한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2025, Vol. 37, No. 3, 91-101
http://dx.doi.org/10.22172/cogbio.2025.37.3.001

Original Article

OPEN ACCESS

The Effect of Response Time-Weighted Scoring on Alternate Forms Reliability: Focusing on New Reading Comprehension Items*

Sunlin Kwon[1], Hwimin Kim[1], Jungyeon Park[2], Jooyong Park[1†]

[1]Department of Psychology, Seoul National University
[2]SNU College, Seoul National University

Reading comprehension is a key path to knowledge acquisition and is closely related to academic achievement. Therefore, in order to devise an appropriate intervention plan, the test taker's actual reading ability must be accurately assessed. The present study proposed a new scoring method that reflects the examinee's response time for each item in order to increase the reliability of reading comprehension tests. The new score was obtained from the raw score multiplied by the response time-weight. The response time-weight is the ratio of an individual examinee's response time to the mean response time of all the examinees. It was hypothesized that time-weighted scores would demonstrate higher alternate forms reliability than the raw scores. In Experiment 1, 73 university students completed two alternate tests with paragraph jigsaw puzzle items that required arranging paragraphs in the correct order. The correlation between the two time-weighted scores was significantly higher than that between the two raw scores. In Experiment 2, 77 participants took two tests where they read the passage and then solved multiple-answer items without the passage. The result showed the same pattern as in Experiment 1. The results of the two experiments suggest that the time-weighted scoring method can measure the examinees' reading comprehension ability more consistently.

Keywords: Reading comprehension assessment, response time, time-weighted score, alternate forms reliability

1차원고접수: 25.01.17; 수정본접수: 25.05.27; 최종게재결정: 25.05.28

대부분의 학습은 글을 통해 이루어지기 때문에, 독해 능력은 학습의 기초이자 핵심이다. 그럼에도 불구하고, 전 세계적으로 독해 능력이 향상되기보다는 쇠퇴하고 있다는 연구 결과가 보고되고 있다(OECD, 2023; Provasnik, 2018). 2022년 국제 학생 평가 프로그램(Programme for International Student Assessment: PISA) 결과에 따르면 만 15세의 문해력 수준은 2000년대에 비해 20점 정도 하락하였다(OECD, 2023). 또한 미국 성인 중 약 48.8%는 추론을 요구하는 글을 이해하지 못하고 있으며(Provasnik, 2018), 16세에서 65세를 대상으로 한 성인 문해력 검사(Programme for the International Assessment of Adult Competencies) 결과, 2017년에 비해 유의미한 수준으로 하위권의 비율은 상승하고 상위권의 비율은 줄어들었다(National Center for Education Statistics [NCES], 2024).

* 본 연구는 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었다(No.RS-2023-00229780, 맞춤형 교육을 위한 과정 중심 평가(학습진단) 인공지능 기술 개발).

† Corresponding author: 박주용, 서울대학교 심리학과, (08826) 서울특별시 서울시 관악구 관악로 1, E-mail: jooypark@snu.ac.kr

낮은 독해 능력을 개선하기 위해 다양한 독해 교수법이 연구·개발되고 있지만(Israel & Duffy, 2017; McNamara, 2007; Mohr et al., 2020), 그로 인한 변화를 찾아보기 어려운 실정이다. 이런 맥락에서 본 연구에서는 독해 능력 향상을 위해 독해 평가 방식에 초점을 두었다. 이것은 독해 교수법이 연구·개발되더라도 평가가 정확하지 않으면 그 효과를 확인하기 어렵기 때문이다. 또한 독해 평가를 통해 사람들이 독해 과정에서 겪는 다양한 어려움을 진단하고, 이를 극복할 수 있는 개입 방안을 제시할 수 있기 때문이다.

실제로 기존의 독해 평가는 개선의 여지가 많다. 우선, 기존 독해 평가에는 선다형 문항이 주로 사용된다. 선다형 문항은 출제와 채점이 용이하기 때문에 시간과 비용을 절감하는 측면에서는 효율적이나, 지문을 이해하지 않고도 추측이나 우연에 의해 정답을 맞힐 가능성이 높다(Burton, 2001; Burton, 2005; Bush, 2006; McKenna, 2019). 또한 지문에 대해 깊게 이해하기보다 선지끼리 비교하면서 정답을 추론할 수 있으며(Magliano et al., 2007; Rupp et al., 2006), 문항을 먼저 보고 문항과 관련된 내용만을 지문에서 찾아내는 전략을 사용할 수도 있다(O'Reilly et al., 2018). 기존 선다형 문항의 문제점을 극복하는 방안으로, 여러 대안적 독해 문항이 제안되었다. 뒤에서 소개될 문단 직소 퍼즐 (paragraph jigsaw puzzle: PJP) 문항과 지문과 분리된 복수 정답 문항은 그 대표적인 예이다.

또 다른 한계는 독해 능력을 측정하는 검사가 대부분 독해 능력 외의 다른 언어 능력과 관련된 문항을 포함하는 경우가 많다는 것이다. 예를 들면, 중등 국어 시험에서는 문법 문항과 독해 문항이 함께 제시된다. 또한 미국의 대학원 입학 자격 시험(Graduate Admission Examination)의 언어 능력 검사는 독해 문항 외에도 문장에 들어갈 적절한 어휘를 고르는 문장 동등성 문항을 포함한다. 위와 같은 독해 검사 형태에서, 어휘력이 높거나 문법 지식이 많은 수험자는 관련된 문항을 빨리 풀고 독해 문항에 더 많은 여유시간을 투자할 수 있다. 어휘력과 문법 등은 독해 능력에 영향을 미칠 수 있지만, 이 요소들과 관련된 문항의 비중에 따라 독해 능력 이외의 요소가 과도하게 독해 점수에 반영될 수 있다. 이 경우 수험자들의 순수한 독해 능력만을 진단하기 어렵다.

