanee (2010), a more realistic implication to draw from these results would be that the MC format had the effect of distorting the measurement of the language based abilities which were used by the participants in answering the CR items. This conclusion was supported by the findings from the data derived from questionnaires which were individually prepared asking each of the participants to comment on one item where their response was found to have been different in the two tests they sat. The results indicated that whereas in the CR test 91% of the responses investigated were based on knowledge or lack of it, 64% of the equivalent responses in the MC items were based on cued recall, a test taking strategy or guessing, rather than being based primarily on knowledge.

Therefore, any claim about construct-equivalence of MC and CR based on both mean-difference and correlational studies should be interpreted cautiously before it could be shown that the two test forms are measuring the same or substantially the same construct

2.2. Cognitive and metacognitive strategy use and motivation

Purpura (1999) examined the relationship between perceived cognitive and metacognitive strategy use and language test performance through the application of the structural equation modeling (SEM) approach. Purpura found that metacognitive processing had significant, direct and positive effects on all the three components of cognitive processing which directly

affected language test performance. This finding suggests that the effect of cognitive strategies on test performance is mediated by metacognitive strategies.

Phakiti (2006 & 2008) examined the extent to which test takers perceived they used a set of cognitive and metacognitive strategies in a specific test context. Phakiti found that metacognitive strategies were statistically positively related to cognitive. With regard to the relationships between strategies and test performance, cognitive and metacognitive strategy use was positively correlated with the reading test performance, explaining about 15–22% of the test score variance.

Haydel and Roeser (2002) investigated the link between different motivational patterns and performance on different types of science assessments. One-way analyses of covariance revealed that motivational pattern was significantly related to achievement on MC items from the science test when math and verbal achievement were controlled for. As the authors expected, ego-success and intrinsic-mastery students performed better than helpless students on the MC test; however, this difference was only significant between ego-success and helpless students. The same pattern of results did not occur for the CR items. Ego-success students outperformed the intrinsic-mastery and helpless students, and this difference in scores was only significant between the ego-success and intrinsic-mastery students.

Kupermintz and Roeser (2002) conducted a correlational analysis between CR and MC modes of assessment, and course grades and some motivational variables. They found that different assessment methods for science achievement, which represented different situational demands, had differential patterns of associations with motivational processes.

The combination of the variables studied in the present study and the rigorous statistical method (structural equation modeling) used to study the relationships is unique to this study. Studies mentioned above have either investigated the effect of test method or personal attributes on test takers' performance. None of the above studies has investigated construct equivalence of MC and CR from a nomothetic span perspective, in the words of Emberetson (1983). The present study investigated trait equivalence of MC and CR from the pattern of their relationships with other variables such as motivation and cognitive and metacognitive strategy use.

3. Relationships among variables of the study

Gardner et.al. (1997) in a study aimed at investigating relationships among individual differences and second language achievement found strong relationships between motivation and language achievement. They found that motivation affects L2 achievement both directly

and indirectly through cognitive and metacognitive strategy use. Pintrich (1999) points out that the more motivated the students are the more they invest in using metacognitive strategies. Schmidt and Watanaba (2001) studied the relationship between motivation, reported use of language learning strategies, and learner preferences for various kinds of pedagogical activities. They found that of the different types of learning strategies, the use of cognitive and metacognitive strategies was most affected by motivation. Purpura (1999) reported metacognitive strategies directly influence cognitive strategies and cognitive strategies in turn influence reading comprehension performance. Similar findings have been repeated in a series of studies by Phakiti (2003b, 2006a and 2008).

