Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 2, October 2021
ISSN 2476-5880

Inferencing Complexity of the Iranian TEFL Ph.D. Entrance Exam under the Lens of the Construction-Integration Model of Inference Processing

Hamid Khosravany Fard[1], Mohammad Davoudi[2]*

Received: 30 March 2021 Accepted: 8 August 2021

Abstract

The present study aims at scrutinizing the inference complexity level (henceforth ICL) of test items of the Iranian State University TEFL Ph.D. entrance exam (ISUTPEE) from 2010 to 2017 through the lens of Kintsch's *construction-integration theory* (C-I) (Kintsch, 1988, 1998). Though there is ample research on inferencing in the field of reading comprehension, the existing literature reveals a serious gap in relation to inferencing complexity of test items in high-stakes exams that exert profound effects on the academic achievements of the individuals. Inferencing is examined in this study to explore the ICL of the test items of the Special Knowledge Test (SKT) according to three levels of memory representations of Kintsch's model: the surface model, the textbase, and the situation model. To this end, the test items for eight consecutive years of the ISUTPEE were examined in relation to the three distinct kinds of mental representations. To ensure the reliability of coding by the researchers, two other specialist coders assessed the ICL of 33% of the items. The intraclass correlation among the three sets of codes was 0.91. The results of the study showed that a large number of questions, accounting for more than 80% of the items, merely activate the surface and the textbase model of information representations in memory. Furthermore, the ICL for each of the four parts of SKT was examined. This analytical study carries a stark warning regarding a deficiency of systematic attention to ICL in the development of test items.

Keywords: Construction-integration model; Inferencing, Iranian state university TEFL Ph.D. Entrance Exam; Mental representations; Situation model; Surface model; Textbase model

1. Introduction

Discourse comprehension, as an inevitable enterprise wrapping up extended language segments namely novels, articles, conversations, textbooks, and other everyday materials involves us in complex cognitive processes to reach a final interpretation. According to Traxler (2011), discourse comprehension closely tied to building mental representations via intermediate products is not merely limited to the exact words in the text, word orders, and syntactic structures. He believes that

[1] Department of English Language and Literature, Hakim Sabzevari University. Email: hamidkhosravani8@gmail.com
[2] Department of English Language and Literature, Hakim Sabzevari University. Email: davoudi2100@gmail.com; m.davoudi@hsu.ac.ir

it requires much substantial amount of mental work to go beyond the text itself in order to structure the input in a way that manifests the ideas conveyed in the text.

As Graesser, Singer, and Trabasso (1994) and Kintsch (1988) argue, comprehension of a narrative text engages the individual in attributional processing to figure out event happenings, the motives for people's actions, and the likelihood of occurrence of particular events in the future in our daily lives. This process is similar to reading stories wherein we intend to unravel the why and how of events, the character's intentions in doing particular actions and reactions, sequence of events and how they fit together.

Basic reading proficiency, as a prerequisite for becoming adept at realizing more in-depth comprehension and learning 21-century skills (Goldman & Pellegrino, 2015; Graesser et al., 1994), paves the way for a successful and fulfilling education and career. Reading comprehension, as a multidimensional and complex human activity (Sosinski, 2020), has been described in a wide range of theoretical models that have addressed comprehension difficulties, the component processes, the integral components of inferential processes and background knowledge (Kendeou, McMaster & Christ, 2016). As Kendeou et al. (2016) argue, the construction of meaning during reading comprehension is dependent on two integral components of inferential processes. They facilitate meaning extraction from the text and sources of knowledge that expedite the extraction and construction of meaning. Deriving and using the overall meaning of a text in both its explicit and implicit modes require the construction of meaning via inferential processes (Davoudi, 2005; Ghanizadeh, Pour, & Hosseini, 2017; Hall & Barnes, 2017).

Although deciphering the underlying meaning of a sentence relies on visual processing of words, identification of phonological, orthographic, and semantic representations, and word connections using syntactic rules (Perfetti & Stafura, 2014), they are not sufficient for mastering deeper comprehension and one needs to integrate the meaning across sentences, make use of relevant background knowledge, generate inferences, identify the text structure, and take into consideration the authors' goals and motives (Graesser, 2015; Srisang, Fletcher, Sadeghi, & Everatt, 2018). The combination of these processes generates a situation model as a mental representation that resonates the overall meaning of the text (Hosoda, 2017; Kintsch & Van Dijk, 1978). However, as Snow (2002) argues, factors such as reader characteristics, text properties, and the demands of the text at hand are among the interacting components that certainly have a significant role in bringing these processes to fruition.

Though many research studies consider inferencing as an isolated phenomenon and attempt to explore the kind of inferences made during reading (Noordman & Vonk, 2015), inference-oriented research considers it as a component of discourse comprehension (Vonk & Noordman, 2001). Inferencing as an inseparable and dynamic component of discourse comprehension is resonated in various models and theories of discourse comprehension (Suvorova, 2020). Walter Kintsch's Construction-Integration Theory (Kintsch, 1988, 1998; Kintsch & Van Dijk, 1978), Gernsbacher's Structure Building Framework (Gernsbacher, 1990), Zwaan's Event Indexing Model (Zwaan, Magliano, & Graesser, 1995), McKoon & Ratcliff's Minimalist Hypothesis (McKoon & Ratcliff, 1992) are the theories and models that underscore the four main aspects of mental representations in the generation of meaning: identifying exact content of the text, connecting the actual text words with those ideas they refer to (known as referential processes), connecting different pieces of text together (establishing cohesion and coherence), and building a representation of what the text is

about (processes responsible for discourse representation or mental model) (Khemlani, Byrne, & Johnson-Laird, 2018).

According to Perfetti and Stafura (2014), modern studies of reading comprehension tend to apply either of the two complementary models of the reader's situation model (Van Dijk & Kintsch, 1983) and the construction-integration (C-I) model (Kintsch, 1988). The former is concerned with an enriched level of comprehension that goes beyond the literal meaning of a text and the latter deals with cognitive dynamics of text comprehension. In this study, the C-I model is used. The reason behind preferring the C-I approach in analyzing the test items compared to other models of text comprehension is its close adherence to the identification of cohesion cues that renders interrelationship between local and distal constituents in discourse analysis. The C-I processing model represents a production system like a computer program which scans the content of the short-term or working memory (Traxler, 2011). The discourse processing system functioning according to a set of if-then rules applied to the content of the active memory buffer shapes the contents of the working memory and constructs a coherent, organized mental representation to be held permanently in the long-term memory. The C-I model advances the notion of levels in the reader's mental representation.

