HNUE JOURNAL OF SCIENCE
Educational Sciences 2026, Volume 71, Issue 1, pp. 128-138
This paper is available online at https://hnuejs.edu.vn/es
DOI: 10.18173/2354-1075.2026-0013

A REVIEW OF MEASUREMENT PROPERTIES OF INTELLIGENCE ASSESSMENT TOOLS FOR CHILDREN

Dao Minh Duc, Nguyen Thi Huyen Cham, Nguyen Thi Duyen, Nguyen Thi Linh Duong, Ta Thi Tra Giang*, Tran Thanh Ha, Dang Khanh Ha, Nguyen Van Hoc, Quach Thi Phuong Lan and Nguyen Thi Tra My
Faculty of Psychology - Education, Hanoi National University of Education, Hanoi city, Vietnam
*Corresponding author: Ta Thi Tra Giang,
email: stu735614031@hnue.edu.vn

Received October 12, 2025.
Revised November 13, 2025.
Accepted January 29, 2026.

Abstract. The purpose of this study is to systematically review and compare the measurement properties (reliability and validity) and application scope of five commonly used intelligence assessment tools for children: WISC-V, SB-5, RPM, KABC-II, and UNIT-2. Using a systematic literature review combined with a comparative analysis approach, the study synthesizes and objectively evaluates academic evidence related to these five individual intelligence scales. The results indicate that WISC-V, SB-5, and KABC-II are comprehensive tools with high reliability and validity, suitable for clinical diagnosis and specialized educational interventions. In contrast, RPM and UNIT-2 demonstrate advantages in large-scale screening and assessment of specific populations due to their nonverbal nature and cultural fairness. The study concludes that selecting the most appropriate tool should be based on specific assessment objectives, individual child characteristics, and contextual factors, while emphasizing the critical role of professional expertise in interpreting results.

Keywords: assessment, intellectual ability, comparison, reliability, validity.

TỔNG QUAN CÁC ĐẶC TÍNH ĐO LƯỜNG CỦA CÁC CÔNG CỤ ĐÁNH GIÁ TRÍ TUỆ ĐỐI VỚI TRẺ EM

Đào Minh Đức, Nguyễn Thị Huyền Châm, Nguyễn Thị Duyên, Nguyễn Thị Linh Dương, Tạ Thị Trà Giang*, Trần Thanh Hà, Đặng Khánh Hà, Nguyễn Văn Học, Quách Thị Phương Lan và Nguyễn Thị Trà My
Khoa Tâm lí Giáo dục, Trường Đại học Sư phạm Hà Nội, thành phố Hà Nội, Việt Nam
*Tác giả liên hệ: Tạ Thị Trà Giang,
email: stu735614031@hnue.edu.vn

Ngày nhận bài: 15/10/2025.
Ngày sửa bài: 13/11/2025.
Ngày nhận đăng: 29/1/2026.

Tóm tắt. Mục đích của nghiên cứu này là tổng quan và so sánh hệ thống các đặc tính đo lường (độ tin cậy, độ hiệu lực) và phạm vi ứng dụng của năm công cụ đánh giá trí tuệ trẻ em phổ biến: WISC-V, SB-5, RPM, KABC-II và UNIT-2. Bằng phương pháp tổng quan tài liệu theo hướng tiếp cận hệ thống kết hợp phân tích so sánh, nghiên cứu tổng hợp và đánh giá khách quan các bằng chứng học thuật liên quan đến năm thang đo trên. Kết quả cho thấy WISC-V, SB-5 và KABC-II là các công cụ toàn diện, có độ tin cậy và hiệu lực cao, phù hợp cho chẩn đoán lâm sàng và can thiệp giáo dục chuyên sâu. Trong khi đó, RPM và UNIT-2 tỏ ra vượt trội trong sàng lọc quy mô lớn và đánh giá đối tượng đặc thù nhờ ưu thế phi ngôn ngữ và tính công bằng văn hóa. Nghiên cứu kết luận rằng việc lựa chọn công cụ tối ưu phải dựa trên mục tiêu đánh giá cụ thể, đặc điểm cá nhân của trẻ và bối cảnh thực tế, đồng thời nhấn mạnh vai trò quyết định của chuyên gia trong việc diễn giải kết quả.

Từ khóa: đánh giá, năng lực trí tuệ, so sánh, độ tin cậy, độ hiệu lực.

1. Mở đầu

Đánh giá trí tuệ là quá trình sử dụng các công cụ đo lường chuẩn hóa nhằm xác định năng lực nhận thức và tiềm năng phát triển của cá nhân. Trong các lĩnh vực tâm lí học, giáo dục học và tâm lí học lâm sàng, hoạt động này đóng vai trò quan trọng vì cung cấp cái nhìn sâu sắc về năng lực nhận thức và khả năng học tập của trẻ em. Nhiều nghiên cứu đã chỉ ra rằng việc sử dụng các thang đo trí tuệ chuẩn hóa giúp phát hiện sớm năng lực và khó khăn của học sinh, từ đó điều chỉnh chương trình giáo dục phù hợp. Một nghiên cứu cho thấy rằng việc áp dụng các thang đo trí tuệ như Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V) và Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition (WPPSI-IV) có thể giúp nhận diện học sinh có năng khiếu cũng như gặp khó khăn trong học tập, từ đó tạo điều kiện cho việc phát triển các chính sách giáo dục hỗ trợ (Flanagan & Alfonso, 2017) [1].