The present study aimed at exploring the relationships among cognitive and metacognitive strategy use and motivation on the one hand and stem-equivalent multiple choice and constructed response forms of a reading comprehension test. According to the available literature the following relationships were posited among the variables of the study. In Figure 1, the variables Mot and MSU are the exogenous variables of the model. Exogenous variables are the variables that are not affected by any other variable(s). The variables CSU, MC and CR are the endogenous variables. Endogenous variables are the variables that are influenced by other variables though they may influence other variable (s) as well. Figure 1 represents the initial model schematically.
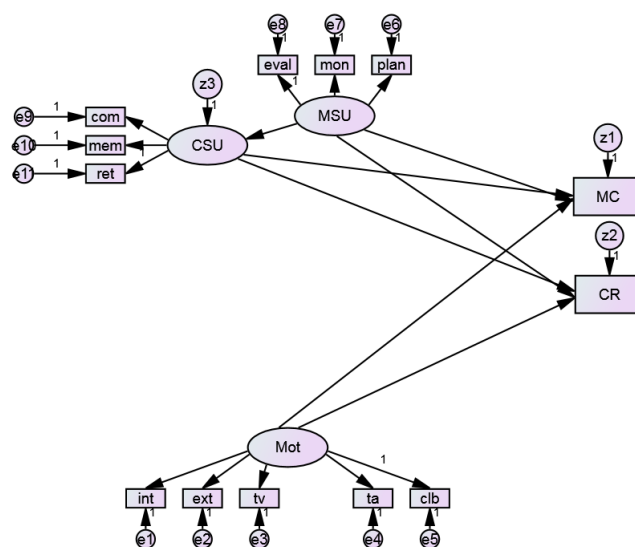

Figure 1. The initially hypothesized model

Figure 1 depicts different elements of measurement and structural models used in the present study. In structural equation models, by convention, observed and latent variables are shown by rectangles/squares and circles/ellipses, respectively. In the hypothesized model, Mot, MSU, and CSU are the latent variables and MC, CR and the other variables shown by rectangles are observed variables. In the measurement models (Mot, MSU, and CSU) directional (one-headed) arrows from factors to observed variables or indicators signify factor loadings. In the structural model directional (one-headed) arrows show regression effects. Measurement errors, are signified by e and prediction errors, by Z. As Figure 1 shows, one of the factor loadings in each measurement model, and the paths from measurement errors 'e' or 'z' to the indicators of each factor of the observed variables are set to one. Because errors and the latent factors are unobserved, there is no specific measurement unit. As such, a value must be set, and a value of 1.0 is most common. Therefore, the scale of measurement of the observed variable is transferred to the latent variable. The value could be set to a different number to scale the measurement error; however, most software programs are designed so that 1 is the default setting.

The following research questions were posed in the present study:

Q1. Does Iranian EFL test takers' performance on MC and CR forms of the reading comprehension test significantly differ?

Q2. Is test takers' performance on MC and CR significantly affected by their cognitive and metacognitive strategy use and motivation?

Q3. Are performances on MC and CR equally affected by test takers' cognitive and metacognitive strategy use and motivation?

3. Methodology

3.1. Participants

Four hundred and nine EFL undergraduates participated in the main phase of the present study. The participants were randomly chosen from three state universities: Shiraz University, Shahid Bahonar University, and Vali Asr University. They were from two fields of study: Translation Studies and Literature. From among the initially selected participants only 366 students completed all the test and questionnaire items and were therefore included in subsequent analyses. The participants were between 23 and 25 years of age.

3.2. Instruments

Two questionnaires and two different forms of a reading comprehension test were used in the main phase of the study. Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et. al. 1991) was used to measure test takers' motivation and Cognitive and Metacognitive Strategy Questionnaire (CMSQ) (Phakiti, 2008) was given to the participants to measure their cognitive and metacognitive strategy use. Both instruments were piloted before the start of the main study and according to the results of exploratory factor analysis some modifications were made (for detailed steps taken in the pilot study refer to Ravand (2013a).

Multiple choice (MC) and constructed response (CR) reading comprehension questions used in this study were stem equivalent and composed of ten items each. The two texts were selected from TOEFL past papers. The original MC items for the texts were kept and CR items were written by removing the choices of MC items. The developed CR items were later reviewed by expert judges for the quality of the items and were modified accordingly.