Though the Iranian State University TEFL Ph.D. Entrance Exam (henceforth ISUTPEE) is among the most challenging high-stakes multiple-choice (MC) tests, review of related literature shows that it has not received adequate attention from academics and researchers regarding the complexity level of cognitive processing that test-takers face to find the correct options. ISUTPEE as an MC test consists of three distinct parts of educational aptitude, a general English proficiency test, and a specialized knowledge test (henceforth SKT). The level of inferential processing might be a satisfying yardstick to assess the difficulty level of questions that engage test-takers in retrieving relevant knowledge from memory. Therefore, this study intended to explore the efficiency of the MC SKT section in ISUTPEE under the lens of Kintsch's Construction-Integration Theory (Kintsch, 1988, 1998). This SKT section of ISUTPEE includes four parts: teaching methodology, linguistics, testing, and research methodology.

2. Review of Literature

According to Perfetti and Stafura (2014), an acceptable theory of reading comprehension takes account of both cognitive and linguistic processes and also makes rigorous, verifiable predictions; however, building such a theory with the required precision is far beyond our ability due to the inherent complexity of this activity. Therefore, a wide array of theoretical models and frameworks have been proposed, each of which pinpoints specific components and processes of reading comprehension.

Component models of reading comprehension such as Simple View of Reading (Hoover & Gough, 1990) and more complex ones such as the Direct and Inferential Mediation Models (Cromley & Azevedo, 2007) depict reading comprehension as the product of decoding and language comprehension. The former involves those processes that break down written codes, especially phonological and orthographic processing and word recognition, while the latter employs those processes that help build a mental representation, such as vocabulary and inference making. These models address the identification of component skills that clarify reading comprehension performance, for example, word decoding (Ehri, 2014), vocabulary knowledge (Quinn, Wagner,

Petscher, & Lopez, 2015), working memory (Sesma, Mahone, Levine, Eason, & Cutting, 2009), and prior knowledge (Kintsch, 1988).

Inference generation models such as Construction–Integration (C-I) model (Kintsch, 1988, 1998), the Structure Building model (Gernsbacher, 1995), Landscape model (Linderholm, Virtue, Tzeng, & van den Broek, 2004), Resonance model (Albrecht & Myers, 1995), the Event-Indexing model (Zwaan, Langston, & Graesser, 1995), the Causal Network model (Langston & Trabasso, 1999; Suh & Trabasso, 1993; Trabasso & Suh, 1993), and Constructionist model (Graesser et al., 1994) address the identification of various cognitive processes employed in reading and are concerned with the construction of mental representations during reading.

The C-I model (Kintsch & van Dijk, 1978), as one of the most influential reading comprehension models, describes reading comprehension as the generation of a coherent mental representation or situation model as a consequence of activation and integration of text information and relevant background knowledge. Kintsch introduced the basics of reading comprehension in 1988 and later he advanced the original model of reading comprehension in 1998. The general processing framework for cognition regarded as the most comprehensive and well-formed model for text comprehension (McNamara & Magliano, 2009) is founded on Kintsch and van Dijk's (1978) psychological theory of discourse comprehension.

Kintsch's C-I model shifted the attention from the memorial-based descriptions of text in 'schema-based models of comprehension' (Rumelhart, 1977) towards those processes and strategies that constitute comprehension. This theory intends to represent the iterative processes as central to comprehension in mapping current input to prior discourse context. The C-I model posits that comprehension goes beyond the links between explicit discourse constituents. This necessitates the generation of inferences that incorporate relevant schemata into the mental representation. The process of making inferences gives rise to the construction of a situation model of the text in the readers' memory (Kintsch, 1988; Tzeng, Van Den Broek, Kendeou, & Lee, 2005). Simply put, a deep comprehension of discourse calls for grasping not only the explicitly mentioned constituents in the content of a text, but also for the apprehension of referenced and implied situations (McNamara & Magliano, 2009).

The notion of the situation model has received increasing attention among researchers due to its importance in discourse comprehension. The inferences that result in the construction of situation models (Graesser et al., 1994), the representation of their nature and the way they mirror spatial, temporal, causal, and motivational relationships (Zwaan & Radvansky, 1998), and the impact of bottom-up processes on building these models (e.g., Graesser et al., 1994; Magliano & Radvansky, 2001; Myers & O'Brien, 1998) are among the aspects of C-I that have been studied.

The C-I model, as its name conveys, explains the process of reading comprehension in two phases: construction and integration. The construction phase refers to the process of information activation of both relevant and irrelevant knowledge in regard to the intended context. Kintsch (1998) argues that automatic memory activates a retrieval-based process to comprehend discourse and current input, the previous sentence or proposition, related knowledge. Reinstatements from the prior text are potential sources of knowledge activation for each round of input.

In the second phase, the integration process, activation of information occurs across the network until the activation values for propositions settle. Activation sources that are limited by the capacity of the working memory, iteratively integrate via constraint satisfaction mechanism. This

brings about a larger extent of activation for those concepts that are linked to other concepts. And an activation loss occurs for those peripheral concepts that have fewer connections in mental representations. A normalization process provoking the activation of connected concepts and the loss of activation of less connected follows constraint satisfaction. The dynamic iterations result in the emergence of particular activation patterns for the whole text or across a group of sentences or propositions in the C-I process (McNamara & Magliano, 2009).

Kintsch's C-I *theory* assumes three distinct levels of representations in the reader's mind for a sentence: *the surface structure*, *the propositional textbase*, and *the situation model* (Kintsch, 1988, 1998; van Dijk & Kintsch, 1983). The surface model, as the least abstract mental representation, represents a phrase structure tree for each sentence that is constituted by the exact words in the text content and their syntactic relations (Kintsch 1998). Assumed to have only a slight effect on comprehension, the surface model, functioning as the input is captured by the interface processes to produce a set of propositions that are represented by the text.

The second level of processing is called the textbase. It results in the generation of propositions. A proposition consists of a predicate and an argument reflects a representation close to the exact wording of the original text; however, it can encompass some information that is not explicitly mentioned in the text. Therefore, the construction of the textbase representation of the text requires the loss of some information from the surface model representation.

And finally, the C-I system represents the situation model as the ultimate goal of comprehension at the highest level of abstraction in mental representations. This abstract level of comprehension process follows the assumption that people are mainly interested in knowing what happened and why rather than the exact wording of a text. However, situation model and textbase representations should be regarded as different aspects of the episodic memory for a text rather than totally separate and compartmentalized mental representations of a text (Graesser & Clark, 1985; Kintsch, 1988; van Dijk & Kintsch, 1983) owing to the fact that, as Zwaan, Magliano, et al. (1995) declare, propositions in the textbase are associated on the basis of a situation model.