Trong thực hành tâm lí học lâm sàng, đánh giá trí tuệ cung cấp thông tin quan trọng cho các chuyên gia trong việc theo dõi sự phát triển nhận thức của trẻ. Theo Wechsler (2014), các kết quả từ các bài kiểm tra trí tuệ không chỉ giúp xác định mức độ rối loạn mà còn hỗ trợ trong việc xây dựng các kế hoạch can thiệp cá nhân hóa [2]. Nghiên cứu đã chỉ ra rằng việc theo dõi tiến trình thông qua đánh giá trí tuệ có thể cải thiện hiệu quả của các can thiệp giáo dục, giúp trẻ tiến bộ và phát triển tốt hơn trong môi trường học tập (Fuchs & Fuchs, 2014) [3].

Hơn nữa, việc phát hiện sớm các rối loạn phát triển như chậm phát triển trí tuệ và rối loạn phổ tự kỉ có thể tạo ra tác động tích cực đến sự can thiệp và hỗ trợ cho trẻ. Nghiên cứu của Guralnick (2001) cho thấy rằng can thiệp sớm có thể làm giảm thiểu các tác động tiêu cực đến sự phát triển xã hội và học tập của trẻ, đồng thời cải thiện chất lượng cuộc sống cho trẻ em và gia đình [4]. Việc xây dựng một môi trường học tập và phát triển phù hợp không chỉ giúp trẻ vượt qua những khó khăn mà còn tạo điều kiện cho chúng phát triển các kỹ năng xã hội và nhận thức cần thiết.

Bài viết này sẽ cung cấp một cái nhìn tổng quan về các thang đo trí tuệ phổ biến, tập trung vào khả năng đánh giá trí tuệ logic và giải quyết vấn đề. Thông qua việc phân tích và so sánh đặc tính đo lường của năm công cụ: WISC-V, SB-5, RPM, KABC-II và UNIT-2, bài viết nhấn mạnh tầm quan trọng của việc lựa chọn thang đo phù hợp trong việc phát hiện và phát triển năng lực trí tuệ của trẻ em. Kết quả từ nghiên cứu này không chỉ có giá trị lí thuyết mà còn mang tính ứng dụng cao trong thực tiễn giáo dục, giúp đánh giá chính xác khả năng trí tuệ và tạo tiền đề cho việc phát triển các chương trình học tập phù hợp.

2. Nội dung nghiên cứu

2.1. Cơ sở lí thuyết về trí tuệ

Trí tuệ được định nghĩa là năng lực tinh thần tổng quát bao gồm khả năng lập luận, giải quyết vấn đề, và học hỏi từ kinh nghiệm (Gottfredson, 1997) [5]. Wechsler (1939) [6] cũng coi trí thông minh là năng lực toàn diện giúp cá nhân hành động có mục đích và thích ứng hiệu quả với môi trường. Hai định nghĩa này nhấn mạnh tính đa diện của trí tuệ, làm nền tảng cho các mô hình đánh giá hiện đại.

Mô hình Cattell-Horn-Carroll (CHC) là lí thuyết có ảnh hưởng sâu rộng nhất, tổ chức trí tuệ thành ba tầng: khả năng g tổng thể và các năng lực rộng như lí luận lưu loát (Gf - Fluid Reasoning), hiểu biết - kiến thức (Gc - Crystallized Intelligence), trí nhớ ngắn hạn (Gsm - Short-Term Memory) và tốc độ xử lí (Gs - Processing Speed). Từ đó, các thang đo hiện đại như Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V) và Stanford-Binet Intelligence Scales - Fifth Edition (SB-5) được thiết kế dựa trên mô hình này nhằm tạo ra hồ sơ nhận thức toàn diện.

Trong khi đó, Raven's Progressive Matrices (RPM) phản ánh thuyết trí tuệ chung (yếu tố g - General Intelligence Theory) và tập trung vào lí luận phi ngôn ngữ, còn KABC-II linh hoạt giữa mô hình CHC và thuyết thần kinh - tâm lí của Luria. Việc tổng quan các thang đo WISC-V, SB-5, RPM, KABC-II và UNIT-2 cho phép phân tích cách mỗi công cụ cân bằng giữa tính toàn diện và công bằng văn hóa trong đánh giá trí tuệ.

2.2. Phương pháp nghiên cứu

Nghiên cứu được thực hiện bằng phương pháp tổng quan tài liệu theo cách tiếp cận hệ thống và phân tích so sánh, nhằm tổng hợp và đánh giá khách quan các bằng chứng học thuật liên quan đến năm thang đo trí tuệ cá nhân hóa gồm WISC-V, SB-5, RPM, KABC-II và UNIT-2. Đối tượng nghiên cứu là các công trình học thuật đã công bố về cơ sở lí thuyết, cấu trúc, đặc tính tâm trắc học (độ tin cậy, độ hiệu lực) và giá trị ứng dụng của các thang đo này.