3.3. Procedure

The MC and CR forms of the reading comprehension tests, MSLQ, and CMSQ, were given to the participants in the same session. As for the order of presenting the MC and CR to the participants, studies of method effect are of two types in the literature; within-subject design studies and between-subject design studies (Rodriguez, 2003). In the within-subject designs all the participants undertake all the versions of the test based on the same text. These studies are either counterbalanced or non-counterbalanced, that is, participants take all the versions in the same order or the order is counterbalanced (In'nami and Koizumi, 2009). The results of these studies may be contaminated because the participants get familiar with the texts when reading them for the second and third time (Shohamy, 1984).

On the other hand, in the between-subject design, only one version of the test is randomly given to any subject (e.g. Kobayashi, 2002).This kind of design has also been criticized by Chen (2004) on the grounds that the passages in the different formats have different content and that the comparison of passages across text types would be problematic. As a remedy for this, Chen suggested that "it would be necessary to have all the participants perform the same tasks on each of the passages in the various formats, or equivalent passages appearing variously in the several formats…" (2004, p.233).

To have all the participants perform the same task on each of the passages in the various forms, would result in the same problem as that of the within subject design. In other words, the validity of the research would be undermined since each participant would read

the same text more than once. In the present study, to avoid such a problem, each participant received both test forms but based on different passages. For example, one participant would receive MC based on passage A, and CR based on passage B while another would take MC based on passage B and CR based on passage A. Thus the order in which participants received different forms was counterbalanced. The data collected in the present study were analyzed using SPSS 18 and Mplus 6 (Muthen & Muthen, 2010).

4. Results

4.1. Comparison of performance on MC and CR

As Table 1 shows, the mean performance of participants on MC was higher than that of CR which indicates that MC items were easier for the subjects.

Table 1. Descriptive Statistics

	Mean	N	Std. Deviation	Std.Error Mean
MC	6.312	366	1.9971	.1104
CR	4.859	366	2.1375	.1182

A paired sample t-test was conducted to evaluate the impact of the response format on participants' performance on reading comprehension tests. The result indicated a significant difference between MC (M=6.312, SD= 1.9971) and CR (M=4.859, SD= 2.1375), t(365) =9.502, p< 000.5 (two-tailed), eta squared = 0.23.

4.2. Impact of TTCs on MC and CR

To answer the second and third research questions, the initial model in Figure 1 was tested. The model consists of five variables: CSU, MSU, Mot, MC and CR. In this model, CSU affects both MC and CR only directly, MSU affects them only indirectly through CSU, and Mot affects MC and CR both directly and indirectly.

It is common practice in SEM analysis to break such models into a measurement model, which shows how latent variables are made up of some manifest subscales and a structural model, which tests the relationship among the latent variables (Schumacker and Lomax, 2004). In this study, accordingly, the measurement models were analyzed first and then the fit of the structural model was explored. In the interest of space, the details on the

analysis of the measurement models are not presented here. Interested readers are referred to Ravand (2013 a).

Having explored the fit of the measurement models, we proceeded with analyzing the structural model, which dealt with the relationships among the constructs of the study. The overall indices of fit for the hypothesized model are presented in Table 2.

Table 2. Fit Indices for the Hypothesized Model.

Index	Mediocre Fit	Good Fit	Observed value	Comment
SRMR(Standardized Root Mean Square Residual)	_____	<.05	. 10	No fit
Root Mean Square Error of Approximation (RMSEA)	.05-.08	< .05	.082	Mediocre
Tucker-Lewis Index (TLI)	.90	.95	.92	Mediocre
Comparative Fit Index (CFI)	.90	.95	.94	Mediocre

Generally, the above indices indicate a poor fit between the data and the model therefore the model needs some modifications.