According to the C-I processing model, the best retrieval cues are made by the words from the same proposition rather than those made from different propositions. Therefore, those elements that come together in a proposition are more likely to establish a tight relationship in memory and enjoy a faster rate of memory retrieval compared to that of elements from two separate propositions (Ratcliff & Mckoon, 1978). Moreover, as Forster (1970) asserts, the number of words in a sentence that are retrieved from memory is dependent on the number of propositions. He argues that individuals tend to recall more words from one-proposition sentences than from two-propositional sentences.

A number of studies have been carried out on various aspects of TEFL Ph.D. Entrance Exam (TPEE) in the context of Iran. Barati and Ahmadi (2012) investigated differential item functioning (DIF) in relation to the effects of gender and subject matter on the Special English Section of TPEE. The results of their study revealed the presence of DIF on this test; the lowest DIF was related to the Cloze test while the highest DIF was observed in the section related to language functions. Moreover, some general gender DIF patterns were explored across the subject area. Females outperformed males in three sections of grammar, language function, and cloze test while males surpassed females in vocabulary and word order sections. Most notably, it was concluded that the degree and direction of DIF are dependent on both item format and the subject area and the

interaction between the two is of significant importance.

Ahmadi, Darabi Bazvand, Sahragard, and Razmjoo (2015) investigated the content validity of the Iranian Ph.D. entrance exam of TEFL (IPEET) in light of argument-based validity and theory of action. The results from the analysis of mixed-methods data including test scores, questionnaires, and focus-group interviews, demonstrated that the IPEET test instrument suffers from validity requirement.

Amirian, Ghonsooly, and Amirian (2020) examined fairness of Special English Test of Iranian National University Entrance Exam by analyzing Differential Item Functioning (DIF) with reading comprehension section of this test. The results revealed that only 6 items in the reading comprehension skill showed DIF. Further analysis manifested that the effect size of DIF for all six items were category A or negligible.

Ashraf, Tabatabaee Yazdi, and Samir (2016) conducted a study to investigate the cognitive processes that X-Tests and C-Tests involved in the test-takers via think-aloud and retroactive interviews. The results showed that X-Test was more difficult for the test-takers and tapped participants' use of mental strategies.

Zhang and Lin (2021) examined the relationship between morphological knowledge and reading comprehension ability among college-level EFL students. The results indicated the salient mediating effect of morpheme-meaning knowledge and also the direct and indirect effects of morphological knowledge on the EFL reading comprehension via reading vocabulary breadth and lexical inference ability.

Moreover, Abbasian and Nassirian (2015) evaluated the Iranian State University EFL Entrance Examination Test (UEEET) based on Bachman and Palmer's (1996) 'Usefulness Six-faceted Model 'accommodating Reliability, Validity, Impact, Interactiveness, Authenticity, and Practicality. Analysis of data gathered through questionnaires and a structured interview with TEFL university professors and freshmen showed that both groups evaluated the exam as mainly less reliable, less practical and entailing negative impacts, while they expressed positive views towards its validity, authenticity, and interactiveness. Overall, UEEET was evaluated as a test enjoying an acceptable degree of the Usefulness criteria.

Reviewing the relevant literature reveals a disturbing dearth of studies dedicated to analyzing the inferential processes that the test items, especially those in the SKT, require the test takers to activate. Given the fact that applicants compete to gain admittance into the top-ranking universities, this study attempts to shed more light on the efficiency of MC SKTs. And for the first time, it adopted a memory-based model of discourse comprehension in assessing the complexity level of test items in ISUTPEE that is of significant impact on both professional and personal levels.

3. Purpose of the Study

This study aims at identifying the mental representation level of ISUTPEE test items in terms of inferences that test-takers are required to make to answer MC questions correctly. The study drew on the *C-I model* (Kintsch, 1988) to analyze the data. ISUTPEE, which is held once in a year, is one of the high-stakes tests that welcomes thousands of test-takers to compete for the seats in the state-funded universities all over the country. The process of admittance to the universities is planned by the Ministry of Science, Research and Technology. In recent years, 50 percent of the total score has been determined through the administration of a nationwide exam under the control

of the National Organization for Educational Testing in which all applicants have the permission to participate. A significantly important section of this exam is the SKT wherein the test takers take MC items including tests from four major fields of TEFL: Teaching Methodology, Linguistics, Testing, and Research Methodology. Then, a limited number of participants that have passed the cut-off score which is decided and determined by each particular university are invited to take part in an interview held by the faculty members of host universities that complement the first 50 percent score. And finally, the results of the interview are reported to the Assessment Organization to rank the participants according to their scores. The overall evaluation is determined through the test takers' scores in the MC exam (50%) and the remaining score has been determined based on the entire range of scores for research qualifications, educational background, and general and scientific interviews carried out by the host universities.

Despite the outstanding consequences that this exam has in the lives of stakeholders, it is noteworthy that only few studies have focused on ISUTPEE and examined it from a wide range of dimensions including validity (Rezvani & Sayyadi, 2016), washback effects (Salehi & Yunus, 2012), usefulness (Abbasian & Nassirian, 2015) and stakeholders' perspectives (Kiany, Shayestefar, Samar, & Akbari, 2013). However, to the best knowledge of the researchers, the inferencing complexity level of test items in ISUTPEE which includes a test of Academic Aptitude, a General English Proficiency Test, and an SKT has not been investigated so far. This study intends to analyze the inferential processes that the test items require the test takers to activate and the efficiency of MC SKTs.

Inferencing complexity level (ICL) or the burden that the questions impose on test-takers' memories would indicate the level of propositional representations. If the question tends to be of low ICL, those test takers with a tendency to store and organize information in memory as presented in the text are expected to be able to answer the question correctly. In other words, a test taker that crams the Ph.D. sources and amasses a huge amount of information in memory in a fragmented mode is expected to capture verbatim text words and syntax. If all the questions are of low ICL, that test-taker can easily ace the test. This would indicate that the test is not functioning appropriately. To make this point clear, two examples are provided below:

Table 1.
Test Items with Low ICL

1) Sociolinguistic competence, in Bachman's (1990) model of communicative competence, includes all of the following EXCEPT…….
1) sensitivity to naturalness 2) sensitivity to different registers 3) ability to interpret the intention of the speaker 4) ability to interpret cultural references and figures of speech
(Question 28, Testing Section, Ph.D. exam in 2017)
2) What kind of classification error occurs when a test taker is classified as a master when his or her domain score is below the cut-off score?
1) type I 2) type II 3) false positive 4) false negative
(Question 35, Testing Section, Ph.D. exam in 2017)

Detailed analysis of the wordings in the above questions in reference to one of the primary sources of questions for Ph.D. tests, that is, Fundamental Considerations in Language Testing (Bachman, 1990), reveals the fact that the exact words and their syntactic relations in the textbook are represented in the question. Therefore, this question taps into a level of representation in memory that is of low ICL. These two questions, most probably extracted from the following texts, respectively, are presented below:

1) Without attempting to identify and discuss the features of the language use situation that determine the conventions of language use, I will discuss the following abilities under sociolinguistic competence: sensitivity to differences in dialect or variety, to differences in register and to naturalness, and the ability to interpret cultural references and figures of speech (Bachman, 1990, p. 95).