Nguồn tài liệu được thu thập từ các cơ sở dữ liệu uy tín như EBSCO, ScienceDirect, SpringerLink, ProQuest, Google Scholar và các tạp chí chuyên ngành như *Journal of Psychoeducational Assessment, Psychological Assessment, Learning and Individual Differences, Perceptual and Motor Skills, và Journal of Autism and Developmental Disorders*. Các từ khóa tìm kiếm được sử dụng bằng cả tiếng Anh và tiếng Việt, bao gồm: *"intelligence assessment", "intelligence test", "psychometric properties", "reliability", "validity", "standardization", "WISC-V", "SB-5", "RPM", "KABC-II", "UNIT-2"* và *"đánh giá trí tuệ", "thang đo trí tuệ", "độ tin cậy", "độ hiệu lực", "chuẩn hóa công cụ"*. Sau khi thu thập, tài liệu được hệ thống hóa thành ba nhóm nội dung chính gồm: cơ sở lí thuyết, tổng quan từng thang đo, và so sánh đặc tính tâm trắc học cùng giá trị ứng dụng, qua đó xây dựng khung so sánh chi tiết giữa các công cụ đánh giá trí tuệ.

2.3. Giới thiệu về các thang đo có trong nghiên cứu.

2.3.1.Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V)

Wechsler Intelligence Scale for Children (WISC) là một thang đo trí tuệ dành cho trẻ em/thiếu niên từ độ tuổi 6:0 đến 16:11, phát triển bởi David Wechsler. Trắc nghiệm trí thông minh WISC-V được xây dựng dựa trên lí thuyết của Cattel - Horn - Carroll (CHC). Phiên bản (WISC-V; Wechsler, 2014) là phiên bản mới nhất. WISC-V nhắm đo năng lực nhận thức tổng quát (FSIQ) cùng năm (5) miền nhận thức chính: Verbal Comprehension Index (VCI), Visual Spatial Index (VSI), Fluid Reasoning Index (FRI), Working Memory Index (WMI), and the Processing Speed Index (PSI).

2.3.2. Stanford-Binet Intelligence Scales - Fifth Edition (SB-5)

Stanford-Binet Intelligence Scale phiên bản thứ 5 (SB-5) là thang đo trí tuệ cá nhân dành cho người từ 2 đến 89 tuổi. Dựa trên mô hình đo lường kép, SB-5 đánh giá năm lĩnh vực nhận thức chính: Lập luận linh hoạt, kiến thức, lập luận định lượng, xử lí trực quan - không gian và trí nhớ làm việc, với cả thang ngôn ngữ và phi ngôn ngữ. Bài kiểm tra cho ra chỉ số IQ tổng thể, được sử dụng rộng rãi trong giáo dục và lâm sàng để xác định năng lực trí tuệ, phát hiện trẻ năng khiếu hoặc khuyết tật trí tuệ.

2.3.3. Raven's Progressive Matrices (RPM)

Raven's Progressive Matrices (RPM) được John C. Raven phát triển, là một bộ công cụ đánh giá trí tuệ phi ngôn ngữ dành cho cá nhân từ 5 tuổi đến người lớn (thông qua ba phiên bản: CPM, SPM, và APM). RPM được xây dựng dựa trên lí thuyết của Charles Spearman về g factor (Trí tuệ chung), đặc biệt là khả năng suy diễn các mối quan hệ (eduction of relations). RPM được thiết kế như một công cụ công bằng văn hóa (culture-fair), nhằm giảm thiểu ảnh hưởng của ngôn ngữ và kiến thức đã học. Thang đo này chủ yếu đo lường trí tuệ linh hoạt (Fluid Reasoning - Gf), được thể hiện qua các nhiệm vụ giải quyết vấn đề bằng hình ảnh ma trận trừu tượng, yêu cầu người làm bài nhận diện và hoàn thành các mẫu hình logic.

2.3.4. Kaufman Assessment Battery for Children - Second Edition (KABC-II)

Kaufman Assessment Battery for Children (K-ABC) do A.S. và N.L. Kaufman phát triển, là thang đo năng lực nhận thức dành cho trẻ từ 2,5 đến 12,5 tuổi. Dựa trên lí thuyết trí tuệ lỏng và trí tuệ kết tinh, K-ABC được thiết kế nhằm giảm ảnh hưởng của ngôn ngữ và văn hóa, giúp đánh giá toàn diện hơn khả năng tư duy của trẻ. Thang đo gồm các lĩnh vực chính như xử lí tuần tự, xử lí đồng thời, lập kế hoạch, học tập và kiến thức, được thể hiện qua các nhiệm vụ nhận thức và giải quyết vấn đề bằng hình ảnh.

2.3.5. Universal Nonverbal Intelligence Test - Second Edition (UNIT-2)

Universal Nonverbal Intelligence Test - Second Edition (UNIT-2) là một thang đo trí tuệ đa chiều, được thiết kế cho trẻ em và thanh thiếu niên từ 5:0 đến 21:11. Đặc điểm cốt lõi của UNIT-2 là sử dụng phương thức hoàn toàn phi ngôn ngữ (qua cử chỉ và vật thao tác), nhằm loại bỏ ảnh hưởng của rào cản ngôn ngữ và văn hóa trong đánh giá. Thang đo này nhắm đo trí tuệ tổng quát (g factor) cùng ba (3) khả năng nhận thức chính đã được phân tích yếu tố xác nhận: Trí nhớ (Memory), Lí luận (Reasoning), và Định lượng (Quantitative).