In this section the fit statistics will be reported at the level of individual paths. In terms of the direct relationships among variables in the structural part of the model, data analysis showed that Mot was significantly associated with MC reading comprehension test performance (coefficient parameter=.43, t=6.72, P<.01). Its association with CR was equally strong (coefficient parameter=.41, t=6.34, P<.01). The above standardized coefficients are interpreted as follows: If Mot was increased by one standard deviation while CSU was held constant, MC and CR would increase by .43 and .43 standard deviations respectively. Cognitive strategy use is significantly associated with MC (coefficient parameter=.54, t=10.38, P<.01) and slightly less with CR test performance (coefficient parameter=.49, t=7.42, P<.01). But the paths leading from MSU to both MC and CR neither were significantly different from zero nor exhibited the correct sign, with the effect of MSU on CR (coefficient parameter= -0.213, t=0.897, P<.01) higher than that of MSU on MC (coefficient parameter= -0.134, t= -1.195, P<.01). These two coefficients were both below the 1.96 cut-

off and their negative signs are not substantively supported. Finally MSU is strongly associated with CSU (coefficient parameter=.83, t=23.78, P<.01). The path coefficients for the hypothesized model are reported in Table 3.

Table 3. Direct Path Coefficients for Variables of the Structural Model

Path	Estimated parameter	Standard Error	t- value	p<.01
Mot→ MC	.43	.065	6.62	.00
Mot→CR	.43	.068	6.34	.00
CSU→MC	.54	.052	10.38	.00
CSU →CR	.49	.066	7.42	.00
MSU→ MC	-.134	.112	-1.19	.23
MSU→CR	-.213	.118	-1.80	.072
MSU→CSU	.828	.035	23.77	.00

As Table 3 shows, all the paths, except for the ones from MSU to MC and CR, indicate a strong association between the variables of the study.

The next step in structural equation modeling is to consider possible changes to a specified model that has poor model fit indices, that is, model modification. Having determined the inadequate fit of the structural model, both overall and at parameter level, we found it both reasonable and logical to move into an exploratory mode and attempt to modify the structural model in a sound manner. Schumacker and Lomax (2004) argued that substantive interest must be the guiding force in a specification search. If a parameter has no substantive meaning to the applied researcher, then it should never be included in a model. In search of a better-fitting model, we decided to eliminate the two paths from MSU to MC and CR for the following reasons: (1) the path coefficients were not significantly different from zero (<1.96), (2) the negative signs of the coefficients were contrary to what is expected according to the literature, and (3) substantive evidence in the literature supported only an indirect relationship between MSU and performance on reading comprehension tests. The usual procedure in specification search is to add or delete paths one at a time and check the chi-square difference. Because each change may affect other parameters in the model and may, therefore, alter the relative importance of modification indices in a revised model.

Accordingly, we deleted one path at a time, which resulted in significant improvement in the fit of the model. Table 4 depicts the modification process.

Table 4. Chi Square Difference Test for the Revised Models.

Modified Models	χ^2	$\Delta\chi^2$	df	RMSEA	Sig. of chi-square difference test P < .01
Hypothesized model	186.632	----	58	.082	----
Respecified model A	148.325	38.307	59	.071	Significant
Respecified model B	119.442	28.883	60	.057	Significant

Although, deleting the paths from MSU to MC and CR improved the overall fit of the model significantly, the SRMR and RMSEA values of the model, as shown in Table 5, were still slightly above the .05 cut-off suggested for a well-fitting model.