2) Whenever we make a mastery/nonmastery classification decision, there are two possible types of errors that can occur. A 'false positive' classification error occurs when we classify the test taker as a master when his domain score is in fact below the cut-off score (Bachman, 1990, p. 215).

Those test takers that have just crammed the resources and have engaged in merely literal text processing might be able to easily answer such tests through word activation processes in the verbatim memory of the text.

The questions regarded as medium ICL demand higher levels of inferencing (Noorbakhsh, Ghonsooly, & Ghanizadeh, 2018), that is, they tap into those mental representations that capture the words and propositions that are encoded in memory (the textbase model). The words and propositions in the questions resonate well with those of memory storage which are highly accessible. Though the concept of the reader's mental representation and the three levels of inferencing complexity seem to be illuminating, the boundary among the levels of mental representations, especially between the textbase and the situation model is blurry and this calls for precise benchmarks to identify them. The fundamental unit of processing in the textbase model is the proposition as one complete idea including a predicate and argument(s). Instances of questions that draw on textbase level of inferencing are presented below:

Table 2.
Test Items with Textbase Level of Inferencing

1) Which of the following beliefs marks the advent of Error Analysis as the systematic investigation of second language learners' errors? 1) The idea that all learners' errors are rooted in their second language. 2) Developments in first language studies and disillusionment with contrastive analysis. 3) The change in overall philosophy behind language learning in the 1990s 4) Popularity of contrastive analysis and its usefulness *(Question 86, Testing Methodology Section (Second Language Acquisition), Ph.D. exam in 2012)*
2) All of the following are true about Grammar Translation Method EXCEPT that …… 1) the sentence is the unit of teaching and language practice

2) rules of grammar are explicitly presented
3) learners' mother tongue is suppressed
4) oral skills are disregarded

(Question 86, Testing Methodology Section, Ph.D. exam in 2012)

The test items shown in Table 2 would activate the textbase model of representations containing a proposition that represents a complete idea. Simply put, such questions require the activation of the underlying meaning of explicitly expressed information in the text. As can be seen in these two questions, the model of mental representation is dependent on the links conveyed by the predicates and the overlap between arguments. Notably, associating ideas in the textbase model is prominently facilitated through overlap between arguments. In other words, the overlap between predicates (verbs, modifiers) is not conducive to the integration of ideas. Therefore, establishing a textbase model of representation is not driven by events and action, it demands an argument overlap which results in text cohesion.

Lastly, the situation model that has attracted much more attention compared to the other two models of mental representations in discourse comprehension research (van Dijk & Kintsch, 1983) incorporates all inferences that transcend the explicitly mentioned concepts in the text. Regarding the Ph.D. exam items, those questions that bring together propositions of distinct concepts and ideas or probe the deepest meanings of the terms by urging the test takers to delve into the representations stored into their episodic memory of a text call for the presence of situation models in their memory. Hence, questions of high ICL do not merely rely on inquiring about the definition or features of a notion or process; instead, they adhere to the knowledge-base that a less active reader might not be able to come up with any mental representation for. Two examples of such question type are given below:

Table 3.
Test Items with High ICL

1) A task in which a prospective salesperson is asked to participate in a role-play to sell a product is described as one with ……
   1) A low degree of both authenticity and interactiveness
   2) A high degree of both authenticity and interactiveness
   3) A low degree of authenticity and a high degree of interactiveness
   4) A high degree of authenticity and a low degree of interactives

(Question 22, Testing Section, Ph.D. exam in 2016)

2) An implication of chaos complexity science for developing and evaluating language teaching models would be ……
   1) considering models as linear appraisal processes
   2) foregrounding certain problems while obviating others
   3) working for local models with higher certainty
   4) searching for simple local solutions to universal challenges

(Question 13, Teaching Methodology Section, Ph.D. exam in 2012)

The test items that activate situation models of representations in episodic memory basically depend on building up connections between propositions in the textbase model (Zwaan, Magliano, et al., 1995). Situation models are comprised of links among ideas in the text and connections to prior knowledge. Therefore, a less active reader assumingly structures a less coherent situation model of the discourse which results in a predominantly textbase level rather than situation level of understanding. Given the lack of information on the ICL of test items in the SKT, the research question that is to be addressed in this study is as follows:

To what extent are the test items in the SKT of ISUTPEE of low, medium, and high ICL?

4. Method

4.1. Instrumentation

In an effort to maximize the relevance of the study, ISUTPEE during 8 successive years from 2010 to 2017 were brought under close examination drawing on the C-I model of inference processing (Kintsch, 1988). To explore the required level of propositional representations in test takers' memories and capture the inferencing complexity level (ICL) or the burden that the questions impose on test-takers memories, this study drew on the C-I processing model.

4.2. Data Collection Procedure

National ISUTPEE includes three distinct sections: the Educational the Aptitude test, the general English proficiency test, and the specialized knowledge test. The number of test items for each of these fields is presented in the following table:

Table 4.
TOEFL Ph.D. Domains and Item Distribution

Year	Number of items for each domain				Total Items
	Teaching Methodology	Linguistics	Testing	Research methodology	
2017	30	18	22	20	90
2016	30	20	20	20	90
2015	20	10	---	---	30
2014	45	25	15	15	100
2013	45	25	15	15	100
2012	45	25	15	15	100
2011	55	15	15	15	100
2010	30	30	---	---	60
					Total 670

Table 4 shows the item distributions for each of the three domains in detail. A total of 670 TOEFL Ph.D. test items from 2010 to 2017 were evaluated and classified according to the ICL of C-I model of inferencing.