2.4. Độ tin cậy (Reliability)

2.4.1. Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V)

Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V) thể hiện độ tin cậy cao qua nhiều nghiên cứu. Hệ số Cronbach's α dao động từ 0.81-0.99 cho các chỉ số chính (Wechsler, 2014) [2], với α trung bình lần lượt là 0.93 cho VCI (Verbal Comprehension Index - Hiểu lời nói), 0.92 cho VSI (Visual Spatial Index - Nhận thức không gian), 0.91 cho WMI (Working Memory Index - Trí nhớ làm việc) (Canivez và cộng sự, 2017), và 0.96 cho PSI (Processing Speed Index - Tốc độ xử lí) (Lecerf & Canivez, 2018) [7], [8]. Hệ số test-retest sau khoảng 25 ngày đạt từ 0.84-0.93 (Wechsler, 2014) [2], trong khi nghiên cứu khác ghi nhận 0.90 cho FSIQ (Full Scale Intelligence Quotient - Chỉ số trí tuệ toàn phần), 0.87 cho VCI và 0.85 cho FRI (Fluid Reasoning Index - Lí luận lưu loát) (Daniel, 2019) [9]. Độ tin cậy giữa các người chấm cho câu trả lời mở đạt 0.95-0.98 (Ryan & Schnakenberg-Ott, 2003) [10].
Các nghiên cứu này được thực hiện trên mẫu chuẩn hóa gồm 2.200 trẻ em Hoa Kỳ, độ tuổi từ 6-16, đại diện cho nhiều vùng miền, giới tính và chủng tộc, phản ánh bối cảnh văn hóa phương Tây.

2.4.2. Stanford-Binet Intelligence Scales - Fifth Edition (SB-5)

Thang đo Stanford-Binet 5 (SB-5) cũng cho thấy các chỉ số tâm trắc học mạnh mẽ. Độ nhất quán nội tại được ghi nhận với hệ số Cronbach's α trung bình từ 0.95-0.98 cho các chỉ số và điểm tổng (Roid, 2003) [11]. Một nghiên cứu về cấu trúc của thang đo cũng xác nhận hệ số α là 0.96 cho điểm tổng (Roid & Barram, 2004) [12]. Về độ ổn định, hệ số test-retest sau 5-8 tuần cho điểm tổng FSIQ (Full Scale Intelligence Quotient - Chỉ số trí tuệ toàn phần) nằm trong khoảng 0.90 đến 0.95 (Roid, 2003) [11]. Một nghiên cứu khác báo cáo hệ số ổn định là 0.91 cho nhóm mẫu giáo và 0.93 cho nhóm người trưởng thành (Sanders và cộng sự 2007) [13]. Độ tin cậy giữa các người chấm được báo cáo ở mức cao, từ 0.90-0.98 cho các câu trả lời mở sau khi được đào tạo (Roid, 2003) [11].

Các dữ liệu này chủ yếu dựa trên mẫu 4.800 người Mỹ trong độ tuổi từ 2-85, phản ánh đa dạng về độ tuổi, trình độ học vấn và bối cảnh văn hóa phương Tây (Hoa Kỳ).

2.4.3. Raven's Progressive Matrices (RPM)

Raven's Progressive Matrices (RPM) là thang đo trí tuệ phi ngôn ngữ có độ tin cậy cao nhưng thay đổi theo thời gian. Hệ số Cronbach's α dao động từ 0.86-0.93 cho các phiên bản tiêu chuẩn (Raven và cộng sự, 2000) [14], với α = 0.82 trên mẫu 807 trẻ em Bồ Đào Nha và 0.87-0.85 trong nghiên cứu 2 năm tại Lithuania (Court & Raven, 1995) [15]. Tuy nhiên, độ ổn định test-retest giảm dần theo thời gian: 0.87 (sau 1 tuần), 0.86 (2-3 tuần), 0.80 (2 tháng), 0.75 (3

tháng) và 0.71 (1 năm) (Court & Raven, 1995) [15]; thậm chí giảm xuống 0.50 sau 2 năm, 0.61 sau 4 năm và 0.46 sau 11 năm (Kazlauskaite & Lynn, 2002) [16].

Các nghiên cứu chủ yếu tiến hành tại Châu Âu (Bồ Đào Nha, Lithuania), trên mẫu trẻ em và thanh thiếu niên trong độ tuổi 7-18, phản ánh nền văn hóa phương Tây.

2.4.4. Kaufman Assessment Battery for Children - Second Edition (KABC-II)

Thang đo Kaufman Assessment Battery for Children - Second Edition (KABC-II) thể hiện độ tin cậy tốt trong nhiều bối cảnh văn hóa. Hệ số Cronbach's α trung bình dao động từ 0.85-0.93 cho các thang đo chính (Kaufman & Kaufman, 2004) [17], và đạt α = 0.78 trên mẫu trẻ 7-11 tuổi tại Nam Phi (Mitchell và cộng sự, 2017) [18]. Hệ số test-retest sau trung bình 20 ngày nằm trong khoảng 0.75-0.92 (Kaufman & Kaufman, 2004) [17], với giá trị cụ thể 0.86 cho MPI (Mental Processing Index - Chỉ số xử lí trí tuệ) và 0.83 cho FCI (Fluid-Crystallized Index - Chỉ số lí luận lưu loát - kết tinh) (Reynolds và cộng sự 2007) [19]. Độ tin cậy giữa các người chấm cho các tiểu test có câu trả lời mở như *Riddles* rất cao, đạt 0.96-0.99 (Kaufman & Kaufman, 2004) [17].