Table 5. Overall Fit Statistics for the Respesified Model B

	SRMR	RMSEA	(%90 CI)	CFI	TLI
Respecified model B	.057	.057	(.059 .085)	.96	.96

Therefore, we consulted the modification indices offered by Mplus to see where in the model, modifications, if any, were needed. Most of the modification indices provided by Mplus suggested crossloading the indicators of some factors on the other factors, which theoretically didn't make sense. One of the MIs that pointed to a considerable decrease in the model chi square (41) was related to the addition of a path from Mot to CSU. Figure 2 depicts the relationships among the variables of the study in the final respecified model.
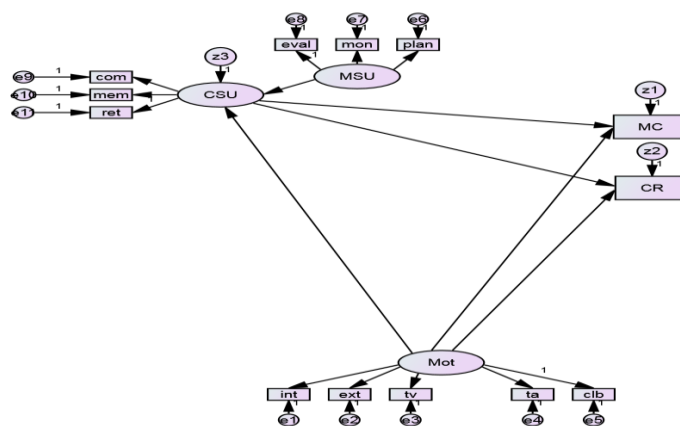
Figure 2. The final best model

Accordingly, a path from Mot to CSU was added to the model. Addition of this path resulted in a significant decrease in chi square ($\Delta\chi^2_{(1)=}$ 41.324). Table 6 depicts the modification process.

Table 6. Chi Square Difference Test for the Best-Fitting Model.

Modified Models	χ^2	$\Delta\chi^2$	df	RMSEA	Sig. of chi-square difference test P < .01
Respecified model B	119.442	---	60	.057	----
Best-fitting model	78.118	41.324	59	.033	Significant

Goodness of fit statistics for the respecified model are presented in Table 7. Once the path from Mot to CSU was added, significant improvements in fit indices emerged in the respecified model.

Table7. Comparison of Fit Statistics for the Hypothesized and Respecified Models

	SRMR	RMSEA	(%90 CI)	CFI	TLI
Respecified model B	.057	.057	(.059 .085)	.96	.96
Best-fitting model	.033	. 028	(.042 .072)	.99	.98

As Figure 2 shows, there is one new path in the model: Mot to CSU. Besides, the paths leading from MSU to MC and CR were deleted.

4.3. Equality of effects

Having established a best fitting model to the data, we proceeded to a more rigorous test of the effects of CSU and Mot on the two different forms of reading comprehension test. Although all structural paths in the final respecified model were significantly different from zero at P <.01, what was important in the present study was whether the effects of the CSU and Mot were the same or different across different forms of the reading comprehension tests. To this end in a series of steps we introduced equality constraints on the paths leading from CSU and Mot to MC and CR, the results of which are presented in Table 8. Chi square difference tests were run between the final respecified model, depicted in figure 2, (taken to be the baseline) and each of the constrained models to assess the improvement or worsening of the fit of the new model as compared to the baseline model.

Table 8. Equality Constraints Tests

Models	χ^2	$\Delta\chi^2$	df	RMSEA	Significance of chi-square difference P < 0.005
Besline model	78.118	---	59	.033	----
Strategy constrained	78.750	.63	60	.028	Not significant
Mot constrained	78.758	.71	60	. 028	Not significant

The imposition of equal effects for strategy resulted in a negligible increase of 0.63 in chi square value compared to the baseline model which was found to be statistically non-significant ($\Delta\chi^2_{(1)=}$ .63, P < .05) which is indicative of equal effect of strategy on all test forms. As for the effect of Mot, imposing equality constraint also led to a negligible increase of .71 in chi square value($\Delta\chi^2_{(1)=}$ .71, P < .05) compared to the baseline model, meaning the equality constraint holds, hence the effect of Mot on different reading comprehension test forms in the present study is the same.

