Original Article

한국심리학회지 : 발달
The Korean Journal of Developmental Psychology
2023, Vol. 36, No. 3, 113-136

ISSN(Print) 1229-0718
ISSN(Online) 2671-6542
https://doi.org/10.35574/KJDP.2023.9.36.3.113

OPEN ACCESS

The Korean Journal of Developmental Psychology

점수 산출 방식에 따른 K-MMSE 검사점수의 유사성 및 분류 일치도*

Received: April 5, 2023
Revised: July 15, 2023
Accepted: August 29, 2023

유현종[1], 장승민[2]
성균관대학교 심리학과/ 석사과정[1]
성균관대학교 심리학과/ 교수[2]

교신저자: 장승민
성균관대학교 심리학과
서울시 종로구 성균관로 25-2

E-MAIL:
jahngs@skku.edu

Correlations and Classification Agreements among K-MMSE Test Scores based on Different Scoring Methods

Hyeonjong Yu[1], Seungmin Jahng[2]
Department of Psychology, Sungkyunkwan University/ Graduate Student[1]
Department of Psychology, Sungkyunkwan University/ Professor[2]

* 이 논문 내용의 일부는 2022년 한국심리측정평가학회 춘계학술대회에서 발표되었음.

ABSTRACT

중노년층의 치매 선별을 위한 심리검사인 K-MMSE는 문항 점수를 단순 합산한 총점으로 인지능력의 수준을 평가하고 치매 위험군을 선별한다. 그러나 인지능력의 수준과 문항 점수 사이의 관계를 선형적으로 가정하느냐 비선형적으로 가정하느냐, 문항에 따른 가중치를 부여하느냐 그렇지 않으냐에 따라 산출되는 검사점수와 선별 분류 결과가 달라질 수 있다. 총점을 검사점수로 사용하기 위해서는 선형성과 비가중이 전제되어야 하지만 많은 심리검사들은 이러한 가정에 부합하지 않는다. 본 연구는 중노년층 6,548명을 대상으로 실시한 K-MMSE 자료를 이용하여 서로 다른 가정에 따라 산출한 네 가지 검사점수 간의 유사성과 분류 일치도를 확인하였다. 선형성을 가정하는 고전검사이론 기반의 검사점수 사이에서, 그리고 비선형성을 가정하는 문항반응이론 기반의 검사점수 사이에서 피어슨 상관계수가 높게 나타났다. 비가중 산출 방식인 총점과 부분점수모형 점수는 선형성 가정 여부가 일치하지 않음에도 분류 결과가 완전히 일치했지만 가중 산출 방식인 요인점수와 일반화부분점수모형 점수는 분류 일치도가 가장 낮았다. 또한 합산점수 분포의 비대칭성이 클수록 각 방식 간 검사점수의 유사성과 분류 일치도가 낮아지는 양상을 확인하였다. 마지막으로 검사의 특성과 목적에 부합하는 검사점수 산출 방식 선택에 대한 고려사항을 논의하였다.

주요어 : 검사점수, 총점, 요인모형, 부분점수모형, 일반화부분점수모형

대한민국의 노인 인구가 해마다 급격히 증가하면서 치매 환자의 수도 빠르게 증가하고 있다. 중앙치매센터에서 발간하는 대한민국 치매 현황(이지수 등, 2023)에 따르면 전국의 65세 이상 추정 치매 환자 수는 2017년 약 71만 명에서, 2021년 약 89만 명으로 해당 기간 연간 4만에서 5만 명이 증가했다. 치매 환자의 빠른 증가세에 따라 인지기능의 비정상적 저하를 조기에 선별하고 적절한 조치를 취하는 것의 중요성도 높아지고 있다. 특히 초기 치매 환자의 인지기능 저하는 노화로 인한 일반적인 인지기능 저하와 구분하기 어렵기 때문에 적절한 검사 도구를 사용하여 치매로 인한 인지기능 손상 가능성을 조기에 탐지하는 것이 중요하다.

Mini-Mental State Examination(MMSE, Folstein et al., 1975)은 10분 내외의 검사 시행으로 인지기능의 저하 여부를 간단히 평가하는 표준화된 검사 도구이다. MMSE는 다섯 개의 인지기능 영역을 측정하는 여러 문항으로 구성되어 있으며, 각 문항에서 요구하는 과제에 대한 정답 반응을 합산한 총점으로 검사점수를 산출한다. MMSE 총점의 만점은 30점이며 일반적으로 검사점수가 23점 이하이면 인지기능의 손상이 의심된다고 평가한다(Anthony et al., 1982). MMSE는 검사 소요 시간이 짧고 채점이 쉬워 국제적으로 임상 현장에서 치매의 조기 선별을 위해 널리 사용되어왔으며, 국내에서도 한국판이 표준화되어 사용되어왔다(예, K-MMSE, 강연욱 등, 1997). 최근에는 개정판(MMSE-2, Folstein et al., 2010)이 발간되었고 개정판의 한국판도 표준화되어 출판되었다(K-MMSE-2, 강연욱 등, 2020).

MMSE와 같은 인지검사의 점수는 수검자의 인지기능의 상대적 수준을 나타내는 것으로 해석된다. 따라서 검사자는 이 검사점수에 근거하여 수검자 인지기능의 수준을 추정하고 필요에 따라 수검자를 분류한다. 마찬가지로 검사점수의 차이는 인지기능의 수준의 차이를 그대로 반영하는 것으로 해석하는 것이 일반적이다. 그러나 검사점수의 분포가 검사 대상인 인지기능의 (가설적) 분포와 다른 형태를 가질 때에는 검사점수의 차이가 인지기능 수준의 차이를 그대로 반영하지 않게 된다. 예를 들어 개인 간 인지능력의 분포는 대칭적인데 그 검사점수의 분포는 비대칭적이라면, 같은 크기의 검사점수의 차이가 반영하는 인지능력에서의 차이는 점수 구간의 위치에 따라 달라진다. 따라서 검사점수를 해석할 때는 검사점수의 분포가 검사의 대상인 특질 수준의 분포와 일치하는 형태를 가지는지를 고려해야 하며, 필요한 경우 특질 수준의 분포와 같은 형태를 가지도록 검사점수를 변환하거나 대안적인 점수 산출 방식을 검토해야 한다.

사람들이 인지능력에서 가지는 개인차는 일반적으로 정규분포의 형태를 띠는 것으로 가정된다(Plomin, 1999). 웩슬러 지능검사(Wechsler adult intelligence scale, Fourth Edition, WAIS-IV; Wechsler, 2008)와 같은 일반적인 지능검사에서는 인지능력의 개인차를 규준집단에서의 상대적 위치로 평가하는 것이 중요하기 때문에 최종적으로 산출되는 검사점수가 정규분포의 형태를 가지도록 개발된다. 그러나 MMSE의 검사점수인 총점의 분포는 -2 이상의 높은 부적 왜도를 가지는 것으로 보고되어 왔다(예, Anthony et al., 1982; Huppert et al., 2005; O'connor et al., 1989). 이것은 이 검사의 주목적이 중등도의 인지기능 저하자를 선별하는 것이어서 난이도가 쉬운 문항으로 구성되어 있기 때문이다(Lopez et al., 2005; Mitchell, 2009; Tombaugh & McIntyre, 1992). 따라서 MMSE의 검사점수를 해석할 때는 정규분포를 가정

하는 일반적인 지능검사 점수의 해석 방식을 그대로 적용하기 어렵다.

특질 수준과 검사점수의 관계가 선형적이면 이들의 분포는 같은 형태를 가지지만 그 관계가 비선형적이면 이들의 분포는 다른 형태를 가진다. 예를 들어 인지능력의 개인차가 정규분포를 따를 때, 검사 문항과 인지능력의 관계가 선형적이면 문항 합산점수가 정규분포를 따르지만, 그 관계가 비선형적이면 합산점수는 정규분포를 따르지 않는다. 이런 경우에는 특질(인지능력)의 수준과 문항 점수의 관계를 적절히 반영하는 비선형 측정모형을 이용하여 문항 응답 자료로부터 정규분포로 가정된 특질점수를 추정할 수 있다. 이렇게 추정된 특질점수를 검사점수로 사용하면 정규분포에서 벗어난 총점을 검사점수로 사용할 때 발생하는 해석상의 문제를 줄일 수 있다. 예컨대 비선형 측정모형의 일종인 문항반응이론(item response theory, IRT)을 적용하면 특질점수를 추정하여 정규분포에 가까운 검사점수를 산출할 수 있다. 요컨대 검사점수는 특질 수준과 문항 응답의 관계가 선형적이냐 비선형적이냐에 따라 다른 방식으로 산출할 수 있다.

문항 가중치의 적용 여부도 검사점수 산출에서 고려해야 할 중요한 요소다. MMSE 총점은 검사를 구성하는 각 문항의 점수를 단순 합산한 것인데 이때 모든 문항은 가중 없이 총점에 동일하게 반영된다. 그러나 인지기능의 수준을 반영하는 정도가 문항마다 다르다면 비가중 합산점수보다 가중 합산점수가 인지능력의 수준을 더 적절히 반영할 것이다. 예를 들어 표준적인 요인모형에서는 문항 가중치가 반영된 합산점수의 형태로 요인점수를 산출할 수 있다. IRT를 적용하여 특질점수를 추정할 때도 각 문항이 인지능력의 수준을 반영하는 정도를 다르게 허용할 수 있다. 라쉬모형(Rasch model)의 일종인 일모수모형(one-parameter model)이나 부분점수모형(partial credit model, PCM; Masters, 1982)에서는 모든 문항의 변별도가 같다고 가정하고 특질점수의 추정에 문항 가중이 반영되지 않는다. 이와 달리 비라쉬모형(Non-Rasch model)인 이모수모형(two-parameter model)이나 일반화 부분점수모형(generalized partial credit model, GPCM; Muraki, 1992)에서는 문항마다 변별도가 다르며 특질점수의 추정에 문항별 가중이 반영된다.