Các dữ liệu này được thu thập trên mẫu 2.025 trẻ em Hoa Kỳ (3-18 tuổi) và 230 trẻ em Nam Phi (7-11 tuổi), cho thấy độ tin cậy ổn định trong cả bối cảnh phương Tây và phương Nam (châu Phi).

2.4.5. Universal Nonverbal Intelligence Test - Second Edition (UNIT-2)

Universal Nonverbal Intelligence Test - Second Edition (UNIT-2) là một công cụ đánh giá phi ngôn ngữ với các chỉ số độ tin cậy ấn tượng. Độ nhất quán nội tại của thang đo rất cao, với Cronbach's α trung bình từ 0.85 đến 0.96 cho các chỉ số và điểm tổng (Bracken & McCallum, 2017) [20]. Một nghiên cứu trên mẫu Canada cũng xác nhận độ tin cậy nội tại tốt, với α trung bình là 0.93 cho phiên bản đầy đủ và 0.90 cho phiên bản rút gọn (Richardson, 1995) [21]. Về độ ổn định, hệ số test-retest cho điểm tổng FSIQ (Full Scale Intelligence Quotient - Chỉ số trí tuệ toàn phần) sau khoảng 21 ngày dao động từ 0.84 đến 0.91 (Bracken & McCallum, 2017) [20]. Độ tin cậy giữa các người chấm được báo cáo là rất cao, trên 0.95, nhờ vào hệ thống hướng dẫn chấm điểm bằng cử chỉ rõ ràng và được chuẩn hóa (Bracken & McCallum, 2016; McCallum & Bracken, 2012) [20], [22].

Các nghiên cứu này được tiến hành trên mẫu 1.747 trẻ em Hoa Kỳ (5-21 tuổi) và một mẫu phụ tại Canada (5-17 tuổi), đại diện cho bối cảnh văn hóa phương Tây.

2.4.6 So sánh về độ tin cậy

Bảng 1. So sánh độ tin cậy các thang đo

Thang đo	Độ nhất quán Nội tại (Internal Consistency - Cronbach's α)	Độ ổn định (Test-Retest Reliability - Hệ số r)	Độ tin cậy Chấm điểm (Inter-rater Reliability)
WISC-V	α = 0.81-0.99 (Wechsler, 2014) α = 0.91-0.96 (Canivez và cộng sự, 2017) α = 0.95-0.96 (Lecerf & Canivez, 2018)	r = 0.84-0.93 (Wechsler, 2014) r = 0.85-0.90 (Daniel, 2019)	r = 0.95-0.98 (Ryan & Schnakenberg-Ott, 2003)
SB-5	α = 0.95-0.98 (Roid, 2003) α = 0.96 (Roid & Barram, 2004)	r = 0.90-0.95 (Roid, 2003) r = 0.91-0.93 (Sanders và cộng sự, 2007)	r = 0.90-0.98 (Roid, 2003)
RPM	α = 0.86-0.93 (Raven và cộng sự, 2000) α = 0.82-0.87 (Court & Raven, 1995)	r = 0.87-0.71 (Court & Raven, 1995) r = 0.50 (2 năm)	r > 0.95 (Raven và cộng sự, 2000)

		r = 0.61 (4 năm) r = 0.46 (11 năm) (Kazlauskaite & Lynn, 2002)	
KABC-II	α = 0.85-0.93 (Kaufman & Kaufman, 2004) α = 0.78 (Mitchell và cộng sự, 2017)	r = 0.75-0.92 (Kaufman & Kaufman, 2004) r = 0.83-0.86 (Reynolds và cộng sự, 2007)	r = 0.96-0.99 (Kaufman & Kaufman, 2004)
UNIT-2	α =0.85-0.96 (Bracken & McCallum, 2017) α = 0.90-0.93 (Richardson, 1995)	r = 0.84-0.91 (Bracken & McCallum, 2017) r = 0.87 (Richardson, 1995)	r > 0.95 (Bracken & McCallum, 2017)

Phân tích so sánh:

Về độ nhất quán nội tại, cả năm thang đo đều đạt tiêu chuẩn tốt đến xuất sắc (Cronbach's α > 0.80). Trong đó, SB-5 và WISC-V nổi bật với độ tin cậy cao nhất (α = 0.95-0.99), tiếp theo là UNIT-2 và KABC-II (α = 0.85-0.96), và RPM cũng duy trì được mức độ đáng tin cậy (α = 0.82-0.93) bất chấp sự khác biệt về văn hóa khi sử dụng.

Xét về độ ổn định, WISC-V và SB-5 một lần nữa thể hiện ưu thế với hệ số tương quan test-retest rất cao trong các khoảng thời gian ngắn (r = 0.84-0.95). KABC-II và UNIT-2 cũng cho kết quả ổn định ở mức tốt đến rất tốt (r = 0.75-0.92). Điểm đáng chú ý nhất thuộc về RPM, khi độ ổn định của nó suy giảm rõ rệt theo thời gian, từ r = 0.87 (sau 1 tuần) xuống chỉ còn r = 0.46 (sau 11 năm), phản ánh sự ảnh hưởng mạnhmẽ của các yếu tố phát triển nhận thức và kinh nghiệm học tập lên kết quả đánh giá.