5. Discussion and conclusion

1. Does Iranian EFL test takers' performance on MC and CR forms of the reading comprehension test significantly differ?

Mean performance of the participants showed that overall, MC format was easier for the participants than the CR format. T-test analysis showed a significant difference between performance on MC and CR. The conclusion is that testing methods of MC and CR can make a difference in the assessment of second language reading comprehension. It may be that the two forms are measuring different skills or different aspects of the same skill. Evidence supporting the above finding can be discussed both empirically and theoretically.

Several empirical studies in the literature have suggested the same conclusion (e.g., Shohamy, 1984; Graves et. al., 1991; Shohamy and Inbar, 1991; Kobayashi, 2002; Rupp et.al. 2006). According to Shohamy (1984) the higher mean score on the MC version may imply that a different skill is called for to fulfill the MC task. She argued that doing MC items involves comprehension and selection, while CR requires comprehension and production, production probably being a higher-level and thus more difficult task than MC. Kobayashi (2002) argued that different test formats, or even different types of items within the same format, seem to measure different aspects of reading comprehension. Van den Bergh (1990) viewed the differences in performance on MC and CR formats as related to differences in strategies test takers use to answer the items. Rupp et. al. (2006) showed that the construct of reading comprehension is assessment specific and is fundamentally determined through item design and text selection. They found that the action of assessing textual comprehension with MC questions changes the process itself and induces supplementary process that are, in their intensity, unique to the test method.

2. Is test takers' performance on MC and CR significantly affected by their cognitive and metacognitive strategy use and motivation?

There were both direct and indirect paths going from Mot to MC and CR in Figure 2. Motivation had a significant direct effect on both MC and CR. Mot was significantly associated with the MC reading comprehension test performance (coefficient parameter=.43, t=6.62, P<.01). Its association with CR was almost equally strong (coefficient parameter=.43, t=6.34, P<.01). The indirect paths from Mot to MC and CR through CSU were also significant with the indirect path from Mot to MC (coefficient parameter =.62, t=12.93, P<.01) negligibly higher than the one from Mot to CR (coefficient parameter =.60, t=12.21,

P<.01). Both theoretical and empirical evidence in the literature point to the association of affective factors in general and motivation in particular to L2 test performance.

Theoretically the above finding is in line with Bachman and Palmer's (1996/2010) model of language ability wherein language users' affective schemata, in combination with the characteristics of the particular language use task, determine, to a large extent, their affective response to the task and can either facilitate or inhibit how they use the language in a given context.

Empirical studies also point to significant effect of motivation on second language test performance in general and performance on MC and CR in particular. Haydel and Roeser (2002) found a strong link between different motivational patterns and performance on MC and CR forms of science assessments. Sundre (1999) found that the effect of motivation on test performance was stronger in the case of CR tests. She argued that the difficulty associated with the CR tasks might result in lower motivation and performances. Kupermintz and Roeser (2002) found that different assessment methods (i.e., MC and CR) for science achievement, which represented different situational demands, had differential patterns of associations with motivational processes.

The only independent variable in the present study which did not have a significant direct effect on MC and CR was MSU. In terms of magnitude, coefficients for the paths from MSU to both MC (coefficient parameter = -0.13, t= -1.19, P<.01) and CR (coefficient parameter = -0.21, t= -1.80, P<.01) were statistically nonsignificant. The results also showed that the relationship between MSU and CSU is significant, (coefficient parameter=.82, t=23.77, P<.01). Theoretically the findings are in line with Bachman and Palmer's (1996 and 2010). Bachman and Palmer (2010) argued that metacognitive strategies along with language knowledge and topical knowledge are involved in arriving at a plan for accomplishing the communicative goal or for completing a language use or assessment task. They further argued that the execution of this plan in language use, involves cognitive strategies.

There is also empirical support for this assertion. Purpura (1997) found a significant and direct influence of metacognitive strategies on cognitive strategies. Similar results have been reported by Phakiti (2003, 2006, and 2008). The CSU had a large size of direct effect on MC reading comprehension test performance (coefficient parameter=.54, t=10.38, P<.01). Its association with CR was also strong. This relationship was significant at (=.49, t=7.42, P<.01). The findings are consistent with earlier research. Purpura (1998) found a significant and direct influence of CSU on L2 test performance. Similar results have been reported by

Phakiti (2003, 2006, 2008). Phakiti (2003) found a strong association between cognitive strategy use and reading achievement.