본 연구는 선다형의 대안 문항을 사용하면서, 후자의 한계를 극복하는 방안을 제안하고 그 효과를 검증하였다. 그 방안은 문항 당 시간을 측정하고 이를 점수에 반영하는 것이다. 동일한 독해 점수를 받더라도 지문의 요지를 빠르게 파악하는 학습자와 지문에 과도한 시간을 소요한 학습자는 독해 능력에서 차이가 있다. 실제로 대부분의 독해 상황은 제한된 시간 내에 이루어지므로, 짧은 시간 안에 글의 내용을 효율적으로 이해하는 것은 독해 능력의 중요한 요소이다. 따라서 독해 문항 풀이 시간을 평가에 반영하면, 수험자의 독해 능력을 더 정확히 진단할 수 있다. 또한 어휘 및 문법과 관련된 문항이 포함된 검사라 하더라도, 다른 언어 능력과 별도로 독해 능력만을 평가할 수 있게 된다.

본 연구는 시간을 반영한 점수가 시간으로 인한 개인차를 추가로 반영하기 때문에, 측정에 따른 오차를 줄일 수 있어, 결과적으로 원점수보다 수험자의 독해 능력을 더 신뢰도 높게 측정할 것이라고 예상하였다. 교육 및 심리 검사 표준(American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014)에 따르면, 신뢰도의 넓은 의미는 검사 절차의 반복을 통해 얻은 점수의 일관성을 가리킨다. 검사의 신뢰도가 어느 정도 높아야 타당도에 대해 논의할 수 있다는 점을 고려할 때, 검사 신뢰도는 학습자 능력을 진단하기 위한 중요한 조건이다.

본 연구는 시간을 반영한 점수가 신뢰도에서 가지는 이점을 확인하기 위해 고전 검사 이론에 기반을 둔 동형검사 신뢰도를 사용하였다. 동형검사 신뢰도는 동일한 내용과 형식으로 이루어진 두 개의 동형검사 간 상관으로 측정된다. 이에 따라 시간을 반영한 점수가 원점수보다 동형검사 상관이 더 높을 것이라는 가설을 설정하였다. 이를 검증하기 위해 연구 1에서는 문단 직소 퍼즐 문항으로 구성된 두 개의 동형검사를, 연구 2에서는 지문과 분리된 복수정답 문항으로 구성된 두 개의 동형검사를 각각 풀게 하였다. 다만, 검사의 동등성에 대한 논란을 최소화하기 위해, 동형검사 신뢰도를 추정할 때, 본 연구에서는 제약이 가장 적은 일반 공통 검사 모형(congeneric tests model)을 가정하였다(Furr, 2021). 일반 공통 검사 모형에 따라 두 동형검사의 평균과 오차 변량이 동일하다고 가정하지 않되, 두 검사가 동일한 구인을 측정하고 한 검사의 진점수가 다른 검사의 진점수와 선형적으로 관련된다고 가정하였다. 연구에 대한 보다 상세한 소개에 앞서, 관련된 배경 연구 동향을 살펴보면 다음과 같다.

독해 능력과 시간 활용의 관계

독해란 심적 모형을 형성하는 것을 의미하며, 심적 모형에는 글의 핵심 정보(Kintsch, 1998) 및 적절한 추론(Graesser et al., 1994)이 포함되어 있다. 이 심적 모형을 형성하는 데 있어 배경지식(Chiesi et al., 1979; Kintsch, 1988), 구조 파악 능력(Gernsbacher, 1990) 등의 요소들이 중요한 기여를 할

수 있다.

그러나 독해는 대부분 한정된 시간 내에 이루어지므로, 시간 활용 능력 역시 독해에서 중요한 요소다. 일반적인 학습 상황을 고려해보면 이는 쉽게 이해될 수 있다. Tekin 등(2022)의 학습 시간 할당(study time allocation) 연구에 따르면, 학습 수준을 향상시키기 위해서는 이해도가 높은 부분과 낮은 부분을 구분하고, 이해가 충분히 된 영역에는 시간을 적게 투자해야 한다. 그래야 어려운 부분에 더 많은 시간을 투자할 수 있기 때문이다. 어려운 항목 역시 무제한적으로 시간을 투자하는 것은 지양해야 한다. 어려운 항목에 시간을 더 많이 들여도 성과가 더 향상되지 않는 헛수고 효과(labor-in-vain effect)가 발생할 수 있기 때문이다(Ariel et al., 2011; Dunlosky & Ariel, 2011; Metcalfe, 2002). 독해 과정에서도 역시 시간 활용은 중요한 역할을 한다. 독자가 긴 글 속에서 쉬운 내용과 어려운 내용을 구분하고 어려운 내용에 집중해야 성공적인 독해가 이루어질 수 있기 때문이다.

독해 능력에서 시간을 반영한 기존의 연구들은 주로 글을 소리 내어 읽는 시간에 주목하였다. 그 중 하나로, 청소년을 대상으로 한 문해력 검사에서는 1분당 얼마나 많은 단어를 정확하게 읽는지로 독해 능력을 평가하였다(Good & Kaminski, 2002; Wiederholt & Bryant, 2012). 단어를 읽는 속도가 지문을 읽고 이해하는 독해 능력과 높은 상관관계를 보인다는 연구 결과도 있다(Álvarez-Cañizo et al., 2015; Hosp & Fuchs, 2005; Jenkins et al., 2003). 이 연구 결과는 소리 내어 단어를 읽는 데 걸리는 시간이 독해 능력을 반영하는 간접적인 지표로 활용될 수 있음을 보여준다.

그러나 소개된 연구들은 주로 초등학생을 대상으로 단순히 단어와 문장을 읽는 속도만을 측정했다는 한계가 있다. 이에 반해, 자연스러운 독해 상황에서 시간과 독해 수준을 살펴본 연구는 찾아보기 어렵다. 예외적으로, Wallot 등(2014)이 컴퓨터 기반 시험에서 독해 점수를 독해 시간으로 나눈 점수(ratio comprehension score)를 사용한 바 있다. 이 연구에서 참여자들은 하나의 지문을 읽은 뒤 6개의 문제를 풀었고, 이 문제들의 총점수를 전체 시간으로 나눈 점수가 사용되었다. 하지만 이런 점수 부여 방식이 원점수에 비해 어떤 이점을 갖는지에 대한 논의는 이루어지지 않았다.