Về độ tin cậy giữa các người chấm, tất cả các thang đo đều có chỉ số rất cao (r > 0.90). Đặc biệt, KABC-II và WISC-V đạt gần mức hoàn hảo (r = 0.95-0.99) đối với các tiểu test có câu trả lời mở, trong khi các bài test phi ngôn ngữ như RPM và UNIT-2 cũng duy trì được độ tin cậy cao nhờ vào tính khách quan trong quy trình chấm điểm.

Kết luận về độ tin cậy:

WISC-V và SB-5 là những công cụ mạnh mẽ toàn diện cho các đánh giá chẩn đoán đòi hỏi độ chính xác cao; KABC-II và UNIT-2 tỏ ra linh hoạt và đáng tin cậy trong các bối cảnh đa văn hóa hoặc với đối tượng có hạn chế về ngôn ngữ; còn RPM, dù có hạn chế về độ ổn định dài hạn, vẫn là một công cụ hữu ích cho sàng lọc và đánh giá ngắn hạn. Việc lựa chọn thang đo nào cần dựa trên mục đích đánh giá cụ thể, đặc điểm của đối tượng và bối cảnh thực tế.

2.5. Độ hiệu lực

2.5.1. Wechsler Intelligence Scale for Children - Fifth Edition (WISC-V)

Thang đo WISC-V thể hiện hiệu lực cấu trúc, hội tụ và dự báo vững chắc. Phân tích yếu tố xác nhận (CFA) cho thấy mô hình 5 nhân tố cốt lõi (hiểu biết lời nói, không gian thị giác, lí luận lưu chất, trí nhớ làm việc, tốc độ xử lí) phù hợp rất tốt với dữ liệu, với CFI = 0.96 và RMSEA = 0.045 (Canivez, McGill & Dombrowski, 2017) [7], đồng thời được xác nhận ổn định xuyên văn hóa trên mẫu trẻ em Trung Quốc (Guo, Aveyard & Dai, 2009) [23]. Về hiệu lực hội tụ và dự báo, điểm FSIQ của WISC-V tương quan mạnh với thành tích học tập trên WIAT-III (r ≈ 0.70) (Wechsler, 2014) [2] và có khả năng dự báo khoảng 40% phương sai điểm học tập sau 2-3 năm (Watkins và cộng sự, 2022) [24]. Nhìn chung, với CFI > 0.95, RMSEA < 0.05, r ≈ 0.70 và R ≈ 0.40, WISC-V đáp ứng đầy đủ các tiêu chuẩn hiệu lực của AERA-APA-NCME (2014) cho một công cụ đánh giá tâm lí chuẩn hóa quốc tế.

2.5.2. Stanford-Binet Intelligence Scales - Fifth Edition (SB-5)

SB-5 có độ hiệu lực vững chắc, đảm bảo thang đo thực sự đo lường cấu trúc trí tuệ g như thiết kế. Độ hiệu lực cấu trúc (Construct Validity) được khẳng định thông qua phân tích nhân tố khẳng định (CFA), (tư duy lưu loát, tư duy định lượng, tư duy thị giác-không gian, trí nhớ làm

việc và kiến thức.) phù hợp tốt với dữ liệu chuẩn hóa (Roid, 2003) [11] . Độ hiệu lực hội tụ (Convergent Validity) được thể hiện qua các tương quan cao giữa điểm FSIQ của SB-5 và các thang đo trí tuệ uy tín khác (như WISC-IV). Ngược lại, độ hiệu lực phân biệt (Discriminant Validity) được thể hiện qua tương quan thấp hơn với các cấu trúc không liên quan (Roid, 2003) [11]. Ngoài ra, chỉ số nhạy cảm với thay đổi (CSI) cũng hỗ trợ cho độ hiệu lực về hậu quả (Consequential Validity) của SB-5, giúp các nhà lâm sàng theo dõi sự thay đổi nhận thức một cách hiệu quả trong quá trình can thiệp (Roid, 2003) [11].

2.5.3. Raven's Progressive Matrices (RPM)

Thang đo Raven's Progressive Matrices (RPM) thể hiện độ hiệu lực cao, được khẳng định qua nhiều khía cạnh. Về hiệu lực cấu trúc, các phân tích nhân tố khẳng định (CFA) cho thấy mô hình đơn nhân tố phù hợp tốt, với RMSEA ở mức chấp nhận được; dù CFI đôi khi thấp hơn lí tưởng, điều này được lí giải bởi tính chất tăng dần về độ khó của các mục câu hỏi (Brouwers, Van de Vijver & Van Hemert, 2009) [25]. Hiệu lực hội tụ được chứng minh qua hệ số tương quan mạnh (r = 0.60-0.75) với các chỉ số trí tuệ linh hoạt khác (Mackintosh & Bennett, 2005) [26]. Hiệu lực phân biệt cũng được xác nhận, khi RPM cho thấy tương quan thấp với Gc nhưng cao với Gf - đúng với mục tiêu đo lường năng lực lí luận phi ngôn ngữ (Raven, 2003) [27]. Ngoài ra, thang đo được chuẩn hóa với Mean = 100, SD = 15, giúp so sánh thuận lợi với các thang IQ khác, dù điểm thô có SD dao động khoảng 6.5-7.5.