 ISSN 2476-5880

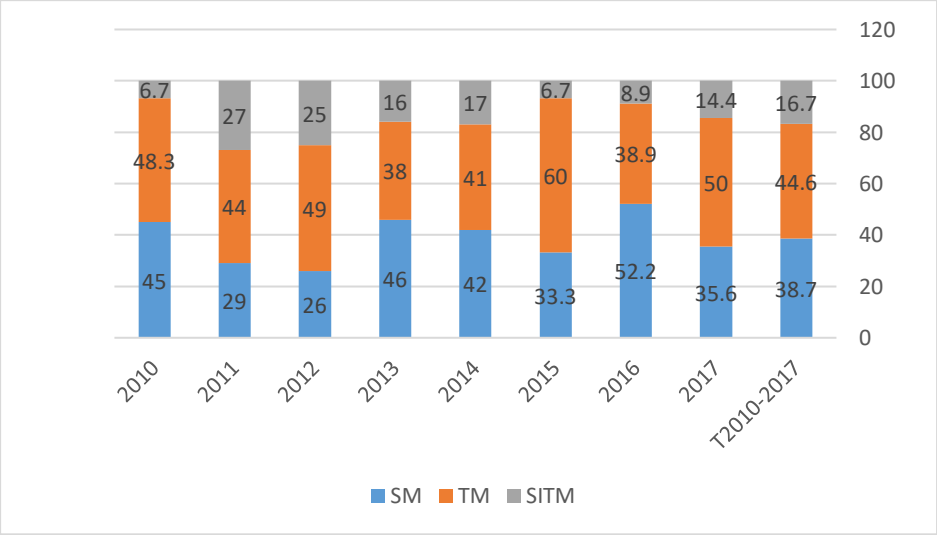
4.3. Data Analysis

The researchers conducted a detailed analysis to classify test items based on their ICL. So, those test items that represent the lowest ICL are coded 1, those with medium ICL are coded 2, and those with high ICL are coded 3. Each of these three codes corresponds to the three levels of inferencing in Kintsch's model of inference processing: the surface model, the textbase, and the situation model. To assess the accuracy of coding for the test items, two TEFL university teachers that were professionally familiar with the C-I model of inferencing independently assessed a subset of 222 (nearly 33%) test items chosen via stratified random sampling. Each of the two coders was kept blind to the coding of the researchers and the other coder. Then, the degree of inter-rater reliability was calculated to explore the extent of agreement among the three coding sets.

To compute the agreement between the codes, intraclass correlation coefficients (ICCs) were calculated. According to Fleiss (1981), ICC values lower than 0.40 can be interpreted as poor, between 0.41 and 0.75 as fair, and above 0.75 as excellent agreement. ICC estimates based on a mean-rating (k=3), absolute-agreement, 2-way mixed effects model revealed a reliability level of .91, which is an excellent interrater reliability.

5. Results

Drawing on the Kintsch's C-I model of inference processing, the test items were coded regarding their ICL. The results of analysis of the table for the whole tests in all the eight years and each of the four main sections of SKT in ISUTPEE, i.e., teaching methodology, linguistics, testing, and research methodology, are presented in the following figures respectively.



Note. SM=Surface Model, TM=Textbase Model, SITM=Situation Model

Figure 1. Percentage Frequency of Items in ISUTPEE

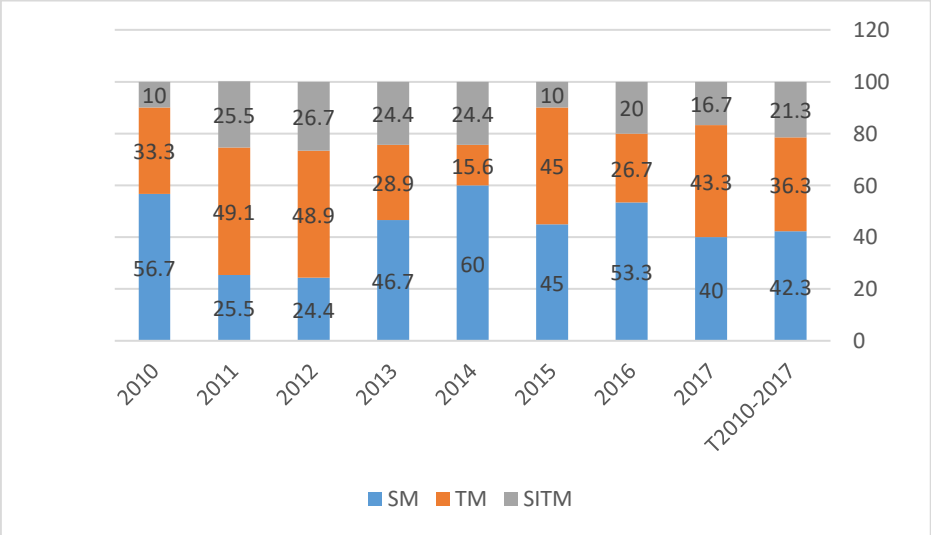
Figure 1, which presents the ICL for test items of ISUTPEEs aggregated across the domains of Teaching Methodology, Linguistics, Testing, and Research Methodology, indicates that among the test items embodied in the exam from 2010 to 2017, items of medium ICL constitute 44.6 percent (N=299) of the whole number of questions. These questions tap into the textbase model of

ISSN 2476-5880

inferencing that test takers store in their episodic memory and to a large extent correspond to the propositions presented in the text. The questions with low ICL constitute a significant portion of the test items, i.e., 38.7 percent (N=259) of 670 test items in ISUTPEE. And finally, the lowest frequent questions in the exams are those with high ICL that embody 16.7 percent (N=112) of the items.

Close analysis of the data presented in Figure 1 shows that in ISUTPEE held in 2016, 2015, and 2010, fewer than 10 items (8, 2, and 4 for each respectively) have an ICL of 3 which reflects items that tap into the situation model of inferencing. Comparing the frequency of test items with the highest ICL among these eight years shows that ISUTPEE that was held in 2012 incorporates the largest number of items with high ICL which accounts for 25 percent (N=25) of the 100 questions of the whole test and ISUTPEE with the lowest frequent test items of high ICL are those held in 2014 and 2010 which account for less than 10 percent (6.7 for both) of the questions of the whole items.

Furthermore, the largest number of items (46 items) with the lowest ICL appeared in 2016 ISUTPEE account for 52.2 percent of the whole items in the exam. Also, the largest number of test items with medium ICL was detected in 2017 ISUTPEE. Additionally, detailed analysis of the data in the figure shows that the number of items with the lowest ICL is larger than those with medium ICL in Ph.D. exams of 2016, 2014, 2013 and for the other five years the number of items with the medium ICL exceed those with the lowest ICL.