본 연구에서는 개별 문항에 소요되는 시간을 이용하여 점수를 산출하고, 이렇게 산출된 점수가 원점수에 비해 어떤 이점이 있는지를 동형검사 신뢰도를 통해 비교하였다. 본 연구에서 시간을 반영한 점수는 효율성 점수 공식을 변형하여 산출하였다. 일반적으로 효율성은 주어진 시간과 노력에 비해 얼마나 수행이 좋은지를 의미하며, 성과를 시간이나 노력으로 나눈 값으로 측정된다(Hoffman & Schraw, 2010; Hoffman, 2012). 본 연구에서는 이를 활용하되, 측정하기 어려운 노력 대신 수험자의 문항 풀이 시간에 초점을 두었다.

구체적으로, 본 연구에서는 문항 풀이 시간 가중치를 사용하여 점수에 시간을 반영하였다. 문항 풀이 시간 가중치란 개별 수험자의 문항 풀이 시간에 대한 전체 수험자의 평균 문항 풀이 시간의 비를 의미한다. 원점수에 문항 풀이 시간 가중치를 부여하는 것은 평균 시간을 기준으로 개별 수험자에게 가점을 주거나 감점을 하기 위해서이다. 즉, 평균 시간보다 짧은 시간에 독해 문항을 풀면 독해 능력이 높다고 보았고, 반대로 그 시간보다 더 긴 시간을 투자하면 독해 능력이 낮다고 보았다. 또한 독해 점수에 문항 풀이 시간을 바로 나누기보다, 평균 시간 대비 비율을 반영함으로써 점수의 편차가 지나치게 작아지지 않도록 만들 수 있다. 이상의 논의를 바탕으로 시간을 반영한 점수의 공식은 다음과 같다.

$$\text{시간을 고려한 점수} = \text{문항의 원점수} \times \left(\frac{\text{전체 수험자의 평균 문항 풀이 시간}}{\text{개별 수험자의 문항 풀이 시간}}\right) \cdots\cdots\cdots\cdots \quad (1)$$

선다형의 대안적 독해 문항

시간을 반영한 점수가 원점수보다 더 일관적인지를 알아보기 위해서는, 독해 문항이 수험자들의 지문 이해 수준을 충분히 반영할 필요가 있다. 따라서 선다형 문항 대신, 대안적 독해 문항을 사용하였다.

Kim 등(under review)은 문단 직소 퍼즐 문항을 대안적 독해 문항으로 제시하였다. 문단 직소 퍼즐 문항에서 수험자는 임의로 제시된 문단들의 순서를 논리적으로 배열해야 한다. 이 문항은 선다형 방식으로 사용해왔던 기존의 문단 배열 문항의 한계를 보완할 수 있다. 예를 들어, 4개의 문단의 논리적 순서를 추론해야 하는 경우, 문단 순서로 가능한 조합은 4! = 24가지인데, 선다형으로는 4개나 5개의 선택지만 제시할 수 있다. 이것은 결국 경우의 수를 줄여 답지를 보고 정답을 맞출 가능성을 높인다. 이에, Kim 등(under review)은 답지를 제시하지 않는 대신, 수험자로 하여금 직접 문단의 순서를 단답식으로 입력하게 한 뒤, 문자열 유사성 알고리즘을 사용하여 부분 점수를 제공하였다. 이 알고리즘은 거품 정렬(bubble sort)방식을 적용하여, 정답이 되기 위해 필요한 이동 횟수만큼 점수를 차감하는 방식으로 부분 점수를 제공한다.

Kim 등(under review)에 따르면, 문단 직소 퍼즐 문항은

독해 능력에 큰 영향을 미치는 구조 파악 능력을 측정할 수 있다는 장점이 있다. 실제로 문단 직소 퍼즐 문항으로 이루어진 검사와 구조 파악 능력을 측정하기 위해 Gernsbach와 Varner(1988)가 개발한 다중미디어 독해 배터리(Mulit-Media Comprehension Battery)점수 간 상관은 0.65로 높았다. 또한 거품 정렬 방식으로 부분 점수를 줄 때와, 선다형 방식으로 점수를 줄 때를 비교했을 때, 전자의 방식을 적용했을 때 동형검사 점수 간 상관이 유의미하게 높음을 관찰한 바 있다. 이에 따라 문단 직소 퍼즐 문항으로 만들어진 검사가 선다형 방식의 검사보다 독해 능력을 더 일관적으로 측정할 수 있을 것으로 가정하였다.

또 다른 대안적 독해 문항으로는, 지문과 분리된 복수정답 문항을 사용하였다. 먼저 지문과 분리된 문항이란 Ozuru 등(2007)이 제안한 문항 형태로, 지문을 읽는 단계에서는 지문만 읽고, 그 뒤에는 지문 없이 문항을 풀게 하는 문항을 의미한다. 이 방식은 학생들의 이해 수준을 효과적으로 포착하기 유리한데, 지문을 보면서 문항에 대한 정답을 찾는 전략을 활용할 수 없기 때문이다. 따라서 전체 지문에 대한 이해 수준이 점수에 영향을 줄 가능성이 높다. 실제로 사람들은 지문과 문항이 분리될 때, 지문을 보면서 문항을 풀었을 때보다 더 많은 시간과 노력을 들여 지문을 읽었다(Ferrer et al., 2017). 또한 지문 없이 문항을 풀 때는, 지문을 보면서 풀 때보다, 같은 지문에 대한 주관식 문항과 객관식 문항 간의 상관이 높았다(Ferrer et al., 2017; Ozuru et al., 2007). 문항의 형태가 달라져도 지문과 문항이 분리되면 지문을 기억하고 이해하는 능력이 동일하게 작동하기 때문이다.

본 연구는 지문과 분리된 문항으로 5지 선다형 문항을 사용하되, 정답의 개수가 최소 1개에서 최대 5개 이하가 되도록 설계하였다. 정답의 개수가 1개 이상인 복수정답 문항을 사용하면 단일 정답 문항보다 우연과 추측에 의한 정답 가능성을 최소화할 수 있다. 또한 단일 정답 문항은 복잡하고 고차원적인 사고 능력을 측정하는 데 한계가 있다는 지적도 있다(Thayn, 2011). 반면, 복수정답 문항은 깊은 학습 및 기억 유지와 관련되어 있었는데, 이것은 사람들이 여러 선지의 상대적 장점을 평가하면서, 자료에 대한 피상적인 이해를 넘어 더 세밀한 이해를 할 수 있기 때문이다(Oc & Hassen, 2024). 이상을 종합하여, 지문과 분리된 복수정답 문항이 단순 정답 찾기 전략을 방지하면서도, 깊은 지문 이해 수준을 측정할 수 있다고 가정하였다.