2.5.4. Kaufman Assessment Battery for Children - Second Edition (KABC-II)

KABC-II cung cấp bằng chứng hiệu lực toàn diện, đặc biệt là về cấu trúc. Độ hiệu lực cấu trúc (Construct Validity) được xác nhận mạnh mẽ thông qua cả phân tích nhân tố khám phá (EFA) và khẳng định (CFA). Kết quả CFA trong sách hướng dẫn (Kaufman & Kaufman, 2004) cho thấy các chỉ số phù hợp mô hình rất tốt cho cả hai mô hình lí thuyết (Luria và CHC), với CFI >0.90 và RMSEA <0.08 [17]. Thêm vào đó, nghiên cứu CFA độc lập đã tái xác nhận cấu trúc 5 yếu tố theo mô hình CHC, cung cấp bằng chứng khách quan mạnh mẽ (Reynolds và cộng sự, 2007) [19]. Về độ hiệu lực hội tụ (Convergent Validity), chỉ số tổng hợp của KABC-II (MPI/FCI) cho thấy mối tương quan rất mạnh mẽ với các chỉ số Toàn thang của các công cụ thuộc hệ thống Wechsler (ví dụ: WISC-III là 0.89, WISC-IV là 0.88), chứng tỏ KABC-II đang đo lường cùng một khái niệm "trí tuệ chung" cốt lõi, mang lại sự linh hoạt cho nhà lâm sàng trong việc lựa chọn công cụ (Mayes & Calhoun, 2007; Lichtenberger và cộng sự, 2009) [28], [29].

2.5.5. Universal Nonverbal Intelligence Test - Second Edition (UNIT-2)

UNIT-2 cung cấp bằng chứng mạnh mẽ về độ hiệu lực, khẳng định công cụ này đo lường đúng cấu trúc trí tuệ phi ngôn ngữ mà nó được thiết kế. Độ hiệu lực cấu trúc (Construct Validity) được xác nhận thông qua phân tích nhân tố khẳng định (CFA). Kết quả phân tích CFA trên mẫu chuẩn hóa cho thấy các chỉ số phù hợp mô hình đều ở mức tốt, với CFI >0.90 và RMSEA <0.08 (Bracken & McCallum, 2017) [20]. Điều này cho thấy dữ liệu thực tế hoàn toàn ủng hộ mô hình cấu trúc lí thuyết của UNIT-2. Về độ hiệu lực hội tụ (Convergent Validity), chỉ số toàn thang (FSIQ) của UNIT-2 cho thấy mối tương quan mạnh mẽ với các công cụ uy tín khác: tương quan 0.84 với FSIQ của WISC-IV và tương quan 0.81 với chỉ số phi ngôn ngữ của SB-5 (Bracken & McCallum, 2017) [20]. Những mối tương quan cao này là bằng chứng cho thấy UNIT-2 đo lường hiệu quả cùng một cấu trúc "trí tuệ chung" như các công cụ tiêu chuẩn vàng khác.

2.5.6. So sánh về độ hiệu lực

Bảng 2. So sánh về độ hiệu lực các thang đo

Thang đo	Hiệu lực cấu trúc (Construct Validity)	Hiệu lực hội tụ (Convergent Validity)
WISC-V	Mạnh mẽ: Mô hình 5 nhân tố được CFA xác nhận (CFI = 0.96, RMSEA = 0.045). Ổn định	Rất tốt: r ≈ 0.70 với thành tích học tập (WIAT-III).

	xuyên văn hóa.	
SB-5	Vững chắc: Mô hình 5 nhân tố được CFA xác nhận.	Tốt: Tương quan cao với FSIQ của các thang đo khác (ví dụ: WISC-IV).
RPM	Tốt: Mô hình đơn nhân tố (*Gf*) được CFA xác nhận (RMSEA = 0.04-0.06).	Tốt: r = 0.60-0.75 với các chỉ số *Gf*khác.
KABC-II	Rất tốt: CFA xác nhận cấu trúc 5 yếu tố (Luria & CHC) (CFI > 0.90, RMSEA < 0.08).	Rất tốt: Tương quan 0.88-0.89 với FSIQ của WISC-III/WISC-IV.
UNIT-2	Tốt: CFA xác nhận mô hình (CFI>0.90, RMSEA < 0.08).	Tốt: Tương quan 0.84 với WISC-IV và 0.81 với SB-5 (phi ngôn ngữ).

Phân tích so sánh:

Độ hiệu lực cấu trúc (Construct Validity): Trong số các thang đo, WISC-V thể hiện bằng chứng cấu trúc rất tốt, với CFI = 0.96 và RMSEA = 0.045. Các chỉ số này đạt chuẩn cao theo khuyến nghị quốc tế (CFI > 0.95, RMSEA < 0.05) (Hu & Bentler, 1999), chứng minh mô hình 5 nhân tố cốt lõi rất vững chắc. Hơn nữa, việc xác nhận tính ổn định xuyên văn hóa (Guo, Aveyard & Dai, 2009) càng củng cố vị thế dẫn đầu của nó [30], [23]. Ba thang đo SB-5, KABC-II và UNIT-2 đều có kết quả CFA rất tốt (CFI > 0.90, RMSEA < 0.08), cho thấy cấu trúc đa nhân tố của chúng phù hợp với mô hình lí thuyết. Trong đó, KABC-II đặc biệt nổi bật vì mô hình CHC của nó đã được xác nhận lại độc lập (Reynolds và cộng sự, 2007), tăng cường giá trị khái niệm của thang đo [19]. RPM được CFA xác nhận mô hình Đơn nhân tố (*Gf*), chứng tỏ đây là công cụ đo lường thuần khiết trí tuệ linh hoạt, không bị pha trộn bởi các thành phần ngôn ngữ hay kiến thức học được.