MMSE는 문항 점수를 단순 합산하여 검사점수로 사용하며 이 점수상의 기준점(23점 이하)을 이용하여 치매 위험군을 선별한다. 그러나 인지기능의 분포에 대한 일반적인 가정과 자료에서 확인되는 MMSE 총점의 분포 특성을 고려할 때 MMSE 총점과 인지기능 수준의 관계는 비선형적이라고 보아야 하며, 따라서 이 검사의 검사점수로는 문항 합산점수보다 IRT를 이용하여 추정한 특질점수를 사용하는 것이 더 적절할 것이다. 또한 MMSE의 문항에 따라 인지기능 수준의 차이를 반영하는 정도가 다르다고 가정하는 것이 더 자연스럽다는 점에서 문항 가중치를 적용한 방식이 그렇지 않은 방식보다 더 적절한 검사점수를 산출할 것으로 기대된다. 그런데도 다른 방식의 검사점수보다 단순 합산점수가 널리 사용되어 온 것은 이 방식의 실용적 편의성이 매우 높기 때문이다. 그렇다면 인지기능의 수준과 문항 점수 관계의 선형성 가정과 문항 가중치 적용 여부에 따라 MMSE의 검사점수를 다르게 산출할 때 이들 사이에는 얼마나 큰 차이가 나타날까? 이러한 차이는 단순 합산점수 방식이 가지는 실용적 편의성을 넘어설 만큼 실질적일까?

본 연구는 전통적 방식으로 산출한 MMSE 또는 K-MMSE의 단순 합산점수와 대안적인 방식으로

산출한 검사점수들을 비교함으로써 기존의 검사점수 산출 방식의 유용성과 적절성을 검토하였다. 이를 위해 검사점수 산출 방식을 (1) 인지능력과 문항 점수 관계의 선형성/비선형성과 (2) 문항 가중치 적용 여부에 따라 네 가지로 구분하였다. 다음 절에서는 먼저 이 네 가지 검사점수 산출 방식의 차이점을 정리하였다. 이어서 실제 자료를 이용하여 네 가지 검사점수를 산출하고 산출된 검사점수와 이를 이용한 선별 분류 결과의 유사성과 차이점을 확인하였다. 이를 통해 MMSE 검사점수의 산출과 활용에 대한 시사점을 찾고 검사의 특성과 목적에 부합한 검사점수 산출 방식의 선택에 관한 일반적인 고려사항을 논의하였다.

특질 수준과 문항 점수 및 합산점수의 관계

여러 개의 문항으로 구성된 심리검사에서 문항 응답 점수를 모두 더한 합산점수는 검사점수를 산출하는 가장 단순한 방식이다. 합산점수를 검사점수로 사용하는 것은 측정값을 진점수와 오차의 합으로 이해하는 고전검사이론(classical test theory)에 근거를 둔다. 고전검사이론에 따르면 진점수에 해당하는 수검자의 특질 수준과 측정값에 해당하는 문항 점수 및 합산점수는 선형적 관계를 가진다(그림 1의 직선).

반면 문항반응이론에서는 측정모형을 통해 특질 점수를 추정하여 이를 검사점수로 사용할 수 있다. IRT의 측정모형은 문항 점수를 얻을 확률을 특질 수준에 따라 달라지는 비선형 함수(문항특성곡선)로 표현하며, 이에 따라 문항 점수를 합산한 총점의 기대값도 특질 수준과 비선형적(검사특성곡선)으로 연결된다. IRT에서 사용되는 검사특성곡선은 로지스틱 곡선의 형태를 가진다(그림 1의 곡선).

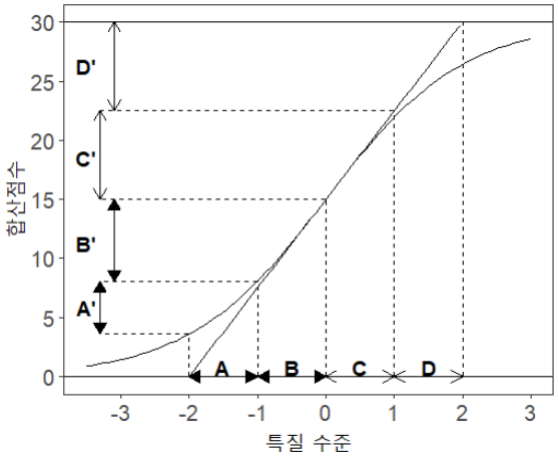

그림 1. 특질 수준과 합산점수 사이의 선형 관계와 비선형 관계

선형 관계와 로지스틱 곡선 관계의 차이는 합산점수의 중심부보다 주변부에서 두드러진다. 직선적 관계에서는 합산점수의 전 범위에서 특질 수준의 차이가 같으면 합산점수의 차이도 같다 ($C = D \rightarrow C' = D'$). 따라서 선형 함수를 사용하는 경우 합산점수의 모든 범위에서 수검자 특질 수준의 차이값과 합산점수의 차이값을 동등하게 해석할 수 있다. 그러나 특질 수준과 합산점수의 직선적 관계는 두 분포의 형태가 일치할 때만 성립한다. 또한 수검자의 특질 수준이 검사가 산출할 수 있는 합산점수의 범위를 넘어설 경우, 이 점수로는 수검자의 특질 수준을 적절히 표현할 수 없게 된다. 반면 로지스틱 곡선의 경우 중심부냐 중심을 벗어난 주변부냐에 따라 특질 수준의 차이가 같더라도 합산점수의 차이는 달라진다($A = B \rightarrow A' \neq B'$). 합산점수 분포의 중심부에서 특질 수준과 원점수의 관계는 선형에 가깝지만, 분포의 중심에서 벗어날수록 특질 수준의 차이에 따른 합산점수의 차이는 작아진다. 이와 같은 비선형적 관계로 인해 합산점수의 분포는 특질 수준의 분포와 다른 형태로 변환

된다. 예를 들어 일모수모형에서 특질 수준이 정규분포를 따를 때 합산점수는 이항분포의 형태를 가진다(Embretson & Reise, 2000).

비가중 및 가중 합산에 의한 검사점수의 산출

특질 수준과 문항 점수 사이에 선형적 관계가 가정되면 결과적으로 특질 수준과 합산점수 사이에도 선형적 관계가 가정되며 합산점수를 그대로 검사점수로 사용할 수 있다. 대표적인 합산점수는 개별 문항 점수를 단순 합산한 총점이다. 총점은 검사를 구성하는 모든 문항의 점수가 검사가 측정하는 특질 수준의 차이를 동등하게 반영한다고 가정한다. 이때는 어느 문항에서 점수를 얻었는지와 관련 없이 문항 점수가 총점에 동일하게 반영되기 때문에, 총점이 같으면 같은 수준의 특질을 반영하는 것으로 간주하며 이는 문항의 배점이 다른 경우에도 마찬가지이다. 인지능력을 측정하는 검사의 경우에는 문항 정답 반응을, 성격 및 태도와 같은 구성개념을 측정하는 검사의 경우에는 수치로 변환한 리커트 척도의 평정 값을 문항 점수로 고려할 수 있다.

총점 자체는 문항 점수를 단순 합산한 값이므로 수검자의 상대적 특질 수준에 대한 정보를 직접적으로 제공하지 않는다. 검사점수 분포 내의 수검자의 특질 수준을 파악하기 위해서는 총점을 환산점수로 변환하여 사용하는 것이 일반적이다(Lezak et al., 2012). 주로 표준점수나 백분위 값이 환산점수로 사용되며, 검사점수 분포 내에서 수검자의 상대적 위치를 단일한 숫자로 나타낸다. 총점 방식은 환산점수를 통해 수검자의 상대적 특질 수준을 쉽고 빠르게 파악할 수 있는 이점이 있어 진단 장면을 비롯하여 심리검사가 활용되는 다양한 장면에서 사용된다.

가중 합산점수 방식은 검사를 구성하는 각 문항이 검사가 측정하는 구성개념을 반영하는 정도가 다를 수 있다고 가정하고 가중치를 적용하여 합산점수를 산출한다. 따라서 검사 총점이 같더라도 가중 합산점수는 다를 수 있다. 가중치를 결정하는 방법은 크게 사전적 방식과 경험적 방식으로 구분한다(Stanley & Wang, 1968). 사전적 방식은 검사가 측정하는 구성개념과 관련된 이론 또는 전문가의 평가를 바탕으로 가중치를 결정한다(Baldwin, 2015; Burt, 1950). 사전적 방식에 의한 문항 가중은 검사의 배경이 되는 이론을 구체적으로 반영할 수 있고, 검사의 목적에 핵심적으로 부합하는 문항에 높은 가중치를 배정함으로써 합산점수에 측정영역의 상대적 중요성을 차등적으로 반영할 수 있다.

경험적 방식은 검사 실시 자료를 분석하여 가중치를 결정한다. 대표적인 경험적 방식은 다중회귀분석을 이용한다(Bobko et al., 2007; Wang & Stanley, 1970). 다중회귀 방식은 타당한 외적 준거가 존재할 때, 이 준거를 결과변수로 설정하고 개별 문항을 설명변수로 설정하여 얻은 회귀계수를 가중치로 사용한다. 이 방식은 검사가 상대적으로 소수의 문항으로 구성되어 있을 때 효과적으로 적용될 수 있다. 그러나 문항 수가 증가할수록 문항 간 독립성이 엄격하게 요구되며, 외적 준거의 선택 과정이 다소 임의적일 수 있으므로 타당한 외적 준거와 적절한 이론적 근거가 뒷받침되어야 한다.

외적 준거가 없을 때는 요인분석을 사용할 수 있다. 요인분석은 상관 행렬 또는 공분산 행렬을 이용해 자료에 반영된 요인 구조를 파악하는 기법으로, 특질 수준으로 정의되는 요인이 여러 문항에 미치는 공통적 영향의 구조적 관계를 파악할 수 있다(Bollen & Bauldry, 2011). 요인모형을 구성하는 문항 간 공분산 행렬과 요인과 문항 간 요인부

하량을 바탕으로 개별 문항에 대한 가중치를 추정할 수 있으며, 문항 점수를 가중 합산한 요인점수를 산출하여 검사점수로 사용할 수 있다.