본 연구에서는 이 두 가지 대안적 독해 문항을 사용하면서, 시간을 반영한 점수가 신뢰도에서 갖는 이점을 알아보고자 하였다. 구체적인 연구 목적은 두 동형검사에서 시간을 반영한 점수 간의 상관이 원점수 간의 상관보다 유의미하게 더 높은지 검증하는 것이다.

한편 통상적으로 점수는 문항에 투자한 시간에 비례한다고 여겨진다. 이 점을 고려할 때, 시간을 적게 쓰도록 유도하는 채점 방식은 독해 문항에 대한 이해도 수준을 낮출 가능성이 있다. 이러한 전제가 성립한다면, 시간을 반영한 채점 방식은 많은 시간을 투자하여 지문의 내용을 깊이 이해하는 능력을 독해 능력으로 간주하는 시각에선 비판받을 수 있다. 그러나 시간을 투자한다고 점수가 반드시 높아지는 것이 아니라면 시간을 짧게 사용하도록 유도하는 방식은 어느 정도 정당화될 수 있다. 이에 따라 문항 풀이 시간과 원점수 간 상관이 유의미한지도 추가적으로 탐색하였다.

연구 1

연구 1에서는 문단 직소 퍼즐 문항으로 구성된 2개의 동형검사를 실시하였다. 이후 두 동형검사의 원점수 간 상관과 시간을 반영한 점수 간 상관을 구한 뒤 이들을 통계적으로 비교하였다. 또한 문항에 투자한 시간과 원점수 간 상관이 유의미한지도 확인하였다.

방 법

참가자

본 연구는 대학 기관심의위원회(IRB)의 승인을 받은 뒤 수행되었다(IRB No. 2212_001-011). 서울 소재 대학교의 심리학 교양과목을 듣는 학부생 77명을 대상으로 연구를 진행하였다. 이 중 문항 풀이 시간 기록에 오류가 발생한 참가자 4명을 제외하고 총 73명(남성 43명, 여성 30명)을 분석 대상으로 삼았다. 분석 참가자들은 30세 이하로, 참가 점수를 받고 연구에 자발적으로 참여하였다.

검사

연구 1에서는 8개의 문단 직소 퍼즐 문항으로 구성된 동형검사 2개를 사용하였다. Figure 1과 같이 문단 직소 퍼즐 문항에서 참가자들은 임의로 배열된 5개의 문단을 보고 논리적 순서를 추론하여 빈칸에 작성한다.

절차

참가자들은 설문 조사 플랫폼인 퀄트릭스(Qualtrics)를 통해 검사 문항을 풀었다. 문항을 푸는 동안 참가자는 화상회의

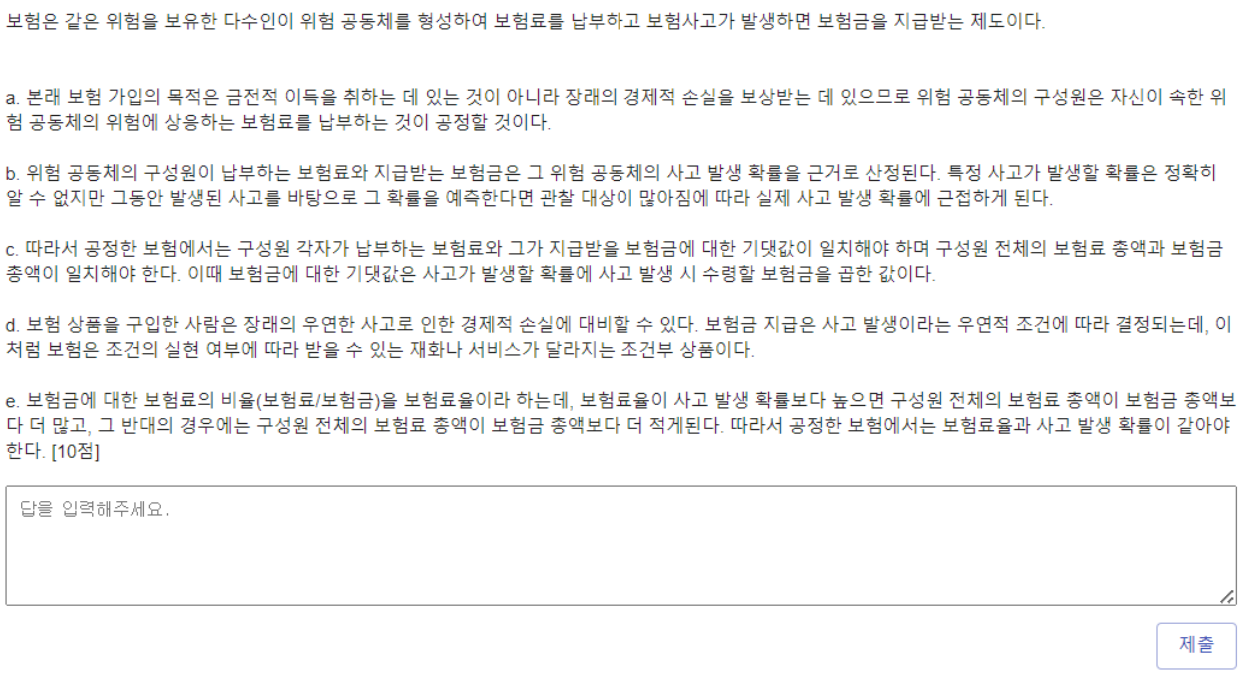

Figure 1. An example of the Paragraph Jigsaw Puzzle Item

플랫폼인 줌(Zoom)에 접속해 연구자의 안내를 받아야 했다. 이 연구는 약 60분의 시간 동안 이루어졌으며, 참가자들은 모두 동일한 절차를 수행했다.

검사를 풀기 전, 참가자들은 모두 전반적인 절차에 대한 안내를 받았다. 또한 각 문항에서 같은 점수를 받아도 문항에 투자하는 시간이 적을수록 더 높은 점수를 받을 수 있음을 안내받았다. 이를 통해 참가자들이 이해도를 높이면서도 가능한 짧은 시간 안에 문항을 풀 수 있도록 유도하였다.