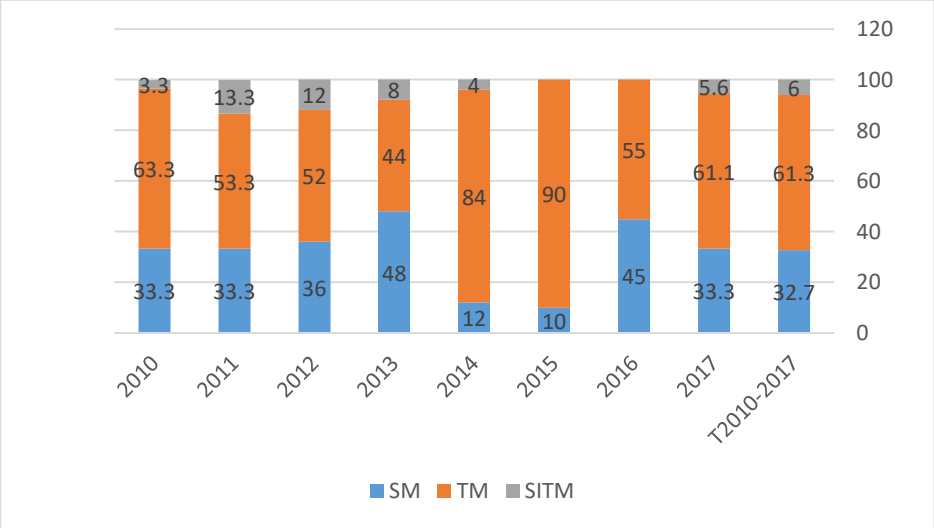
Note. SM=Surface Model, TM=Textbase Model, SITM=Situation Model

Figure 2. Percentage Frequency of Items in Teaching Methodology Section of ISUTPEE

Figure 2, which displays the ICL for Teaching Methodology Section, brings into light the following points: First, analysis of all the test items in the teaching methodology section of the test show that questions corresponding to the lowest ICL account for 42.3 percent (127 out of 300 items) of the total items and hold the highest rank in terms of frequency and the second place is held by questions of medium ICL that account for 36.3 percent (109 out of 300) of the test items, and finally the least frequent questions are those that activate the situation models in test takers memory; they account for only 21.3 percent (64 out of 300) of the questions. Second, items with the highest ICL

 ISSN 2476-5880

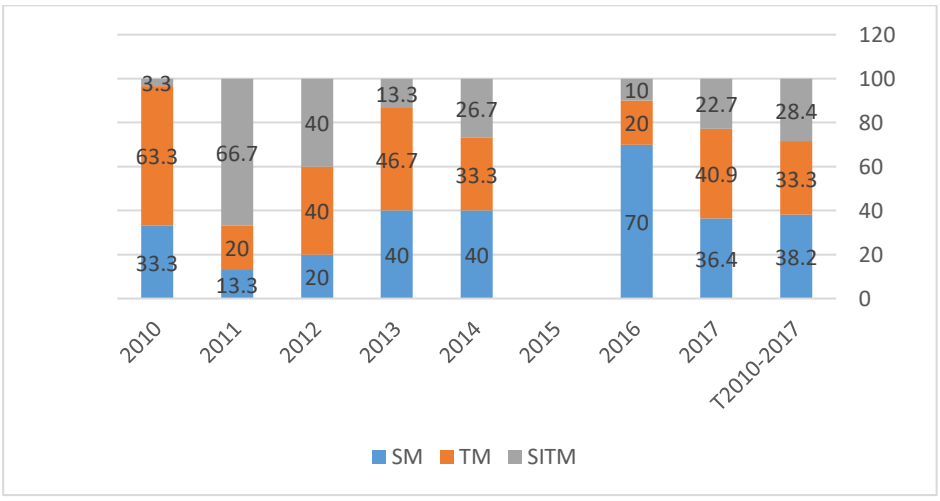
have taken a significant share of the questions in ISUTPEE of 2014 (24.4%), 2013(24.4%), 2012(26.7%), and 2011(25.5%). Third, the teaching methodology section of ISUTPEE in 2015 and 2010 is substantially dominated by questions of low and medium ICL accounting for 90 percent of the test items.



Note. SM=Surface Model, TM=Textbase Model, SITM=Situation Model

Figure 3. Percentage Frequency of Items in Linguistics Section of ISUTPEE

According to Figure 3 which presents the ICL of Linguist Section in ISUTPEEs, merely 6 percent (N=10) of 168 test items of Linguistics are of high ICL, 32.7 percent (N=55) fall into the category of low ICL, and the remaining questions which account for 61.3 percent (N=103) of questions as a considerable proportion of items are of medium ICL. As can be seen, most of the test items in this section lie within the domain of surface and textbase models of inferencing. Moreover, Ph.D. exams held in 2015 and 2016 are overwhelmed with items of low and medium ICL and not a single item checks the taker's situation model of the linguistic text content.

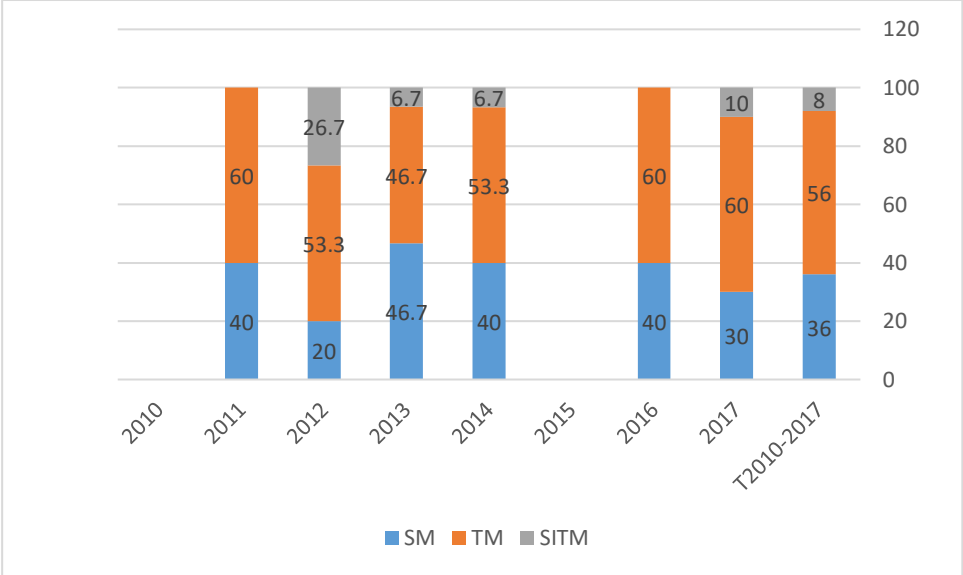


Note. SM=Surface Model, TM=Textbase Model, SITM=Situation Model

Figure 4. Percentage Frequency of Items in Testing Section of ISUTPEE

ISSN 2476-5880

Figure 4 presenting the ICL of Testing Section in ISUTPEEs shows that test items of the low, medium, and high ICL share a relatively equivalent proportion of questions, i.e., 38.2% (N=39), 33.3% (N=34), and 28.4% (N=29) of all the 102 Testing items. Figure 4 reveals that the number of items triggering situation models in test takers' mental representations is more than those that activate merely surface models of representations in 2011 and 2012. As indicated in the figure, out of 15 test items in 2012, 6 items and in 2011 out of 15 test items, 10 items have the highest ICL which account for 40 and 66.7 percent of all questions respectively. Additionally, a large number of items in 2016 ISUTPEE have a 70 percent (N=14) share of the 20 items which is a significant figure. It should be pointed out that 2015 ISUTPEE did not have a Testing Section; therefore, no data is given for this year in Figure 4.