3. Are performances on MC and CR equally affected by test takers' cognitive and metacognitive strategy use and motivation?

The imposition of equal effects for strategy resulted in 0.63 increase in chi square value compared to the baseline model which was found to be statistically non-significant ($\Delta\chi^2_{(1)} = 0.63$, $P < 0.05$) which is indicative of equal effect of strategy on all test forms. As for the effect of Mot, imposing equality constraints also led to a negligible increase of 0.71 in chi square value compared to the baseline model, meaning the equality constraint holds, hence the effect of Mot on different reading comprehension test forms in the present study is the same. To the best of the researchers' knowledge, no other study so far has investigated the effect of CSU on MC and CR forms of reading comprehension test. As for the effect of motivation on MC and CR forms of tests, Kupermintz and Roeser (2002) found that MC and CR assessment methods for science achievement, which represented different situational demands, had differential patterns of associations with motivational processes.

There is also dearth of research on the effect of other test taker characteristics on MC and CR tests. Scores on MC test have been found to be affected by format-specific strategies, sometimes called test-wiseness. Snow (1980) and Bethell-Fox, Lohman, and Snow (1984) argued that examinees may differ in the use of response elimination strategy, in which examinees narrow down the presented response options by eliminating those that are implausible; the remaining set is then evaluated more carefully to yield the best selection given what the examinee knows.

Martinez (1999) argued that CR items are probably also susceptible to their own brands of test-wiseness. Canny test takers are likely to know, for example, how to write responses that capitalize on what they know and that conceal when there are gaps in their knowledge.

Martinez (1999) suggested that CR items are also susceptible to contamination by irrelevant person characteristics. He argued that the debilitating effects of test anxiety may be greater when examinees so predisposed encounter CR items than they attempt MC items. Crocker and Schmitt (1987) found that the negative effects of test anxiety on test scores were moderate on a MC version of a test, but severe on a test that required the construction of responses.

A possible explanation for the discrepancy between the results of the present study and the studies mentioned above is that the effect of test takers' individual characteristics on different modes of assessment might be domain-dependent. Kupermintz and Roeser (2002) Sundre (1999) studied the differential effect of test takers' individual characteristics on MC and CR modes of assessment in the context of science achievement tests but in the present study the same relationship is studied in the context of L2 reading comprehension. Another reason might be that the studies by Snow (1980), Bethell-Fox, Lohman, & Snow (1984), and Crocker and Schmitt (1987) studied the effect of test takers' characteristics other than Mot and CSU on MC and CR modes of assessment. Replicating the present study with the same test taker characteristics and modes of assessment in the context of L2 reading would shed more light on the tenability of the results obtained in this study.

7. Conclusion

The present study explored the effect of cognitive and metacognitive strategy use and motivation on the EFL test takers' performance on multiple choice (MC) and constructed response (CR) tests of reading comprehension through structural equation modeling (SEM). Of further interest in this study was the interaction among the variables of the study (cognitive strategy use (CSU), metacognitive strategy use (MSU), motivation (Mot), MC and CR forms of the reading comprehension test) which was categorized as measurement error by Bachman (1990). Three hundred and sixty six students completed the instruments of the study (CSU, MSU, Mot, and MC and CR forms of a reading comprehension test).

T-test analysis showed a significant difference between performance on MC and CR. The conclusion is that testing methods of MC and CR can make a difference in the assessment of second language reading comprehension. It may be that the two forms are measuring different skills or different aspects of the same skill.

To see whether difference in performance on MC and CR can be attributed to test takers' level of motivation and strategy use, an initial model was hypothesized (see Figure 1) which included direct paths from motivation (Mot), cognitive strategy use (CSU), and metacognitive strategy use (MSU) to MC and CR forms of reading comprehension test and a path from MSU to CSU.