안내를 받은 이후 참가자들은 총 8개의 문단 직소 퍼즐 문항으로 이루어진 검사를 두 차례 풀었다. 첫 번째 검사가 끝나면 3분간 휴식을 취하고, 이후 두 번째 검사를 풀었다. 검사에서 참가자들은 각 문항을 최대 10분까지 풀 수 있었다.

종속변수 측정

문단 직소 퍼즐 문항의 원점수는 거품 정렬(bubble sort)방식을 적용하는 자동 채점 알고리즘에 의해 산출되었다. 거품 정렬 방식은 응답이 정답이 되기 위해 필요한 이동 횟수에 따라 점수를 차감한다. 본 연구의 경우, 한 문항의 점수는 10점이고, 이동 횟수에 따라 2점씩 차감되었다. 예컨대 문항의 정답이 abcde라면, abced와 같은 응답은 정답이 되기 위해 1회의 이동이 필요하다. 이에 따라 2점이 차감되어 8점을 받을 수 있다. 만약 응답이 edcba라면, 정답까지 이동하는데 총 10회의 이동이 필요하므로 10 - (2*10)으로 -10점을 받는다. 거품 정렬 방식으로 문항을 채점했을 때, 각 문단 직소 퍼즐 문항에서 원점수의 범위는 -10점에서 10점이 된다. 단, 본 연구에서는 음수 점수를 없애기 위해 원점수에 10점을 더하였다. 결과적으로 각 문항 점수의 범위는 0점에서 20점까지이다. 8개의 문항을 각각 이러한 방식으로 채점한 다음, 이들을 합산하여 각 검사의 원점수를 산출하였다.

각 문항에서 시간을 반영한 점수는 식 (1)에 따라 원점수에 전체 참가자의 평균 문항 풀이 시간과 해당 참가자의 문항 풀이 시간의 비율을 곱해 산출했다. 여기서 문항 풀이 시간은 학습자들이 해당 문항이 있는 페이지에 머무는 총 시간이다. 퀄트릭스 사이트에서는 한 페이지 당 한 문항이 제시되며, 참가자가 각 문항이 있는 페이지에 도달한 시점부터 다음 페이지로 이동할 때까지의 응답시간이 자동으로 기록된다. 참가자들은 문항에 시간을 적게 쓸수록 점수가 높아진다는 점을 안내받았으므로, 다음 페이지로 이동한 시점을 문항 풀이가 완료된 시점으로 간주하였다. 시간을 반영한 점수 역시 8개 문항의 점수의 총합으로 산출되었다.

결과 및 논의

연구 1에서 참가자들의 점수와 문항 풀이 시간에 대한 기술통계는 Table 1에 제시되었다. 각 검사의 원점수는 160점이 만점이었다. 시간을 반영한 점수는 원점수에 비해 표준 편차가 커졌는데, 이것은 기존의 원점수에 가중치가 부여되어 점수폭이 넓어졌기 때문이다.

두 검사 점수 간의 관계는 산포도(Figure 2)를 통해 시각

Table 1. Descriptive statistics of results of the Experiment 1

	1st Test	2nd Test
Raw score	120.16(13.06)	137.62(13.30)
Time-weighted score	146.62(49.09)	168.31(52.60)
Response time	158.75(43.79)	120.15(33.61)

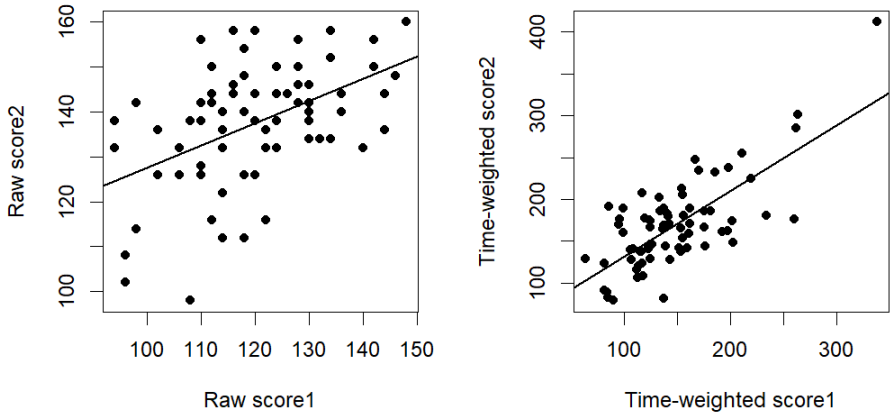

Figure 2. Scatter plot for raw score and time-weighted score

적으로 제시하였다. 동형검사 신뢰도를 확인한 결과, 두 검사의 원점수 간 상관은 .49($r(71)$ = .49, p < .001), 시간을 반영한 점수 간의 상관은 .73($r(71)$ = .73, p < .001)이었다. 두 상관계수의 차이가 통계적으로 유의한지 확인하기 위해 Fisher Z 검정을 수행하였다. 그 결과 단측 검정 시에 시간을 반영한 점수의 상관이 원점수의 상관보다 통계적으로 유의미하게 더 컸다(z = 2.36, p = .009). 이 결과는 시간을 반영한 점수의 상관이 원점수의 상관보다 높다는 본 연구의 가설을 지지한다.

추가적으로, 두 검사에서 문항 풀이 시간과 원점수 간의 상관관계를 확인하였다. 그 결과, 검사 수준에서 문항 풀이 시간과 원점수 간의 상관은 유의미하지 않았다. 문항 수준에서 확인했을 때도, 총 16개 문항 중에서 14개의 문항은 문항 풀이 시간과 원점수 간 상관이 유의미하지 않았다. 나머지 2개의 문항 중 하나는 오히려 문항에 투자한 시간과 원점수 간에 부적 상관이 유의했다($r(71)$ = −.25, p = .03). 다른 한 문항의 경우 문항 풀이 시간과 원점수 간 상관이 .35로 유의했다($r(71)$ = .35, p = .003). 이를 종합하면, 문항 풀이 시간과 점수 간의 비례관계가 성립한다고 보기 어렵다.