문항 점수와 비선형 관계인 특질 수준의 추정

특질 수준과 문항 점수 사이의 관계를 선형적으로 가정할 수 없는 경우에는 IRT와 같이 비선형적인 관계를 가정하는 측정모형을 이용하여 특질 수준을 추정할 수 있다. IRT는 특질 수준에 따라 문항 반응을 얻을 확률을 나타내는 비선형 함수인 문항특성곡선을 바탕으로 수검자의 특질 수준을 추정한다. 문항특성곡선의 위치와 형태는 문항 난이도와 문항 변별도를 통해 결정된다. 문항 난이도는 해당 문항에서 점수를 획득할 확률이 0.5가 되는 지점의 특질 수준 값으로 표현되며 문항특성곡선의 위치를 나타낸다. 문항 변별도는 문항특성곡선의 문항 난이도 지점에서의 기울기 값으로 표현되며 특질 수준 연속선상의 해당 지점에서 수검자들을 변별할 수 있는 민감도를 나타낸다. 문항 변별도가 클수록 문항특성곡선은 더욱 가파른 형태를 띠며, 이 값이 0에 가까울수록 문항특성곡선은 선형 함수에 가까워진다.

IRT는 수검자의 특질 수준에 따라서 수검자가 검사를 통해 획득할 수 있을 것으로 기대되는 총점을 검사특성곡선으로 표현할 수 있다. 예를 들어 그림 1의 로지스틱 곡선은 30개의 이분 문항으로 구성된 검사의 검사특성곡선을 나타낸 것이다. 이 곡선에 따르면 특질 수준이 1인 수검자의 합산점수 기댓값은 21.93점이다. 이처럼 검사특성곡선은 특질 수준과 총점 간의 비선형적 관계를 시각적으로 보여줄 뿐만 아니라 수검자의 특질 수준을 검사자들에게 친숙한 총점 체계에서 해석할 수 있도록 돕는다(de Ayala, 2022).

합산점수 산출에서와 마찬가지로 IRT를 이용한 특질점수 추정에서도 검사점수에 개별 문항 점수가 기여하는 정도를 동등하게 할 수도 있고 차등을 둘 수도 있다. IRT에서는 모든 문항의 변별도가 같으면 특질점수 추정에 문항 가중이 반영되지 않지만, 문항마다 변별도가 다르면 문항 가중이 반영된다. 모든 문항의 변별도가 같다고 가정하는 대표적인 IRT 모형은 이분 응답 점수 문항에 적용할 수 있는 일모수모형과 다분 응답 점수 문항에 적용할 수 있는 부분점수모형이 있다. 이러한 모형에서는 검사 총점이 동일하면 추정된 특질점수도 같다. 반면 이분 문항에 적용하는 이모수모형과 다분 문항에 적용하는 일반화부분점수모형은 문항마다 변별도를 다르게 추정하는 대표적인 모형이다. 이러한 모형에서는 두 수검자의 총점이 같더라도 변별도가 높은 문항을 많이 맞춘 사람이 더 높은 특질점수를 가지는 것으로 추정된다.

특질점수 추정을 통해 검사점수를 산출하는 방식은 최근 출판되는 심리검사에 널리 적용되고 있으며 여기에는 다음과 같은 검사들이 포함된다: Woodcock-Johnson 4판(WJ IV; Schrank et al., 2014), Kaufman Test of Educational Achievement 3판(KTEA-3; Kaufman & Kaufman, 2014), Woodcock Reading Mastery Tests 3판(WRMT-III; Woodcock, 2011), Wechsler Individual Achievement Test 4판(WIAT-4; NCS Pearson, 2020), Clinical Evaluation of Language Fundamentals 5판(CELF-5; Wiig, Semel, & Secord, 2013), Bayley Scales of Infant and Toddler Development 4판(Bayley-4; Bayley, & Aylward, 2019), Goldman-Fristoe Test of Articulation 3판(GFTA-3;

Goldman, & Fristoe, 2015), Preschool Language Scales 5판(PLS-5; Zimmerman, Steiner, & Pond, 2011), Vineland Adaptive Behavior Scale 3판(Vineland-3; Sparrow, Cicchetti, & Saulnier, 2016).

다음 절에서는 실제 MMSE 자료를 이용하여 특질 수준과 문항 점수 관계의 선형성 및 개별 문항 점수 반영의 가중치 적용 여부에 따라서 네 가지 검사점수를 산출한 후 이 검사점수들 사이의 유사성과 차이점을 검토하였다. 먼저 검사점수들 사이의 선형적 관련성을 탐색하기 위해 피어슨 상관계수를 살펴보았다. 또한 검사점수를 이용한 수검자 분류 결과를 비교하기 위해 검사점수 분포에서 준거점수를 설정하여 수검자를 분류하고, 카파 계수를 구하여 분류 일치도를 검토하였다. 추가적인 이해를 위해 연령집단을 나누어 합산점수의 분포 형태가 다를 때 네 가지 검사점수의 유사성과 차이점이 어떻게 달라지는지 확인하였다.

방 법

자료 및 검사 도구

분석에는 한국고용정보원 고령화연구패널 7차자료(Korean Longitudinal Study of Ageing, KLoSA, 7th wave, 2018)가 사용되었다. 총 6,940명 가운데 인지기능 검사에 응답하지 않은 392명을 제외한 6,548명의 자료를 분석하였다. 연령의 범위는 만 55세부터 102세까지였으며 평균은 69.39세, 표준편차는 9.9세였다. 검사점수 산출 결과를 비교하기 위하여 네 개의 연령집단을 구성하였는데 첫 세 집단은 10세 단위로 묶고 85세 이상은 하나의 집단으로 묶었다. 표 1에 각 연령집단에 따른 표본 수를 제시하였다.

표 1. KLoSA 연령집단 및 성별에 따른 표본 수

연령집단	성별		합계
	남	여	
55~64세	1,060	1,409	2,469
65~74세	857	1,080	1,937
75~84세	683	977	1,660
85세~	161	321	482
합계	2,761	3,787	6,548

KLoSA의 K-MMSE 자료에는 다섯 개 측정영역의 점수가 19개 문항의 점수로 세분되어 기록되어 있다(표 2). 합산점수 방식의 검사점수는 다섯 개의 영역점수를 사용하여 가중 및 비가중 점수를 산출하였다. 특질점수 방식의 가중 및 비가중 검사점수 산출에는 문항반응이론의 모형이 적용되었으며 19개의 문항 점수가 사용되었다.

검사점수의 산출 및 비교 절차

합산점수의 산출

합산점수 중 비가중 합산점수는 각 측정영역 점수의 단순 합산으로 산출하였으며 K-MMSE의 검사점수 총점과 같다. 가중 합산점수는 다섯 개의 측정영역 점수에 대한 요인모형(선형 단일 요인모형)의 요인점수를 추정하여 산출하였다.

요인점수는 요인모형의 결과로 얻은 요인부하량을 이용하여 다양한 방식으로 추정될 수 있다(Grice, 2001). 여기에서는 회귀식을 이용한 요인점수 추정 방법을 사용하여 가중 합산점수를 산출하였다(Brown, 2015; Thurstone, 1935). 단일 요인모형은 다음과 같은 식으로 요약된다.

표 2. MMSE의 측정영역 및 KLoSA K-MMSE의 문항 구성

MMSE의 측정영역		KLoSA K-MMSE의 문항 구성		
측정영역	최대점수	문항 내용	최대점수	문항 번호
지남력 (orientation)	10	시간지남력_날짜-연월일	3	1
		시간지남력_요일	1	2
		시간지남력_계절	1	3
		장소지남력_현 위치	1	4
		장소지남력_시/구/동/번지	4	5
기억등록 (registration)	3	기억력 테스트(3개의 단어 암기)	3	6
주의집중 및 계산 (attention and calculation)	5	주의집중 및 계산(뺄셈 1)	1	7
		주의집중 및 계산(뺄셈 2)	1	8
		주의집중 및 계산(뺄셈 3)	1	9
		주의집중 및 계산(뺄셈 4)	1	10
		주의집중 및 계산(뺄셈 5)	1	11
기억회상 (recall)	3	기억력 테스트(암기한 3개의 단어 재확인)	3	12
언어 (language)	9	소지품의 용도_소지품 1	1	13
		소지품의 용도_소지품 2	1	14
		따라서 말하기(발음의 정확성)	1	15
		명령시행_종이뒤집기, 접기, 건네주기	3	16
		명령시행_읽고 눈감기	1	17
		명령시행_기분 또는 날씨에 대해 쓰기	1	18
		명령시행_제시된 그림 똑같이 그리기	1	19

$$\Sigma = \underline{\lambda}\phi\underline{\lambda}' + \Psi \quad (1)$$

식 (1)에서 Σ는 지표변수 간 공분산 행렬, $\underline{\lambda}$는 요인부하량 벡터, ϕ는 요인분산, Ψ는 측정오차간 공분산 행렬을 나타낸다. 회귀식을 이용하여 추정하는 요인점수는 각 측정영역별 가중치를 고려하여 문항 점수를 가중 합산하는 방식으로 산출된다. 여기에서는 단일 요인모형에서 사용되는 다음과 같은 식을 이용하여 요인점수를 산출하였다.

$$\hat{f} = \hat{\beta}_1\tilde{x}_1 + \cdots + \hat{\beta}_j\tilde{x}_j \quad (2)$$

식 (2)에서 $\hat{f}$는 추정된 요인점수, $\tilde{x}_j$는 j번째 문항의 편차 점수, $\hat{\beta}_j$는 j번째 문항에 대한 추정 가

중치를 나타낸다. 회귀 방식에서 추정 가중치 벡터 $\hat{\beta}$은 문항 간 공분산 행렬 Σ의 역행렬과 요인모형을 통해 추정된 요인부하량 벡터 $\hat{\lambda}$의 곱으로 표현된다.

$$\hat{\beta} = \Sigma^{-1}\hat{\lambda} \quad (3)$$

특질점수의 산출

특질점수는 문항반응모형을 이용하여 산출하였으며 모형에 의해 추정된 수검자의 특질 수준을 반영한다. 문항반응모형은 추정하는 모수와 문항의 응답 양식에 따라 구분된다. 가장 단순한 문항반응모형인 일모수모형은 문항의 난이도 모수만을 고려한다. 문항 난이도는 문항특성곡선의 변곡점에 대응하는 특질 수준 값을 바탕으로 산출되며, 값이 클수록 응답을 위해 더 높은 특질 수준이 요구되는 문항으로 간주된다. 일모수모형에서 문항 변별도는 1로 고정되거나 모든 문항이 동일한 변별도를 가지도록 하나의 모수만 추정된다. 일모수모형은 이분 자료에 적용할 수 있으며 문항 변별도가 1로 고정된 일모수모형은 다음과 같은 수식으로 표현된다.

$$P(x_j = 1|\theta, \delta_j) = \frac{e^{(\theta - \delta_j)}}{1 + e^{(\theta - \delta_j)}} \quad (4)$$

식 (4)의 θ는 수검자의 특질 수준, δ_j는 j번째 문항의 난이도 모수를 나타낸다.