The final respecified model (see Figure 2) showed that there was a strong association between Mot and CSU. Motivation also had a significant effect on both MC and CR. The results also showed that metacognitive strategies directly influenced cognitive strategies and cognitive strategies in turn significantly affected performance on both MC and CR. Another

finding of the present study was that performance on both MC and CR was equally affected by Mot and CSU. The conclusion was that differential performance of test takers on MC and CR could not be attributed to their personal characteristics such as MSU, CSU, and Mot.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*: Oxford University Press, USA.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*: Oxford University Press, USA.

Bachmann, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*: Oxford University Press.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ:Lawence Erlbaum As-sociates. Inc.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, 8, 205-238.

Bloom, B. S., Engelhart, M., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay, 19, 56.

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format: A reply to Kobayashi. *Language Testing*, 21(2): 228-234.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* . Hillsdale, NJ: Erlbaum.

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. *The Journal of Experimental Education*, 55, 201-205.

Currie, M. and T. Chiramanee (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing* 27(4): 471-491.

Farhady, H., Jafarpour, A., & Birjandi, P. (1994) *Language skills testing*: *From theory to practice*. Tehran, Iran: SAMT Publishers.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist; American Psychologist, 39*(3), 193.

Graves, M. F., Prenn, M. C., Earle, J., Thompson, M., Johnson, V., & Slater, W. H. (1991). Commentary: Improving instructional text: Some lessons learned. *Reading Research Quarterly*, 110-122.

Haydel, A. M., & Roeser, R. W. (2002). *On the links between students' motivational patterns and their perceptions of beliefs about, and performance on different types of science…assessments: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 573). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *language testing, 26*(2), 219-244. doi: 10.1177/0265532208101006

Katz, I. R., Friedman, D. E., Bennett, R. E., & Berger, A. E. (1996). *Differences in strategies*

used to solve stem-equivalent constructed-response and multiple-choice SAT-mathematics items. College Board Report (95-3).

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *language testing, 19*(2), 193-220.

Kunnan, A. (1995). *Test taker characteristics and test performance: A structural modeling study*: Cambridge: Cambridge University Press.

Kupermintz, H., & Roeser, R. (2002). *Another look at cognitive abilities and motivational. processes in science achievement: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 571). Los Angeles: University of California, National..Center for Research on Evaluation, Standards, and Student Testing.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Messick, S. (1987). *Assessment in the schools: Purposes and consequences* (RR-87-51).Princeton, NJ: Educational Testing Service.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus…choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, Hillsdale, NJ: Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *language testing, 20*(1), 26-56.

Phakiti, A. (2006). Modeling cognitive and metacognitive strategies and their relationships to EFL reading test performance. *Melbourne papers in language testing, 11*(1), 53.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237-272.

Pintrich, P. R., Smith, D. A. Garcia, T. & McKeachie, W. J. (1991). *A Manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor, MI: The University of Michigan.

Purpura, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47*(2), 289-325.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high-and low-ability test takers: A structural equation modelling approach. *Language Testing, 15*(3), 333-379.

Rauch, D. P., & Hartig J. (2010).Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52(4).

Rodriguez, M. C.(2003).Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *language testing, 23*(4), 441-474.

Schumacker, R.E., Lomax, R.G., (2004). *A Beginner's Guide to Structural Equation Modeling*, 2nd ed. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1(2), 147-170.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of

text and question type. *language testing, 8*(1), 23-40.

Snow, R. E. (1980). Intelligence for the year 2001. *Intelligence*, 4.185-199

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test.consequences and test performance?*(Report No. TM029964). Harrisonburg,

Virginia:James Madison University. (ERIC Documentati on Reproduction Service No.ED432588).

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.),   *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, Hillsdale, NJ: Erlbaum.

Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14*(1), 1-12.