연구 1의 결과는 문항 풀이 시간을 고려한 가중치를 부과하면, 수험자의 독해 능력이 더 일관성 있게 평가됨을 시사한다. 이 발견은 문항 풀이 시간을 쉽게 측정할 수 있는 컴퓨터화 검사에서 중요한 함의를 가진다. 그러나 이러한 결과가 다른 유형의 독해 시험에서도 나타나는지 확인할 필요가 있다. 이를 위해 지문과 분리된 복수정답 문항을 활용한 연구 2가 수행되었다.

연구 2

연구 2에서는 지문과 분리된 복수정답 문항으로 구성된 동형검사 2개를 활용하였다. 서론에서 소개된 바와 같이, 이 문항은 제한 시간 동안 지문을 읽은 뒤 지문 없이 제시된다. 각 문항은 기존의 선다형과 달리 정답이 하나 이상일 수 있다. 연구 2에서도, 연구 1과 마찬가지로 두 동형검사에서 원점수 간의 상관과 시간을 반영한 점수 간의 상관을 비교하였다.

방 법

참가자

서울 소재 대학교의 학부생 77명(남성 44명, 여성 33명)을 대상으로 연구를 진행하였다. 학부생들은 모두 30세 이하였다. 마찬가지로 참가자들은 심리학 교양과목을 듣는 학부생들로, 모두 참가 점수를 받고 자발적으로 참가를 희망하였다.

검사

두 개의 동형검사는 각각 4개의 지문으로 구성되어 있으며, 하나의 지문과 관련된 문항은 총 2개였다. 즉 학생들은 하나

의 검사당 4개의 지문과 8개의 문항을 풀어야 했다. 문항은 지문과 분리되어, 지문을 읽은 뒤에 다음 페이지로 넘어가면 지문 없이 문항을 풀게 된다. 여기서 문항은 복수정답 문항을 사용하였다. 복수정답 문항은 제시된 5개의 선지 중 정답의 수가 최소 1개에서 최대 5개인 문항을 의미한다. 사람들은 복수정답 문항에서 정답이라고 생각되는 선지를 빈칸에 적게 된다.

절차

참가자들은 연구 1과 동일하게 줌에 접속한 상태에서 퀄트릭스 사이트를 통해 검사 문항을 풀었다. 단, 검사가 시작되기 전 지문과 문항이 분리되어, 지문이 있는 페이지 다음으로 넘어가면 다시 지문을 읽을 수 없다는 점을 안내하였다. 연구 2에서 참가자들은 첫 번째 검사에서 문항을 푼 이후 5분간 휴식한 뒤 두 번째 검사를 풀었다. 각 검사에서 참가자들은 먼저 최소 100초부터 최대 10분까지 지문을 이해하는 시간을 가졌다. 이후 참가자가 특정 시점에서 다음 페이지로 자유롭게 넘어가면, 지문과 관련된 복수정답 문항 2개를 풀 수 있었다. 참가자는 지문 없이 각 문항에서 답이라고 생각되는 선지를 모두 골라 빈칸에 주관식으로 작성했다. 이후 나머지 지문에 대해서도 같은 과정이 반복되었다.

종속변수 측정

복수정답 문항에서 원점수를 채점할 때 다음과 같은 방식으로 부분 점수를 제공했다. 부분 점수를 줄 때 기본 원칙은 정답을 정답으로 인식했는지와 오답을 오답으로 인식했는지를 모두 점수 계산에 반영하는 것이다. 예컨대 선지가 a, b, c, d, e인 문항에서 정답이 a, b일 때, 참가자의 응답이 a, c인 경우를 생각해볼 수 있다. 정답의 측면에서, 응답에서 정답 중 일부인 a를 선택했으므로 1점을 부여할 수 있다. 그러나 정답 중 하나인 b를 선택하지 않았으므로 0점을 부여한다. 오답의 측면에서 볼 때, 응답에서 오답인 d, e를 선택하지 않았으므로 각각 1점씩 2점을 부여할 수 있다. 그러나 오답인 c를 선택했으므로 0점을 부여한다. 이에 따라 이 문항에서의 점수는 총 3점이 된다. 이 원칙에 따라 참가자는 한 문항당 0점에서 5점까지를 받을 수 있다. 각 지문에 대한 이해도 수준을 측정하는 문항은 총 2개로, 지문마다 2개의 문항의 점수를 합친 값을 원점수로 산출했다. 각 검사의 원점수는 4개의 지문마다 산출된 점수의 총합으로 계산되었다.

각 문항의 시간을 반영한 점수는 식 (1)에 따라 원점수에 전체 참가자의 평균 문항 풀이 시간과 해당 참가자의 문항 풀이 시간의 비율을 곱해 계산하였다. 이때, 문항 풀이 시간은 지문이 있는 페이지에 머문 시간을 의미했다. 문항 없이 지문을 이해하는 시간이 실질적으로 문항을 풀기 위해 필요한 시간이라고 할 수 있기 때문이다. 연구 1과 마찬가지로 각 지문이 있는 페이지에 도달한 시점부터 다음 페이지로 이동할 때까지의 응답 시간이 자동으로 기록되었고, 이 시간을 공식에 반영하였다. 각 검사의 시간을 반영한 점수 역시 4개의 지문마다 산출된 점수의 총합으로 계산되었다.

결과 및 논의

연구 2의 기술 통계는 Table 2에 제시되었다. 원점수의 총점은 40점이었다. 연구 1과 마찬가지로 시간을 반영한 점수는 원점수에 식 (1)과 같은 가중치가 부여되었으며, 이로 인해 원점수에 비해 점수폭이 더 넓어지고 표준편차도 증가하였다.

두 검사 점수 간의 관계는 산포도(Figure 3)로 제시하였다. 연구 2에서 동형검사 신뢰도를 확인한 결과는 다음과 같다. 두 검사의 원점수 간 상관은 0.31이었고 ($r(75) = 0.31$, $p = .006$), 시간을 반영한 점수 간의 상관은 0.55이었다 ($r(75) = 0.55$, $p < .001$). Fisher Z 검정 결과 단측 검정 시 두 상관계수는 유의미한 차이가 있었다($z = -1.86$, $p = .03$). 따라서 복수정답 문항에서도 시간을 반영한 점수가 원점수보다 더 일관적으로 나타났다고 볼 수 있다.