이모수모형은 문항마다 서로 다른 문항 난이도와 문항 변별도를 가질 수 있는 문항반응모형이다. 문항 변별도는 서로 다른 특질 수준을 지닌 수검자를 문항이 변별하는 수준을 나타내며, 변별도 모수의 값이 클수록 문항의 변별력이 높은 것을 의미한다. 이모수모형은 이분 자료에 적용할 수 있으며 다음과 같은 수식으로 표현된다.

$$P(x_j = 1|\theta, a_j, \delta_j) = \frac{e^{a_j(\theta - \delta_j)}}{1 + e^{a_j(\theta - \delta_j)}} \quad (5)$$

식 (5)의 θ는 수검자의 특질 수준, δ_j는 j번째 문항 난이도, a_j는 j번째 문항 변별도를 나타낸다.

부분점수모형은 수검자의 정답 반응이 여러 수준으로 구분된 순서형 다분 자료에 적용할 수 있는 문항반응모형이다. 부분점수모형은 범주반응곡선을 이용하여 수검자가 문항에서 특정 반응범주에 속할 확률을 추정할 수 있다(그림 2). 두 범주반응곡선 사이의 경계 지점을 전환위치 모수(δ_j)라고 하며 하위 범주의 반응곡선과 상위 범주의 반응곡선이 같아지는 지점의 특질 수준을 나타낸다. 전환위치 모수의 값이 클수록 상위 범주 응답에 더 높은 특질 수준이 요구되는 것을 의미한다. 부분점수모형은 다음과 같은 수식으로 표현된다.

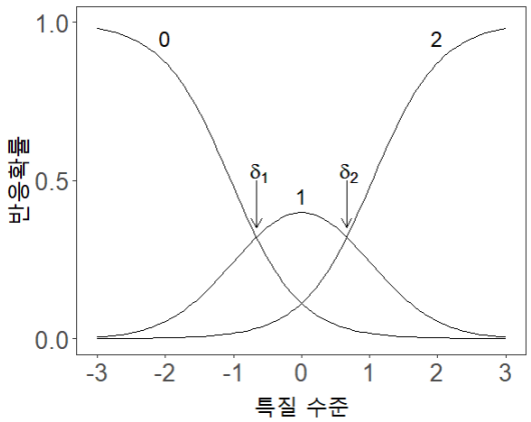

그림 2. 세 범주를 갖는 문항의 범주반응곡선

$$P(x_j|\theta,\delta_{jh}) = \frac{e^{\sum_{h=0}^{x_{jh}}(\theta-\delta_{jh})}}{1+e^{\sum_{r=1}^{m_j}[\exp\sum_{h=0}^{r_h}(\theta-\delta_{jh})]}} \quad (6)$$

식 (6)에서 x_j는 j번째 문항에 대한 문항 반응, θ는 수검자의 특질 수준, δ_{jh}는 j번째 문항의 h번째 전환위치 모수, $r...m$은 범주의 수를 나타낸다.

일반화부분점수모형은 부분점수모형에 문항 변별도 모수를 함께 고려하는 문항반응모형이다. 일반화부분점수모형에서 문항 변별도 모수는 전환위치 모수 지점의 범주반응곡선의 기울기의 값으로 산출되며, 변별도 모수의 값이 클수록 전환위치 모수 지점의 범주반응곡선 기울기는 가팔라진다. 일반화부분점수모형은 다음과 같은 수식으로 표현된다.

$$P(x_j|\theta,a_j,\delta_{jh}) = \frac{e^{\sum_{h=0}^{x_{jh}}a_j(\theta-\delta_{jh})}}{1+e^{\sum_{r=1}^{m_j}[\exp\sum_{h=0}^{r_h}a_j(\theta-\delta_{jh})]}} \quad (7)$$

식 (7)에서 x_j는 j번째 문항에 대한 문항 반응, θ는 수검자의 특질 수준, δ_{jh}는 j번째 문항의 h번째 전환위치 모수, a_j는 j번째 문항 변별도, $r...m$은 범주의 수를 나타낸다.

특질점수 추정은 크게 최대우도 추정방식과 베이지안 추정방식으로 구분된다. 두 방식 모두 문항 모수를 추정한 뒤 수검자의 특질 수준을 추정하는 방식이나, 베이지안 추정방식은 수검자의 특질 수준에 대한 사전분포를 설정한다는 점에서 최대우도 추정방식과 차이가 있다. 베이지안 추정방식의 경우 수검자의 특질 수준을 정확히 반영하지 못한 사전분포를 설정했을 때 특질점수의 잠재적인 편향이 발생할 우려가 있는 것으로 알려져 있다(de Ayala, 2022). 최대우도 추정방식은 사전분포를 요구하지 않으므로 베이지안 방식의 잠재적 편향에서 자유로우나, 원점수를 하나도 획득하지 못하거나 만점을 획득할 경우 적절한 값을 추정하지 못하는 경우가 종종 발생한다는 한계가 있다. 본 연구에 사용된 자료에서 상당한 수의 만점자가 존재하는 점을 고려하여 베이지안 방식인 기대 사후 확률(Expected A Posteriori) 방법을 사용하여 특질점수를 추정하였다.

검사점수의 비교

검사점수 간 선형적 관련성을 측정하기 위해 피어슨 상관계수를 사용하였다. 피어슨 상관계수는 두 검사점수 간 선형적 관련성을 -1에서 +1 사이의 값으로 나타낸다. 피어슨 상관계수의 최댓값은 두 검사점수 분포가 완전히 동일한 분포일 때 획득할 수 있으며, 두 검사점수 분포 형태의 차이가 클수록 가능한 최댓값은 감소한다(Cohen et al., 2003).

검사점수 산출 방식 간 분류 일치도를 살펴보기 위해 카파 계수를 사용하였다. 카파 계수는 두 검사점수를 준거점수에 따라 상호 배타적인 두 범주로 구분했을 때 두 분류 결과 간의 일치도를 나타낸다. 카파 계수는 다음과 같은 수식으로 표현된다.

$$\kappa = \frac{p_o - p_e}{1-p_e} \quad (8)$$

식 (8)에서 p_o는 두 검사점수 간 분류 일치도, p_e는 우연에 의한 분류 일치 확률을 나타낸다.

분석절차

본 연구는 각기 다른 가정을 갖는 검사점수 산출 방식에 따라 산출된 검사점수 간 선형적 관련성과 검사점수를 이용한 분류 일치도를 비교하고자 하였다. 이러한 분석을 시행하기 위해 문항 응답 자료를 이용하여 총점, 요인점수, PCM 특질점수, GPCM 특질점수를 각각 산출하였다. 총점과 요인점수는 MMSE의 5개의 측정영역을 바탕으로 합산점수를 산출하였고, PCM 및 GPCM 특질점수는 KLoSA의 19개의 문항에 대한 응답을 이용하여 특질점수를 추정하였다. 인지기능 저하를 나타내는 기준점으로 총점에서 23점 이하를 준거점수로 설정하였다. 이는 본 자료에서 제29.5 백분위에 해당하였다. 동일한 백분위를 기준으로 각 검사점수에서 제29.5 백분위 이하의 수검자를 인지기능 저하 집단으로 분류하였다. 이후 전체 수검자의 검사점수 간 피어슨 상관계수와 카파 계수를 산출하였다. 추가적으로 MMSE의 경우 연령과 검사점수 간 유의한 수준의 부적 상관이 나타나는 것으로 알려져 있어(Crum et al., 1993), 수검자의 연령집단을 구분하여 검사점수 분포가 서로 다른 조건에서 피어슨 상관계수와 카파 계수를 산출하였다. 모든 분석은 R 프로그램(version 4.3.0; R Core Team, 2023)을 이용하였으며, 요인모형은 lavaan(version 0.6-15; Rosseel, 2012), 문항반응모형은 mirt(version 1.39; Chalmers, 2012) 패키지를 이용하여 분석하였다. 분석에 사용된 자료와 R 코드는 OSF를 통해 공개하였으며 https://osf.io/uxz7y/에서 확인할 수 있다.

결 과

기술통계

각 측정영역 점수의 기술통계 및 요인분석 결과를 표 3에 제시하였다. 각 영역 점수와 총점은 전반적으로 높은 수준의 부적 왜도를 보였고 총점의 내적 일치도 신뢰도 계수는 높게 나타났다(α = 0.85). 측정영역 점수와 총점 간 상관의 범위는 0.76에서 0.91로 전반적으로 높은 수준의 변별도를 보였다.