Note. SM=Surface Model, TM=Textbase Model, SITM=Situation Model

Figure 5. Percentage Frequency of Items in Research Methodology Section of ISUTPEE

Figure 5 which shows the ICL of the test items in the Research Methodology Section of ISUTPEEs reveals that merely 8 percent of items (N=8) have high ICL and 56 percent which account for more than half of the items (N=56) are of medium ICL. Furthermore, all the items in this section are of low and medium ICL in the exams of 2016 and 2011. It should be pointed out that the items of high ICL accounting for 26.7 percent of all the items exceed those of the lowest ICL accounting for 20 percent of all test items in 2012 ISUTPEE. It should be pointed out that 2010 and 2015 ISUTPEE did not have a Research Methodology Section; therefore, no data is given for these years in Figure 5.

6. Discussion

A number of key issues emerged from the analysis of the data in this study. First and foremost, Kintsch's C-I model of inferencing, adopted as a memory-based model of discourse comprehension was applied for the first time in assessing the complexity level of test items in high stakes exams such as ISUTPEE. Although ISUTPEE has a significant impact on both professional and personal levels, there is a serious dearth of research on it in the literature review. In this study, the specialized

knowledge part of the exam was carefully scrutinized in relation to the three levels of inferencing in the C-I model of inference processing, namely, the surface model, the textbase, and the situation model of text representations. The underlying rationale to adopt the C-I model as the lens used to examine the test items in this study is the fact that achievement in this exam, to a large extent, is supposed to be tied to the test takers' ability to create coherent mental models of complex symbols in various text contents. Those who have made the representations at the highest levels of complexity, i.e., the situation models of text are expected to achieve much better results if the test items tap into representations beyond the surface models compared to those who have merely shaped a surface model of the text contents. Therefore, the ICL of test items would be useful in distinguishing those that have only structured a surface model of a text and absorbed a large amount of information in short amount of time from those that have structured a coherent mental model of the text content.

As reconstructing coherent mental representation or discourse models out of complex written symbols is an inseparable process of reading comprehension (Kintsch, 1988), the extent of success in exams such as ISUTPEE is to a large degree supposed to be dependent on the ability of test takers to integrate ideas within and across sentences in the resource books they are supposed to study. In other words, proficient readers make inferences, integrate parts of texts, draw conclusions, and make evaluations about text content (Kendeou et al., 2016). Hence, the standards for reading performance go beyond the fundamental aspects of reading, as Ehri (2014) argues, i.e., word reading and fluency.

The results of this study revealed the fact that ISUTPEE in Iran might not be well-supported by the C-I model of inference processing due to the fact that most of the questions lie in the boundary of the surface and textbase levels of mental representations, i.e., they can be simply answered without deriving and using the overall meaning of text contents read before the exam.

Though there are numerous assessment tools to demonstrate individuals' achievement in comparison to others including, formative, summative, diagnostic, curriculum-based, performance-based, portfolio, continuous, computer-assisted, online, peer, small group, standardized, criterion-based, high stakes, pre-assessment, post-assessment, and self-assessment (Timperley, 2015), the exams such as ISUTPEE are assumed to be of standardized high stakes type. However, it only determines reading performance at a particular point in time to qualify test takers for the interview section as the second phase of the exam. Given that, this exam adheres to the assessment of reading as a product. And as was reported in the result section, the test items of this high-stakes exam during the eight years from 2010 to 2017 fall short of the standards in tapping into the high ICL. Only a few items (16.7%) among 670 questions function as efficient and useful measures that tap into situation mental models of text contents; hence, more thoughtful approaches are to be adopted to develop test items in such an important exam that exerts a significant impact on individuals and serve important social and educational policy purposes.

Actually, this study suggests that level of inference activation should be recognized at the core of standards for the development of test items at the appropriate ICL since it might be of help in even reducing the length of ISUTPEE. Standardized tests are routinely applied in making tracking decisions to provide differentiated instruction suited to students' varied needs, interests, and achievement levels. Contrary to the fact that ISUTPEE plays a significant role in tracking decisions, the final decision on test taker's admittance to the universities is not taken merely based on the

results of the SKT but multiple sources of information are brought into consideration, for example, GPA, educational aptitude, general English proficiency, interview score, research articles, publications, and master's thesis grade are among the list of those factors. However, it is indisputable that SKT score is also a determining factor in the test takers' ultimate achievements.

7. Conclusion and Implications

The goal of the present research study was to scrutinize test items in SKT of ISUTPEE during eight consecutive years with regard to the Kintsch's C-I model of inference processing.

Though such large-scale gatekeeping competency tests are recognized to have intended and unintended consequences at both individual and social level, a paucity of follow-up research regarding the complexity levels of test items in this domain is noticeable. This lends urgency to the requirement that test items in this demanding large-scale national exam be of appropriate ICL and tap into those mental representations that activate situation models that are well beyond surface and textbase levels of text comprehension (Maeda, 2017). Such items intend to measure in-depth understanding of concepts in text in relation to other parts of the text content and those inferences that are beyond the explicitly mentioned concepts in the texts. As the results of this study reveal such large-scale national assessments with significant consequences seriously require the adoption of new valid methods, practices and safeguards of item development in which construction of items abide by a systematic standard procedure in organizing ICL of items. That is to say, test items with appropriate ICL should be developed since the items that summon the situation models of text representations in test takers' memory outperform the items that activate representations at the level of the surface and textbase models in distinguishing test takers that execute in-depth inferential processes in constructing meaning during reading and those that just memorize unanalyzed fragments of texts. Hence, it is assumed that the primary purpose for which the exam is developed, that is, recognizing the most competent test takers as in-depth readers of texts, is more likely to be accomplished. In other words, the probability of false-positive error that mistakenly accredit high levels of achievement to those with a surface and shallow knowledge of resource texts is minimized if more items of high ICL were used.

Development test items based on Kintsch's construction-integration theory that follow guidelines of psychometric and educational standards ensures a much more economical evaluation of the mastery of resource materials by test takers. We would suggest more empirical research studies with a focus on inference processing models of reading comprehension. However, researchers are suggested to increase the number of coders due to the fact that it might raise the reliability of the original coding done by the researchers. This study assessed the reliability of researchers' coding by using only two coders. Moreover, using test takers' responses to the test items to check out the compatibility of the three levels of coding with the mean score of each item would increase our certainty regarding the accuracy of the coding to a large extent.

References

Abbasian, G. R., & Nassirian, H. (2015). Evaluation of the Iranian State University EFL Entrance Examination Test (UEEET). *Journal of Language and Translation*, *2*(10), 43–59.