연구 1과 마찬가지로, 문항 풀이 시간과 원점수 간의 상관을 확인하였다. 그 결과 검사 수준에서는 물론 8개의 지문 각각에서도 문항 풀이 시간과 원점수 간의 상관은 유의미하지 않았다. 따라서 연구 2에서도 역시 참가자들이 지문에 투자한 시간이 늘어날수록 이해도가 그만큼 향상되었다고 보기 어렵다.

Table 2. Descriptive statistics of results of the Experiment 2

	1st Test	2nd Test
Raw score	25.94(4.19)	27.75(4.05)
Time-weighted score	29.27(9.32)	30.84(8.72)
Response time	236.01(82.77)	206.20(58.19)

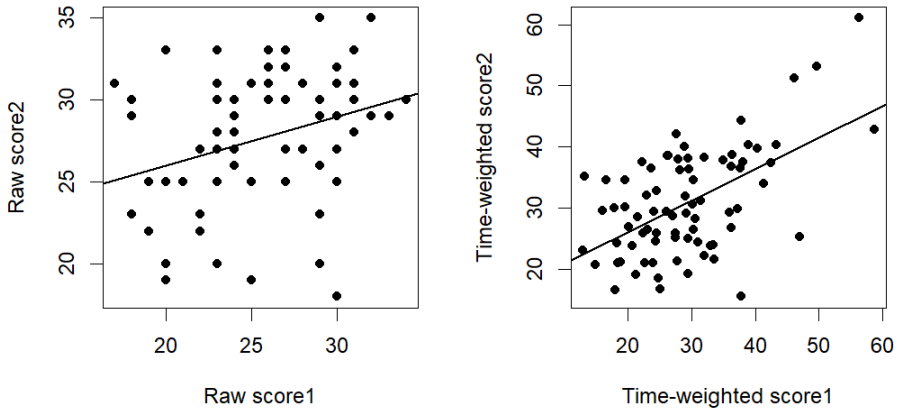

Figure 3. Scatter plot for raw score and time-weighted score

종합논의

본 연구에서는 두 독해 검사에서 시간을 반영한 점수가 원점수보다 더 일관적인지를 검증하였다. 이를 위해 연구 1에서는 문단 직소 퍼즐 문항으로, 연구 2에서는 지문과 분리된 복수정답 문항으로 이루어진 두 개의 동형검사를 사용하였다. 그 결과, 두 형태의 검사에서 모두 시간을 반영한 점수는 원점수보다 동형검사 점수 간 상관이 유의미하게 더 높았다.

이 결과는 대안적 독해 시험에서 시간을 반영한 점수가 원점수보다 더 일관적임을 보여준다. 이것은 원점수로 구분되지 않은 독해 능력의 개인차를 포착할 수 있기 때문으로 보인다. 기존 점수에서 비슷한 독해 능력을 가진 것으로 분류되었던 사람들은, 이해도에 비해 시간을 얼마나 잘 사용했는지를 기준으로 서로 다른 능력을 가진 것으로 구분될 수 있다. 이를 통해 독해 능력의 개인차가 더 잘 반영되어 동형검사 점수 간 상관이 더 높아졌을 수 있다.

이상의 결과는 다음과 같은 시사점을 갖는다. 첫째, 독해 시험에서 기존 점수보다 독해 능력을 더 일관적으로 측정할 수 있는 방안을 제안하였다. 점수에 더해 문항 풀이 시간을 고려한 점수를 산출할 때, 기존 점수보다 수험자들의 현재 독해 능력에 대한 정확한 정보를 제공할 가능성이 높다. 따라서, 본 연구에서 제안한 시간을 반영한 점수는 원점수를 이용할 때보다 독해 능력을 더 안정적으로 평가하는 방안이 될 수 있다. 컴퓨터를 통해 독해 시험에서 사용한 시간을 자동으로 측정할 수 있다는 점 역시 이 평가 방식의 적용 가능성을 높일 수 있다.

둘째, 본 연구는 독해 평가에 있어 시간을 반영하는 새로운 방안을 제시하였다. 기존 연구들에서는 단순히 단어를 정확하게 읽는 속도에 초점을 맞춘 반면, 본 연구는 실제 지문을 이해하는 데 투자한 시간을 독해 능력 평가의 핵심 요소로 삼아, 독해 능력 측정의 범위를 확장하였다. 특히, 기존 연구에서는 주로 초등학생을 대상으로 시간을 측정했다면, 본 연구는 성인 독해 평가에도 시간 요소를 적용하여, 성인 독해 능력 평가에 새로운 관점을 제시하였다. 이 결과는 독해 평가에서 수험자가 소요하는 시간을 반영할 때 얻을 수 있는 장점을 보여주며, 이는 향후 독해 능력 측정 시 시간 요소를 어떻게 반영할지 탐구하는 후속 연구에 기초 자료로서 활용될 수 있을 것이다.

본 연구의 결과는 다음과 같은 후속 연구를 통해 뒷받침될 필요가 있다. 우선 더 많은 형태의 시험에서 같은 결과가 일반화되는지 확인이 필요하다. 선다형 문항과 같은 보편적인 시험에서도 시간을 반영한 점수가 원점수보다 독해 능력을 더 일관적으로 평가하게 만드는지 탐색할 필요가 있다. 또한 시간을 반영한 점수가 기존 점수보다 왜 더 일관적으로 나타나는지 그 기제에 대한 구체적인 설명이 더 필요하다. 시간을 반영한 점수가 구체적으로 어떤 독해 능력의 요소와 관련되어 있는지 확인할 필요가 있다.