전체 표본 및 연령집단별 총점의 분포 특성을 그

표 3. K-MMSE 측정영역점수의 기술통계

측정영역	평균	표준편차	왜도	총점과의 상관	요인부하량	가중치
지남력	9.23	1.58	-2.85	.81	0.73	0.11
기억등록	2.51	0.82	-1.61	.76	0.74	0.23
주의집중 및 계산	3.70	1.69	-1.05	.81	0.70	0.09
기억회상	1.95	1.04	-0.56	.77	0.73	0.18
언어	7.59	1.94	-1.56	.91	0.91	0.30
총점	24.98	5.83	-1.60			
α = 0.85						

주. 가중치는 식 (3)의 $\hat{\beta}_j$에 해당.

림 3에 제시하였다. 상단의 왼쪽에 제시된 총점의 히스토그램에서는 천정효과로 인해 다수의 만점자가 확인되었으며 실선으로 표현된 추정정규밀도곡선과 큰 차이를 보였다. 그 오른쪽에는 Q-Q 도표가 위로 볼록한 형태를 보이고 있어 왼쪽의 히스토그램에서와 같이 부적 왜도가 두드러지는 것을 확인할 수 있다. Q-Q 도표에 왜도의 값을 g_1으로 표기하였다(Cramér, 1946; Joanes & Gill, 1998).

가운데 단에는 연령집단별로 총점의 상자도표를 제시하였다. 이 도표에서는 연령이 높은 집단에서 평균적으로 총점이 감소하는 양상을 확인할 수 있다. 하단에는 연령집단별 Q-Q 도표를 제시하였는데, 낮은 연령집단에서 가장 정규분포 형태를 벗어나는 형태를 보이고, 연령이 높은 집단에서는 총점 분포의 비대칭성이 감소하는 양상을 보였다.

단일 차원 확인적 요인분석에서는 최대우도 방식으로 모수를 추정하였으며, 개별 측정영역에서 나타난 부적 편포를 고려하여 Satorra-Bentler 교정 검정 통계량을 확인하였다(Satorra & Bentler, 1994). 모형 적합도는 CFI $\geq$.950, SRMR $\leq$.080 (Hu & Bentler, 1999), RMSEA $\leq$0.08(Browne & Cudeck, 1993)을 수용기준으로 하였다. 본 연구에서 단일 차원 요인모형의 적합도는 $\chi^2(5)$ = 321.192, p < .001, Robust CFI = .970, SRMR = .031, Robust RMSEA = 0.122로 나타났다. RMSEA 값은 기준을 초과하였지만 나머지 지표에서는 단일 차원 요인모형 적용에 수용 가능한 수준의 모형 적합도를 보였다. 표준화된 요인부하량은 0.70에서 0.91의 범위로 나타났으며, 요인점수의 개별 측정영역 가중치는 0.09에서 0.30의 범위로 산출되었다.

그림 4는 가중 합산점수인 요인점수의 분포 특성을 전체 표본 및 연령집단별로 제시하고 있다. 전반적으로 요인점수의 분포 특성은 그림 3에 제시된 단순 합산점수(총점)의 분포 특성과 매우 비슷하게 나타났다.

문항 합산점수를 사용하지 않고 특질점수를 추정하기 위해 부분점수모형을 적용한 결과는 다음과 같다. 부분점수모형의 적합도는 $M^2(159)$ = 5812.980, p < .001, CFI = .950, SRMR = .100,

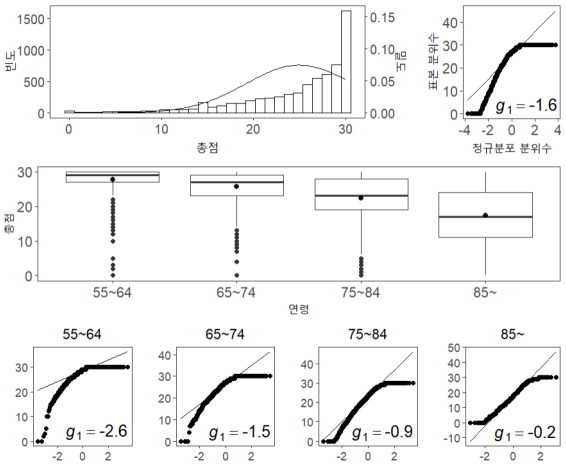

그림 3. 전체 표본 및 연령집단별 K-MMSE 총점의 분포

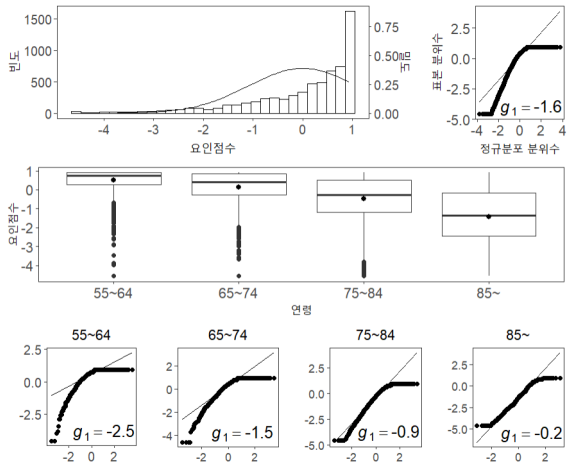

그림 4. 전체 표본 및 연령집단별 K-MMSE 요인점수의 분포

RMSEA = 0.074로 나타나 단일 차원 부분점수모형 적용에 대체적으로 수용 가능한 수준의 모형 적합도를 보였다. 문항모수 분석 결과는 다음과 같다(표 4). 문항의 변별도는 2.2로 높은 수준의 변별도가 추정되었으며 문항 간 서로 동일하게 제약되었다. 각 문항의 전환위치 모수를 추정한 결과 전환위치 모수의 범위는 -3.02에서 0.18로 전반적으로 쉬운 난이도 수준으로 추정되었다. 부분점수모형의 문항 변별도와 전환위치 모수를 바탕으로 표현된 문항특성곡선과 검사특성곡선은 본 논문의 OSF 공개 자료 저장소(https://osf.io/uxz7y)에서 확인할 수 있다(Appendix.pdf 파일의 부록 1). 부분점수모형의 개별 문항 적합도를 파악하기 위해 INFIT과 OUTFIT 분석을 실시하였다. INFIT 및 OUTFIT은 0.50 이상 1.50 이하의 범위에 포함되는 문항을 수용 가능한 수준의 문항 적합도를 지닌 것으로 해

표 4. K-MMSE 부분점수모형의 문항 특성

문항번호	부분점수모형						
	변별도	전환위치 모수				문항 적합도	
		1	2	3	4	INFIT	OUTFIT
1		-2.23	-1.42	-1.23		1.03	0.97
2		-2.02				0.90	1.08
3		-2.96				0.72	0.78
4		-3.02				0.71	0.57
5		-2.57	-1.99	-1.24	-1.03	1.36	1.08
6		-1.90	-1.16	-0.74		1.11	0.83
7		-1.44				0.91	0.66
8		-0.58				0.96	0.71
9		-0.85				0.94	0.67
10	2.20	-0.83				0.90	0.63
11		-0.53				0.96	0.75
12		-1.21	-0.55	0.18		1.02	0.86
13		-2.80				0.81	0.61
14		-2.68				0.86	0.47
15		-2.21				0.87	0.58
16		-1.73	-0.88	-0.35		0.84	0.63
17		-0.66				0.88	0.72
18		-1.53				0.89	0.48
19		-0.96				0.84	0.59

석된다(de Ayala, 2022). 부분점수모형의 개별 문항 INFIT의 범위는 0.71에서 1.36으로 나타나 모든 문항이 적절한 수준의 문항 적합도를 지닌 것을 확인하였다. 개별 문항 OUTFIT의 범위는 0.47에서 1.08로 두 문항을 제외한 모든 문항이 문항 적합도 기준 내에 포함되었다.

일반화부분점수모형을 적용한 결과는 다음과 같다. 모형의 적합도는 $M^2(141) = 5221.916$, $p < .001$, CFI = .955, SRMR = .080, RMSEA = 0.074로 단일 차원 일반화부분점수모형 적용에 적절한 수준의 모형 적합도를 보였다. 문항모수 분석 결과는 다음과 같다(표 5), 문항의 변별도의 범위는 1.31에서 3.14로 전반적으로 넓은 수준의 변별도가 혼재되어 있는 것을 확인하였다. 각 문항의 전환위치 모수를 추정한 결과 전환위치 모수의 범위는 -2.83에서 0.12로 전반적으로 쉬운 난이도 수준

표 5. K-MMSE 일반화부분점수모형의 문항 특성

문항번호	일반화부분점수모형						
	변별도	전환위치 모수				문항 적합도	
		1	2	3	4	INFIT	OUTFIT
1	1.85	-2.25	-1.43	-1.37		0.97	0.89
2	2.42	-1.94				0.92	1.21
3	2.43	-2.82				0.73	0.80
4	2.58	-2.83				0.72	0.59
5	1.31	-2.60	-2.18	-1.27	-1.47	1.02	0.80
6	1.66	-1.95	-1.19	-0.90		0.99	0.75
7	2.79	-1.32				0.96	0.76
8	2.36	-0.57				0.97	0.70
9	2.55	-0.81				0.98	0.67
10	2.78	-0.77				0.96	0.64
11	2.32	-0.52				0.96	0.72
12	1.63	-1.24	-0.58	0.12		0.90	0.78
13	2.56	-2.62				0.82	0.65
14	2.64	-2.48				0.89	0.56
15	2.92	-1.99				0.93	0.77
16	2.42	-1.71	-0.88	-0.32		0.91	0.66
17	2.56	-0.63				0.94	0.75
18	3.11	-1.36				0.97	0.54
19	3.14	-0.86				0.94	0.66

으로 추정되었다. 일반화부분점수모형의 개별 문항 INFIT의 범위는 0.72에서 1.02, OUTFIT의 범위는 0.54에서 1.21로 나타나 모든 문항에서 적절한 수준의 문항 적합도를 보였다. 일반화부분점수모형의 문항특성곡선과 검사특성곡선은 본 논문의 OSF 공개 자료 저장소 내의 Appendix.pdf 파일의 부록 2에서 확인할 수 있다.