Ahmadi, A., Darabi Bazvand, A., Sahragard, R., & Razmjoo, A. (2015). Investigating the Validity

of Ph.D. Entrance Exam of ELT in Iran in Light of Argument-Based Validity and Theory of Action. *Journal of Teaching Language Skills*, *34*(2), 1–37.

Albrecht, J. E., & Myers, J. L. (1995). Role of context in accessing distant information during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1459–1468.

Amirian, S. M. R., Ghonsooly, B., Amirian, S. K. (2020). Investigating Fairness of Reading Comprehension Section of INUEE: Learner's Attitudes towards DIF Sources. *International Journal of Language Testing*, *10*(2), 88-100.

Ashraf, H., Tabatabaee Yazdi M., Samir A. (2016). An in-depth Insight into EFL University Students' Cognitive Processes of C-Test and X-Test: A Case of Comparison. *International Journal of Language Testing*, *6*(2), 101-112.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford university press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Barati, H., & Ahmadi, A. R. (2012). Gender-based DIF across the subject area: A study of the Iranian national university entrance exam. *Journal of Teaching Language Skills*, *29*(3), 1-26.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325.

Davoudi, M. (2005). Inference generation skill and text comprehension. *The Reading Matrix*, *5*(1), 106–123.

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, *18*(1), 5–21.

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss (Ed.). *Statistical Methods for Rates and Proportions* (pp. 212–236). New York, NY: John Wiley.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics*, *8*(4), 215–221.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Gernsbacher, M. A. (1995). The structure-building framework: What it is, what it might also be, and why. In B. K. Britton, & A. C. Graesser, (Eds.). *Models of text understanding* (pp. 289–311). Hillsdale, NJ: Erlbaum.

Ghanizadeh, A., Pour, A. V., & Hosseini, A. (2017). IELTS academic reading achievement: The contribution of inference-making and evaluation of arguments. European Journal of English Language Teaching, 2(2), 1–20.

Goldman, S. R., & Pellegrino, J. W. (2015). Research on learning and instruction: Implications for curriculum, instruction, and assessment. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 33–41.

Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 42–50.

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge*. Norwwod, NJ: Ablex.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text

comprehension. *Psychological Review, 101*(3), 371–395.

Hall, C., & Barnes, M. A. (2017). Inference instruction to support reading comprehension for elementary students with learning disabilities. *Intervention in School and Clinic*, *52*(5), 279–286.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing*, *2*(2), 127–160.

Hosoda, M. (2017). Learning from expository text in L2 reading: Memory for causal relations and L2 reading proficiency. *Reading in a Foreign Language*, *29*(2), 245–263.

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 62–69.

Khemlani, S. S., Byrne, R. M., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based-theory of sentential reasoning. *Cognitive science*, *42*(6), 1887–1924.

Kiany, G. R., Shayestefar, P., Samar, R. G., & Akbari, R. (2013). High-rank stakeholders' perspectives on high-stakes University entrance examinations reform: priorities and problems. *Higher Education, 65*(3), 325–340.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.

Langston, M., & Trabasso, T. (1999). Modeling causal integration and availability of information during comprehension of narrative texts. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 29–69). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Linderholm, T., Virtue, S., Tzeng, Y., & van den Broek, P. (2004). Fluctuations in the availability of information during reading: Capturing cognitive processes using the landscape model. *Discourse Processes, 37*(2), 165–186.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466.

Maeda, M. (2017). The effects of questions on EFL learners' situation models: Types of question, text levels and learners' l2 reading proficiency. *JLTA Journal*, *20*, 37–56.

Magliano, J. P., & Radvansky, G. A. (2001). Goal coordination in narrative comprehension. *Psychonomic Bulletin & Review*, *8*(2), 372–376.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation*, *51*, 297–384.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse processes*, *26*(2-3), 131–157.

Noorbakhsh, F., Ghonsooly, B., & Ghanizadeh, A. (2018). Examining a flow driven program in reading skill and its relation to higher-order reading skill (inference-making) and self-efficacy. *International Journal of Educational Investigations*, *5*(1), 1–22.

Noordman, L. G. M., & Vonk, W. (2015). Inferences in Discourse, Psychology of. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd Ed.) Vol. 12 (pp. 37–44). Amsterdam: Elsevier.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37.

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development*, *86*(1), 159–175.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. *Journal of Memory and Language*, *17*(4), 403–417.

Rezvani, R., & Sayyadi, A. (2016). Ph. D. Instructors' and students' insights into the validity of the new Iranian TEFL Ph.D. program entrance exam. *Theory and Practice in Language Studies*, *6*(5), 1111–1120.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance VI* (pp. 573–603). Hillsdale, NJ: Erlbaum.

Salehi, H., & Yunus, M. M. (2012). The washback effect of the Iranian universities entrance exam: Teachers' insights. *GEMA Online Journal of Language Studies*, *12*(2), 609–628.

Sesma, H. W., Mahone, E. M., Levine, T., Eason, S. H., & Cutting, L. E. (2009). The contribution of executive skills to reading comprehension. *Child Neuropsychology*, *15*(3), 232–246.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand Corporation.

Sosinski, M. (2020). 3 Reading from a Psycholinguistic Perspective. In *Teaching Adult Immigrants with Limited Formal Education* (pp. 30-51). Multilingual Matters.

Srisang, P., Fletcher, J., Sadeghi, A., & Everatt, J. (2018). Impacts of inferential skills on reading comprehension in Thai (L1) and English (L2). *Asia Pacific Journal of Developmental Differences*, *5*(1), 117–136.

Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, *32*(3), 279–300.

Suvorova, E. V. (2020). The principles of inference in discourse comprehension. *XLinguae*, *13*(2), 78-91.

Timperley, H. (2015). Leading teaching and learning through professional learning. *Australian Educational Leader*, *37*(2), 6–9.

Trabasso, T., & Suh, S. (1993). Understanding text: Achieving explanatory coherence through on-line inferences and mental operations in working memory. *Discourse Processes*, *16*(1-2), 3–34.

Traxler, M. J. (2011). *Introduction to psycholinguistics: Understanding language science*. John Wiley & Sons.

Tzeng, Y., Van Den Broek, P., Kendeou, P., & Lee, C. (2005). The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods*, *37*(2), 277–286.

Van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, *39*(2-3), 299–316.

Van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Vonk, W., & Noordman, L. G. M. (2001). *The psychology of inferences in discourse*. In N. J. Smelser, & P. B. Baltes (Eds.), International Encyclopedia of The Social and Behavioral Sciences, section Cognitive Psychology and Cognitive Science (pp. 7427–7435). Amsterdam: Elsevier.

Zhang, H., & Lin, J. (2021). Morphological knowledge in second language reading comprehension: Examining mediation through vocabulary knowledge and lexical inference. *Educational Psychology*, *41*(5), 563-581.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 386–397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.