특히 메타인지적 판단과 시간을 반영한 점수 간의 상관관계를 추가적으로 탐구해 볼 수 있다. Tekin 등(2022)은 학습 시간을 효율적으로 배분하는 데 있어 자신의 이해도 수준을 평가하고, 이해도가 높아지지 않는 시점을 판단하는 메타인지적 판단이 중요하다고 설명하였다. 이러한 메타인지적 판단은 시간을 반영한 점수와 깊은 연관이 있을 가능성이 높다. 따라서 참여자들이 각 지문의 난이도에 따라 시간을 적절하게 할당했는지 추가적으로 분석하여, 메타인지적 판단이 시간을 반영한 점수에 영향을 주는지 탐색할 필요가 있다. 이러한 후속 연구들은 독해 능력을 더 정확하고 일관적으로 평가하는 데 중요한 통찰을 제공할 것으로 기대된다.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Ariel, R., Al-Harthy, I. S., Was, C. A., & Dunlosky, J. (2011). Habitual reading biases in the allocation of study time. *Psychonomic Bulletin & Review, 18*(5), 1015-1021.

Álvarez-Cañizo, M., Suárez-Coalla, P., & Cuetos, F. (2015). The role of reading fluency in children's text comprehension. *Frontiers in psychology, 6*, 1810.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41-50.

Burton, R. F. (2005). Multiple choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72.

Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education, 14*(4), 398-404.

Chiesi, H. I., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 275-290

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 899-912.

Ferrer, A., Vidal-Abarca, E., Serrano, M. Á., & Gilabert, R. (2017). Impact of text availability and question format on reading comprehension processes. *Contemporary Educational Psychology, 51*, 404-415.

Furr, R. M. (2021). *Psychometrics: an introduction.* SAGE publications.

Gernsbacher, M. A., & Varner, K. R. (1988). *The multi-media comprehension battery* (No. 88-07). Tech. Rep.

Graesser, A. C., Person, N. K., & Huber, J. D. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Good, R., III, & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Institute for Development of Educational Achievement.

Israel, S. E., & Duffy, G. G. (Eds.). (2009). *Handbook of research on reading comprehension*. Routledge.

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist, 45*(1), 1-14.

Hoffman, B. (2012). Cognitive efficiency: A conceptual and methodological comparison. *Learning and Instruction, 22*(2), 133-144.

Hosp, M. K., & Fuchs, L. S. (2005). CBM as an indicator of decoding, word reading, and comprehension: Do relations change with grade? *School Psychology Review, 34*(1), 9-26.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729.

Kim, H., Park, J., Song, M., & Park, J. (under review). Constructed Paragraph Jigsaw Puzzle Items for Measuring Structure Building Ability. *Applied cognitive psychology*

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, Cambridge University Press

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 107-136). Lawrence Erlbaum Associates Publishers.

McKenna, P. (2019). Multiple choice questions answering correctly and knowing the answer. *Interactive Technology and Smart Education, 16*(1), 59-73.

McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theories, interventions, and technologies.* Psychology Press.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363.

Mohr, K. A., Downs, J. D., & Mohr, E. S. (2020). Mindful reading: Eye tracking evidence for goal directed instruction. *Journal of Adolescent & Adult Literacy, 64*(3), 301-310.

National Center for Education Statistics. (2024). *Highlights of the 2023 U.S. PIAAC results web report*(NCES 2024-202). U.S. Department of Education.

OECD (2023). *PISA 2022 results (Volume II): Learning during - and from - disruption.* OECD Publishing.

Oc, Y., & Hassen, H. (2024). Comparing the effectiveness of multiple-answer and single-answer multiple-choice questions in assessing student learning. *Marketing Education Review*, 1-14.

O'Reilly, T., Feng, D. G., Sabatini, D. J., Wang, D. Z., & Gorin, D. J. (2018). How do people read the passages during a reading comprehension test? The effect of reading purpose on text processing behavior. *Educational Assessment, 23*(4), 277-295.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25*(4), 399-438.

Provasnik, S. (2018). *Analyzing US Young Adults' Skills by Student and Employment Status: Methodology for a New PIAAC Variable with Initial Results* (NCES 2018-122). National Center for Education Statistics.

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

Tekin, E. (2022). Can learners allocate their study time effectively? It is complicated. *Educational Psychology Review, 34,* 717-748.

Thayn, S. (2011). An evaluation of multiple choice test questions deliberately designed to include multiple correct answers [Doctoral dissertation, Brigham Young University]. ProQuest Dissertations & Theses Global.

Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., & Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1745-1765.

Wiederholt, J. L., & Bryant, B. R. (2012). *Gray Oral Reading Tests: Fifth Edition (GORT-5)*. Pro-Ed.

시간을 반영한 채점이 동형검사 신뢰도에 미치는 영향: 새로운 독해 문항을 중심으로

권선린[1], 김휘민[1], 박정연[2], 박주용[1]

[1]서울대학교 심리학과
[2]서울대학교 학부대학

독해는 지식 습득의 핵심 경로로 학업 성취와 직결된다. 따라서 학습 개선을 위한 적절한 개입 방안을 마련하기 위해서는 독해 평가가 수험자의 실제 독해 능력을 정확히 반영해야 한다. 본 연구는 독해 평가의 신뢰도를 높이기 위해 수험자의 문항 풀이 시간을 반영하는 채점 방안을 제안하였다. 시간을 반영한 점수는 원점수에 문항 풀이 시간 가중치를 곱하여 산출되었다. 문항 풀이 시간 가중치란 특정 수험자의 문항 풀이 시간에 대한 전체 수험자의 평균 문항 풀이 시간의 비를 의미한다. 시간을 반영한 점수는 원점수에 비해 동형검사 신뢰도가 높을 것으로 예상하였다. 이를 알아보기 위해 연구 1에서는 대학생 73명을 대상으로 문단의 올바른 순서를 배열하는 문단 직소 퍼즐 문항으로 구성된 두 개의 독해 검사를 실시하였다. 이후 두 검사의 원점수 간의 상관과 시간을 반영한 점수 간의 상관을 비교하였다. 그 결과, 시간을 반영한 점수의 상관이 원점수의 상관보다 통계적으로 유의미하게 더 높았다. 연구 2에서는 77명의 참가자를 대상으로, 지문을 읽은 후 지문 없이 복수정답 문항을 푸는 두 개의 독해 검사가 실시되었다. 그 결과 연구 1과 같은 패턴의 결과를 얻었다. 두 연구 결과는 문항 풀이 시간을 반영한 채점 방식이 수험자의 독해 능력을 보다 일관성 있게 측정할 수 있음을 시사한다.

주제어: 독해 검사, 문항 풀이 시간, 시간을 반영한 점수, 동형검사 신뢰도