문항반응모형은 특질점수 추정 오차를 분포의 형태로 표현할 수 있다. 정보 I(θ)는 특질점수 추정 오차 분산의 역수로 계산되며, 피험자의 특질 수준을 정확하게 추정하는 수준을 나타낸다. K-MMSE 특질 수준의 연속선상에서 수검자를 가장 정확하게 추정하는 지점을 파악하기 위해 검사정보함수를 분석한 결과는 다음과 같다(그림 5). 두 모형의 모두 약 -1 특질 수준을 지닌 수검자에게 가장 정확한 추정을 보이며, 특질 수준 축의 양쪽으로 멀어질수록 추정치의 정확성이 가파르게 낮아지는 것을 확인하였다. 정보함수는 문항 모수에 의해 형태가 변화하므로, 변별도를 동일하도록 제약한 부분점수모형의 검사정보함수와 개별 문항 변별도를 추정한 일반화부분점수모형의 검사정보함수는 분포의 양극단에서 차이를 보이나 전반적으로 유사한 형태를 나타내었다.

비가중 및 가중 특질점수의 분포특성을 전체 표본과 연령집단별로 그림 6, 7에 제시하였다. 상단의 왼쪽에 제시된 특질점수의 히스토그램에서 천정효과로 인해 많은 수의 만점자가 확인되었으나, 전반적인 분포의 형태는 실선으로 표현된 추정정규밀도곡선과 좀 더 가까운 형태를 보였다. 상단 오른쪽에 제시한 Q-Q 도표를 통해 왼쪽의 히스토그램과 같이 전반적으로 정규분포에 좀 더 가까운 형태를 띠는 것을 확인하였다.

가운데 단은 연령집단별 특질점수 상자도표를 제시하였으며, 연령집단의 증가에 따라 특질점수가 감소하는 양상을 확인하였다. 하단에는 연령집단별 Q-Q 도표를 제시하였고, 낮은 연령집단에서 가장 정규분포에서 벗어나는 형태를 보이고, 연령집단의 증가에 따라 특질점수 분포의 비대칭성이 감소하는 양상을 확인하였다. 추가로 부분점수모형과 일반화부분점수모형 가운데 K-MMSE 자료에 더 적합한 모형을 탐색하기 위해 우도비 검정을 실시하였다. 분석 결과, 본 자료는 일반화부분점수모형과 더

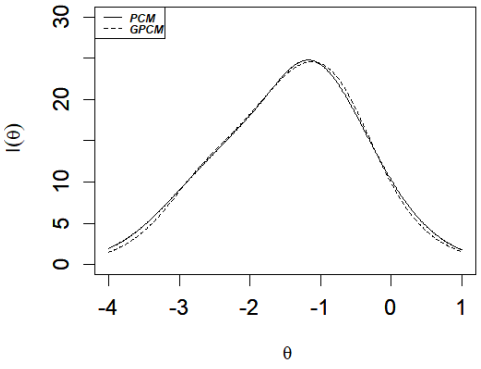

그림 5. K-MMSE 부분점수모형 및 일반화부분점수모형의 검사정보함수

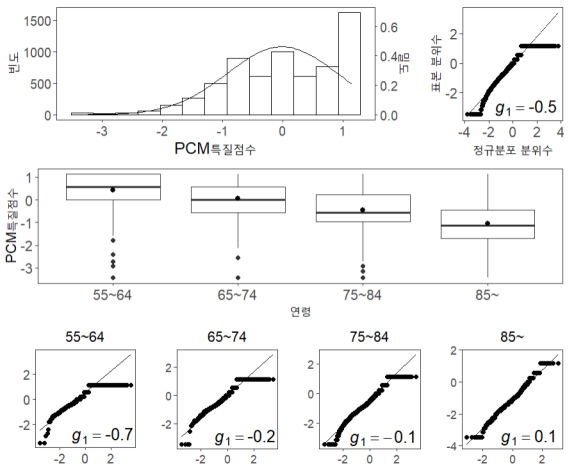

그림 6. 전체 표본 및 연령집단별 K-MMSE PCM 특질점수의 분포

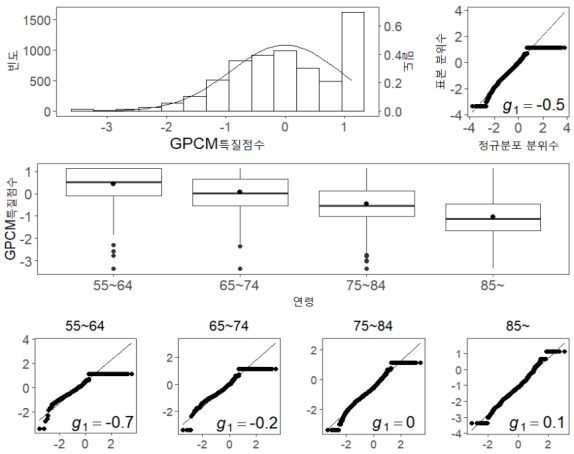

그림 7. 전체 표본 및 연령집단별 K-MMSE GPCM 특질점수의 분포

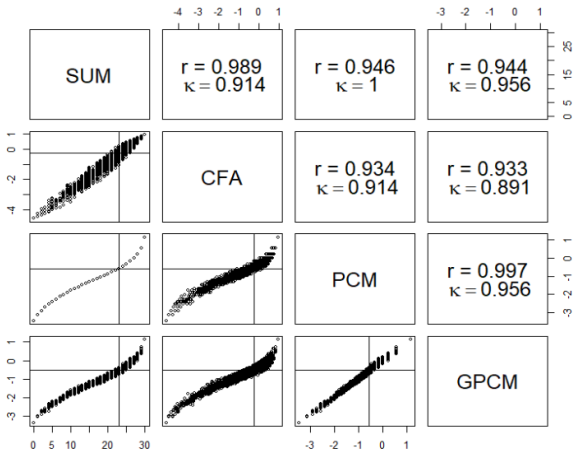

그림 8. 전체 표본의 K-MMSE 검사점수 간 피어슨 상관계수 및 카파 계수

잘 부합하는 것으로 나타났다, $\Delta\chi^2 = 850.632$, $df = 18$, $p < .001$.

검사점수의 비교

전체 표본에서 서로 다른 방식으로 산출된 검사점수들 사이의 산점도를 그림 8 왼쪽 하단에 제시하였다. (가중치 없이 산출된) 동일한 총점에 대하여 가중치를 고려하여 산출된 특질점수(요인점수 및 GPCM 점수)들은 문항반응의 패턴에 따라 서로 다른 값을 가질 수 있다(SUM-CFA, SUM-GPCM 참조). 반면 모든 문항의 가중치가 같다는 제약하에 산출된 검사점수들은 서로의 관계가 비선형적이더라도 일대일 대응을 이루고 있음을 확인하였다(SUM-PCM). 검사점수 산출 방식 사이에 선형적 관계성에 대한 가정이 다른 경우 두 검사점수의 관계는 비선형적인 형태를 보이나, 동일한 선형 관계를 가정하는 경우, 두 검사점수의 산점도는 선형에 가까운 형태를 띠었다(SUM-CFA, PCM-GPCM). 검사점수 산출 방식 간 선형적 관련성을 파악하기 위해 피어슨 상관계수를 산출하여 그림 8 오른쪽 상단에 제시하였다. 선형성에 대한 가정이 일치하는 점수 산출 방식 사이에 가장 높은 수준의 피어슨 상관계수가 나타나 선형 관계 가정의 일치 여부가 검사점수 간 선형적 관련성과 밀접한 관련이 있음을 확인하였다(PCM-GPCM, r = .997; SUM-CFA, r = .989). 가중치 설정 또한 검사점수 간 선형적 관련성에 영향을 미치는 요인으로, 서로 다른 선형 관계를 갖더라도 두 방식 모두 가중치를 설정하는 경우 검사점수 간 선형적 관련성은 가장 낮은 수준으로 나타나는 것을 확인하였다(CFA-GPCM, r = .933). 전체 표본에서 검사점수 간 분류 일치도를 살펴보기 위해 카파 계수를 산출하여 그림 8 오른쪽 상단에 제시하였다. 가중치를 동일하게 제약한 두 방식의 경우 검사점수가 일대일 대응의 형태를 띠므로 선형 관계 가정이 서로 다름에도 불구하고 서로 완전히 일치하는 분류 결과를 나타냈다(SUM-PCM, κ = 1). 가중치를 제약한 두 방식은 서로 동일한 분류 결과를 나타내므로, 두 방식과 다른 방식 간 카파 계수는 서로 동

일한 값을 보였다(SUM-GPCM 및 PCM-GPCM, κ = .956; SUM-CFA 및 PCM-CFA, κ = .914). 가장 낮은 분류 일치도를 보여준 방식은 두 방식 모두 가중치를 설정하고, 서로 다른 선형 관계를 가정한 방식 사이에서 나타났다(CFA-GPCM, κ = .891).

검사점수 분포가 서로 다른 조건에서 검사점수 간 선형적 관련성을 검토하기 위해 연령집단에 따라 검사점수 간 피어슨 상관계수를 산출하여 그림 9에 제시하였다. 각 검사점수 산출 방식을 비교했을 때, 두 방식이 동일한 선형 관계를 가정했을 때 피어슨 상관계수가 모든 연령집단에서 상대적으로 더 높은 수준으로 나타났다(PCM-GPCM, r = .996 ~ .997; SUM-CFA, r = .978 ~ .990). 반면 서로 다른 선형 관계를 가정한 방식에서는 상대적으로 낮은 수준의 피어슨 상관계수를 보였고(SUM-PCM, r = .926 ~ .982; CFA-PCM, r = .897 ~ .973), 가중치를 설정할 경우 그보다 더 낮은 피어슨 상관계수를 보였다(SUM-GPCM, r = .922 ~ .978; CFA-GPCM, r = .893 ~ .972). 또한 연령집단의 상승에 따라 피어슨 상관계수가 전반적으로 상승하는 양상을 확인하였다. 따라서 검사점수 분포의 비대칭성이 완화될수록 각 방식에 따라 산출된 검사점수 분포는 서로 유사하며, 검사점수 분포의 비대칭성이 클수록 각 방식에 따라 산출된 검사점수 분포의 이질성이 전반적으로 증가하는 결과를 확인하였다.

검사점수 분포가 서로 다른 조건에서 검사점수를 이용한 분류 일치도를 검토하기 위해 연령집단에 따라 검사점수 간 카파 계수를 산출하여 그림 10에 제시하였다. 모든 연령집단에서 가중치를 동일하게 제약한 두 방식 간 카파 계수는 1로 나타나(SUM-PCM), 두 방식 모두 가중치를 동일하게 제약한 경우 선형 관계 가정 여부 및 검사점수 분포의 비대칭성은 분류 일치도에 아무런 영향을 미치지 않는다는 것을 확인하였다. 이에 따라 그림 10에서는 가중치를 동일하게 제약한 두 방식과 다른 방식 간 카파 계수는 서로 동일한 값으로 중첩되어 표현되었다(SUM-GPCM 및 PCM-GPCM, κ = .916 ~ .968; SUM-CFA 및 PCM-CFA, κ = .864

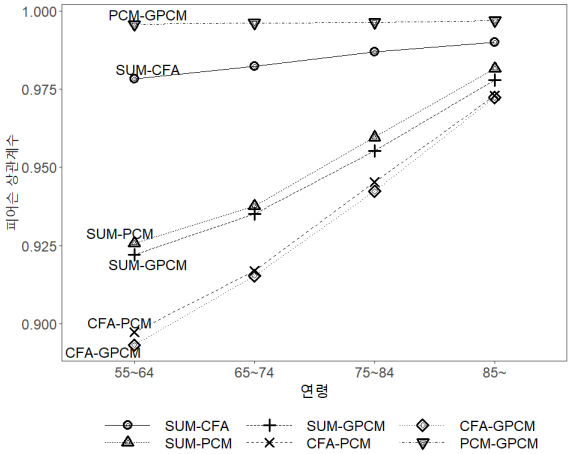

그림 9. 연령집단에 따른 K-MMSE 검사점수 간 피어슨 상관계수

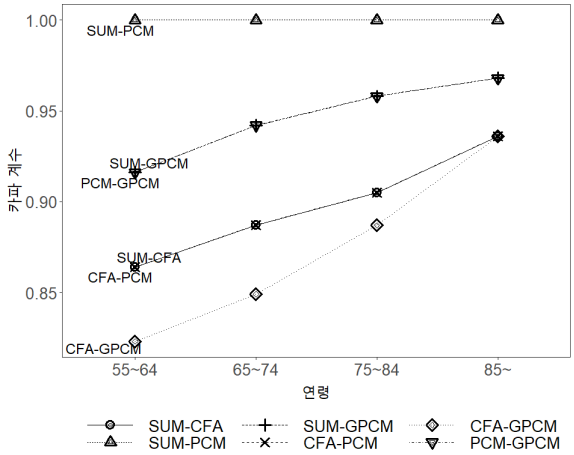

그림 10. 연령집단에 따른 K-MMSE 검사점수별 분류 결과에 대한 카파 계수

~ .936). 모든 연령집단에 걸쳐 가장 낮은 수준의 카파 계수는 가중치 설정과 더불어 서로 다른 선형 관계를 가정한 방식 사이에서 나타나 가중치 설정과 더불어 선형 관계 가정이 서로 다른 경우 가장 낮은 분류 일치도를 나타냈다(CFA-GPCM, κ = .823 ~ .936). 가중치를 동일하게 제약한 두 방식을 제외한 모든 방식 간 연령집단의 상승에 따라 카파 계수가 증가하는 양상이 나타나, 검사점수 분포의 비대칭성 또한 분류 결과 일치도에 영향을 미치는 요인으로 고려된다.

논 의

진단 장면에서 심리검사를 통한 수검자 선별과정은 수검자의 검사점수에 상당 부분 의존하므로 적절한 검사점수 산출 방식의 선택은 검사 개발과정에서 중요하게 다루어져야 할 요소이다. 특히 검사점수를 바탕으로 정상군에 속한 수검자를 위험군으로 진단하거나 위험군에 속한 수검자를 정상군으로 진단하는 것은 수검자 개인과 사회 모두에게 불필요한 시간과 비용의 소모로 이어질 우려가 있다. 국내에서 인지능력을 비롯한 심리적 구성개념을 측정하는 검사들은 대부분이 총점 방식을 채택하여 활용하고 있으나(국립정신건강센터, 2020; 박수빈 등, 2021), 검사 개발과정에서 검사의 목적에 따른 검사점수 산출 방식에 대한 검토와 논의는 충분히 이루어지지 않고 있다.

본 연구는 K-MMSE 검사점수 산출 방식의 적절성을 검토하기 위하여 서로 다른 가정을 지닌 방식으로 산출된 검사점수들 사이의 유사성과 검사점수에 기초하여 수검자를 분류한 결과의 일치도를 확인하였다. 검사점수 산출 방식에 따라 산출된 두 검사점수 간의 유사성에는 선형 관계 가정 여부가 가장 큰 영향을 미치는 요인으로 나타났다. 관계의 선형성에 대한 가정이 다른 점수 산출 방식 사이에서는 유사성이 낮게 나타났으며, 문항 가중 적용 여부가 다를 때에는 그 유사성이 더 낮게 나타났다. 연령집단을 구분하여 분석한 결과는 합산점수 분포의 비대칭성이 증가할수록 합산점수 방식과 특질점수 추정 방식 사이의 유사성이 더욱 낮아진다는 것을 확인하였다.

검사점수 산출 방식에 따라 검사점수를 이용한 분류 일치도는 가중치 제약 여부가 가장 큰 영향을 미치는 요인으로 나타났다. 두 방식 모두 가중치를 동일하게 제약한 경우에는, 선형 관계에 대한 가정이 다른 것은 분류 결과에 아무런 영향을 미치지 않는 결과를 확인하였다. 반면 가중치가 반영된 경우 두 방식 간 분류 결과에 불일치가 발생하였고, 더불어 선형 관계에 대한 가정이 다른 경우 분류 결과의 불일치가 크게 나타났다. 연령집단을 구분하여 분석한 결과 가중치를 동일하게 제약한 두 방식을 제외한 모든 방식 사이에서 검사점수 분포의 비대칭성이 증가할수록 두 방식 간 분류 결과의 불일치가 증가하는 결과를 확인하였다.

검사점수를 이용하여 수검자 간 차이를 명확하게 드러내기 위해서는 검사의 목적에 따라 검사의 특성에 부합하는 검사점수 산출 방식의 검토가 우선적으로 이루져야 한다. 수검자 간 상대적 특질 수준의 차이를 정확하게 파악하기 위해서는 특질 수준과 원점수 간 선형 관계를 주요하게 검토해야 하며, 수검자 분류를 목적으로 하는 경우 문항 가중치 적용 여부를 중요하게 검토해야 한다.

본 연구의 결과는 K-MMSE에 대안적인 검사점수 산출 방식의 적용이 필요하다는 것을 시사한다. 원점수 분포에서 높은 수준의 비대칭성으로 인해

특질 수준과 원점수 간 비선형 관계를 고려할 필요가 있다. 인지기능 저하 조기 선별의 중요성을 고려했을 때, 상대적으로 젊은 연령집단에서 검사점수 분포의 비대칭성이 더욱 크게 나타나므로 수검자의 상대적 인지능력 수준 파악을 위해 비선형 관계의 필요성은 더욱 증가한다. 따라서 본 연구의 문항 분석 결과 K-MMSE의 검사 특성에 부합하는 모형은 부분점수모형 또는 일반화부분점수모형으로 판단된다. 그러나 선별검사로서 K-MMSE의 목적을 고려했을 때, 일반화부분점수모형은 모형 설정에 요구되는 복잡성에 비해 상대적으로 단순한 모형인 부분점수모형과 분류 일치도에서 큰 차이를 보이지 않는다. 단순한 모형에 비해 복잡한 모형에서 제공하는 정보의 차이가 크지 않다면, 모형의 간명성을 고려하여 상대적으로 단순한 모형인 부분점수모형을 적용하는 것이 실용적인 맥락에서 적절한 선택으로 보인다.

검사 제작자를 비롯하여 전통적인 검사점수 산출 방식에 익숙한 연구자들은 복잡한 수학적 구조를 갖는 대안적인 검사점수 산출 방식의 실질적 이점에 대해 회의적인 경우가 많다. 예를 들어 수검자를 선별하기 위한 목적으로 K-MMSE 검사점수를 사용하는 경우, 총점과 부분점수모형 특질점수 방식은 일대일 대응 형태로 인해 분류 결과에 아무런 차이가 나타나지 않는다. 따라서 이 맥락에서는 부분점수모형 방식과 총점 방식 중 어느 방식을 사용하여도 무방하다. 그러나 K-MMSE를 이용하여 수검자의 상대적 능력 수준을 추정하고자 하는 경우에는 수검자의 검사점수 차이를 특질 수준의 차이로 적절하게 해석하기 위하여 비선형 모형 적용을 고려해야 한다. 다만 비선형 모형의 복잡성으로 인한 어려움도 함께 고려해야 한다. 예를 들어 검사특성곡선을 이용하여 수검자가 획득할 수 있는 모든 개별 총점에 대응하는 특질점수를 사전적으로 구성할 수 있다. 총점과 부분점수모형 특질점수는 일대일 대응하는 성질을 띠므로, 대응표를 통해 수검자가 획득한 총점과 특질점수를 일대일 대응 형식으로 간단하게 표현할 수 있다. 검사자는 대응표를 통해 수학적 복잡성에 대한 어려움 없이 검사점수의 사용 목적에 부합하는 방식으로 검사점수를 유연하게 변환하여 사용할 수 있다.

수검자의 검사점수가 갖는 의미를 정확하게 해석하기 위해 검사 제작 과정에서 충분한 표본 크기의 중요성은 지속적으로 강조되어왔다(장승민, 강연욱, 2012; Bridges & Holler, 2007). 검사의 특성을 적절히 반영한 검사점수 산출 방식을 적용하기 위해서는 검사점수 산출 방식이 요구하는 충분한 수의 표본 크기가 필요하다. 확인적 요인분석의 경우 추정 모수의 최소 10배 이상의 표본 크기를 요구하며(Kline, 2023), 부분점수모형의 경우 최소 250명, 일반화부분점수모형의 경우 최소 500명에서 1,200명 수준의 표본 크기를 요구하기도 한다(de Ayala, 2022). 충분한 수의 표본 크기를 확보하지 못할 경우 표집 변동(sampling variability)에 의한 문제가 수반될 우려가 있다. 표집 변동은 무선 표집 과정에 따른 추정치 변화의 범위를 의미하며 표본 크기가 작을수록 표집 변동은 크다. 표집 변동을 고려했을 때, 심리학 일반에서 주로 사용하는 500명 이내 표본 크기 수준에서 획득할 수 있는 가중치를 전적으로 신뢰하기 어렵다. 검사 개발과정에서 충분한 수의 표본 크기를 확보하지 못했다면, 표집 변동의 문제를 안고 불안정한 모형을 추정하는 것보다는 가중치를 동일하게 제약하는 등 강력한 가정을 지닌 검사점수 산출 방식을 통해 안정적인 검사점수를 산출하는 것을 권장한다.

본 연구의 결과는 KLoSA의 2018년도 MMSE

자료에 기초한 것으로 앞서 논의된 내용을 일반화하기 위해서는 추가적인 후속 연구가 필요하다. 모의실험 등을 이용해 검사점수 산출에 영향을 미치는 표본 크기, 문항 수, 응답 범주 수 등 다양한 조건 및 검사점수 산출 방식에 따라 산출되는 검사점수의 선형적 관련성과 검사점수를 이용한 분류일치도를 살펴볼 필요가 있다.

본 연구는 검사점수 산출 방식 선택이 검사 개발자의 임의적 선택이 아닌 검사의 특성과 목적을 고려한 의사결정 과정이라는 것을 강조한다. 현실적인 맥락에서 검사 특성 분석과 검사점수 산출에 필요한 시간 및 인적자원 및 물적자원을 고려했을 때, 모든 진단 장면에서 검사의 특성을 선명히 반영한 검사점수 산출 방식을 적용하는 것은 현실적인 어려움이 따를 수 있다. 그러나 치매 선별과 같이 검사 결과가 개인과 사회에 미치는 영향이 중대할수록 검사점수 산출 방식의 선택은 가능한 정밀하게 검토되어야 한다. 검사의 특성, 진단 장면의 현실적 여건 등을 고려하여 최선의 방법으로 산출된 검사점수가 연구의 기초적인 재료이자 의사결정의 준거로써 보다 신뢰할 수 있는 측정치로 기능할 것을 기대한다.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

참고문헌

강연욱, 나덕렬, 한승혜 (1997). 치매환자들을 대상으로 한 K-MMSE의 타당도 연구. **대한신경과학회지, 15,** 300-307.

강연욱, 장승민, 김상윤, 대한치매학회 (2020). **한국판 간이정신상태검사 2판 사용자 지침서.** 인싸이트.

국립정신건강센터 (2020). **2020년 정신건강 검진 도구 및 사용에 대한 표준지침.** 국립정신건강센터.

박수빈, 이은호, 유빈, 이지현, 김빛나, 김효정, 박승진 (2021). **근거기반 정신건강 평가도구: 통합본.** 국립정신건강센터.

이지수, 강민지, 최민지, 윤혜원, 이옥진, 조현희, 조현성, 이정례, 서지원, 고임석 (2023). **대한민국 치매현황 2022.** 중앙치매센터.

장승민, 강연욱 (2012). 정규분포가 가정된 심리검사의 규준추정을 위한 모형 기반 접근. **한국심리학회지: 일반, 31,** 923-944. https://doi.org/10.15842/kjcp.2012.31.4.004

Anthony, J. C., LeResche, L., Niaz, U., Von Korff, M. R., & Folstein, M. F. (1982). Limits of the 'Mini-Mental State'as a screening test for dementia and delirium among hospital patients. *Psychological medicine*, *12*(2), 397-408. https://doi.org/10.1017/S0033291700046730

Baldwin, P. (2015). Weighting components of a composite score using naïve expert judgments about their relative importance. *Applied Psychological Measurement, 39* (7), 539-550. https://doi.org/10.1177/0146621615584703

Bayley, N., & Aylward, G. P. (2019). *Bayley Scales of Infant and Toddler Development*

fourth edition. Bloomington, MN: NCS Pearson.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689-709. https://doi.org/10.1177/1094428106294734

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological methods, 16*(3), 265. https://doi.org/10.1037/a0024448

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* Guilford publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology*, *13*(6), 528-538. https://doi.org/10.1080/09297040701233875

Burt, C. (1950). The influence of differential weighting. *British Journal of Statistical Psychology, 3*(2), 105-125.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/ correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton university press.

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Jama, 269*(18), 2386-2391. http://dx.doi.org/10.1001/jama.1993.03500180078038

De Ayala, R. J. (2022). *The theory and practice of item response theory, Second edition.* Guilford Publications.

Embretson, S. E., & Reise, S. P. (2000). *item response theory for psychologists*. Psychology Press.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Folstein, M. F., Folstein, S. E., White, T., & Messer, M. A. (2010). *MMSE-2: Mini-mental state examination 2nd*

Edition. Psychological Assessment Resources, Inc., Lutz, FL.

Goldman, R., & Fristoe, M. (2015). *GFTA-3: Goldman Fristoe 3 test of articulation*. Pearson.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6* (4), 430-450. https://doi.org/10.1037/1082-989X.6.4.430

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Huppert, F. A., Cabelli, S. T., & Matthews, F. E. (2005). Brief cognitive assessment in a UK population sample - distributional properties and the relationship between the MMSE and an extended mental state examination. *BMC geriatrics*, *5*(1), 1-14. https://doi.org/10.1186/1471-2318-5-7

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(1), 183-189. https://doi.org/10.1111/1467-9884.00122

Kaufman, A. S. & Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement-Third Edition (KTEA-3).* Bloomington, MN: Pearson.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.

Lopez, M. N., Charter, R. A., Mostafavi, B., Nibut, L. P., & Smith, W. E. (2005). Psychometric properties of the folstein mini-mental state examination. *Assessment*, *12*(2), 137-144. https://doi.org/10.1177/1073191105275412

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Mitchell, A. J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of psychiatric research*, *43*(4), 411-431. https://doi.org/10.1016/j.jpsychires.2008.04.014

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

O'connor, D. W., Pollitt, P. A., Hyde, J. B., Fellows, J. L., Miller, N. D., Brook, C. P. B., & Reiss, B. B. (1989). The reliability

and validity of the Mini-Mental State in a British community survey. *Journal of psychiatric research*, *23*(1), 87-96. https://doi.org/10.1016/0022-3956(89)90021-6

Plomin, R. (1999). Genetics and general cognitive ability. *Nature*, *402*(6761), C25-C29. https://doi.org/10.1038/35011520

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Sage Publications, Inc.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland-3: Vineland adaptive behavior scales*. Pearson.

Stanley, J. C. & Wang, M. D. (1968). *Differential weighting: A survey of methods and empirical studies*. New York: College Entrance Examination Board. https://doi.org/10.3102/00346543040005663

Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

Tombaugh, T. N., & McIntyre, N. J. (1992). The mini mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, *40*(9), 922-935. https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*(5), 663-705. https://doi.org/10.3102/00346543040005663

Wechsler, D. (2008). *Wechsler adult intelligence scale-Fourth Edition (WAIS-IV)*. San Antonio, TX: NCS Pearson, 22(498), 1.

NCS Pearson (2020). *Wechsler individual achievement test* (4th ed.). Bloomington.

Wiig, E. H., Semel, E. & Secord, W. A., (2013). *Clinical evaluation of language fundamentals: CELF-5*. Pearson.

Woodcock, R. W. (2011). *Woodcock reading mastery tests: WRMT-III*. Pearson.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). P*reschool language scales, Fifth edition.* Pearson.

The Korean Journal of
Developmental Psychology

한국심리학회지 : 발달
The Korean Journal of Developmental Psychology
2023, Vol. 36, No. 3, 113-136

Correlations and Classification Agreements among K-MMSE Test Scores based on Different Scoring Methods

Hyeonjong Yu[1] Seungmin Jahng[2]

Department of Psychology, Sungkyunkwan University/ Graduate Student[1]
Department of Psychology, Sungkyunkwan University/ Professor[2]

Traditionally, K-MMSE, a psychological assessment tool used for dementia screening, has been used to evaluate cognitive ability based on summative scores. However, aspects such as whether the relationship between cognitive ability and test scores is assumed to be linear or non-linear and whether item-weight is considered or not may lead to different test scores and different classification. Use of the total score as a test score requires linear relationship and item unweighting, but many psychological tests do not meet these assumptions. The current study examined similarity and classification agreements among test scores derived from different scoring methods, by using K-MMSE data sourced from 6,548 middle-aged and older adults. The Pearson correlation coefficients were high between the scores based on the classical test theory with linearity assumption and between the scores based on the item response theory with nonlinearity assumption. The unweighted scores of total and partial credit model were completely consistent in their classification despite the inconsistency in linearity assumption, but the weighted scores from factor analysis and generalized partial credit model, had the lowest classification agreement. We also found that the greater the asymmetry in the distribution of the total score, the lower the similarity of test scores and classification agreement based on different scoring methods. Lastly, it was emphasized that the selection of appropriate scoring methods should be consistent with the objectives of the test.

Keywords : Test Scores, Sum Scores, Factor Model, Partial Credit Model, Generalized Partial Credit Model