Độ hiệu lực hội tụ và phân biệt (Convergent & Discriminant Validity): KABC-II và UNIT-2 có bằng chứng hội tụ rất ấn tượng. KABC-II có tương quan gần như ngang bằng với FSIQ của WISC-IV (0.88-0.89), cho thấy nó là một sự thay thế rất đáng tin cậy cho hệ thống Wechsler. UNIT-2 (phi ngôn ngữ) có tương quan 0.84 với WISC-IV, một con số cực kỳ cao đối với một thang đo loại trừ yếu tố ngôn ngữ. RPM nổi bật về độ hiệu lực phân biệt; mục tiêu của nó là chỉ đo *Gf*, do đó tương quan thấp với Gc là một bằng chứng hiệu lực quan trọng cho tính thuần khiết của nó.

2.6. Bàn luận

Giá trị ứng dụng của các công cụ đánh giá trí tuệ phụ thuộc vào bối cảnh, đối tượng và mục tiêu đánh giá. WISC-V và SB-5 được coi là tiêu chuẩn vàng trong chẩn đoán và định hướng giáo dục đặc biệt nhờ khả năng đánh giá cả năng lực trí tuệ tổng quát (g) và hồ sơ nhận thức cụ thể, với độ tin cậy và tính ổn định cao ở nhiều nghiên cứu quốc tế [1], [12], [7]. fKhi áp dụng tại Việt Nam, cần chuẩn hóa kỹ lưỡng để đảm bảo công bằng văn hóa và tránh sai lệch so với chuẩn phương Tây. KABC-II cũng có ưu thế tương tự, nổi bật ở tính công bằng văn hóa khi đánh giá trẻ em từ môi trường đa văn hóa hoặc gặp rào cản ngôn ngữ.

Các công cụ phi ngôn ngữ như RPM và UNIT-2 phù hợp khi đánh giá trẻ gặp rào cản giao tiếp hoặc thuộc nhóm đặc thù (khiếm thính, tự kỉ), giúp giảm thiểu ảnh hưởng của ngôn ngữ và văn hóa (Bracken và cộng sự, 2017; Abdel-Khalek, 2005; Handargule và cộng sự, 2024) [31], [20], [32]. Ở phạm vi lâm sàng - giáo dục, WISC-V, SB-5, KABC-II, RPM và UNIT-2 đều có thể điều chỉnh linh hoạt, từ sàng lọc ban đầu đến chẩn đoán chuyên sâu, phản ánh xu hướng đánh giá trí tuệ công bằng và chính xác (APA và cộng sự, 2014) [33]. Tuy nhiên, khả năng thay thế lẫn nhau vẫn có giới hạn: WISC-V, SB-5 và KABC-II đánh giá trí tuệ tổng quát, RPM đo tư duy linh hoạt (Gf), UNIT-2 phù hợp với đánh giá phi ngôn ngữ.

Tại Việt Nam, nghiên cứu cung cấp cơ sở khoa học giúp lựa chọn công cụ phù hợp với đặc điểm văn hóa và ngôn ngữ bản địa. KABC-II và UNIT-2 hứa hẹn ứng dụng cao cho trẻ dân tộc thiểu số hoặc có hoàn cảnh đặc biệt. Đồng thời, kết quả nghiên cứu tạo nền tảng cho chuẩn hóa

và kiểm định hiệu lực xuyên văn hóa các công cụ quốc tế, gợi mở hướng phát triển thang đo trí tuệ bản địa, tích hợp công nghệ số, và tối ưu hóa việc đánh giá trong hệ thống giáo dục - y tế, nâng cao hiệu quả chăm sóc tâm lí - giáo dục cho trẻ em.

3. Kết luận

Tóm lại, việc lựa chọn thang đo đánh giá trí tuệ nào không phải là một quyết định duy nhất và tuyệt đối, mà là một quá trình cần dựa trên sự cân nhắc kỹ lưỡng giữa mục đích đánh giá, đặc điểm cụ thể của cá nhân và bối cảnh thực tiễn. Mỗi thang đo từ WISC-V, SB-5 cho đến RPM, KABC-II và UNIT-2 đều sở hữu những ưu thế riêng biệt, định vị chúng trong những vùng ứng dụng tối ưu khác nhau, từ chẩn đoán lâm sàng chuyên sâu, sàng lọc quy mô lớn, cho đến đánh giá trong các bối cảnh đa văn hóa hoặc với các đối tượng đặc thù. Mặc dù tồn tại khả năng thay thế lẫn nhau trong một số phạm vi hẹp dựa trên sự tương đồng về mặt lí thuyết, thì giá trị cuối cùng và độ tin cậy của kết quả đánh giá vẫn luôn phụ thuộc vào việc lựa chọn công cụ phù hợp nhất, được thực hiện bởi nhà chuyên môn có kinh nghiệm, nhằm đảm bảo rằng bức tranh trí tuệ thu được là chính xác, toàn diện và hữu ích nhất cho sự phát triển của cá nhân.

TÀI LIỆU THAM KHẢO

[1] Flanagan DP & Alfonso VC, (2017). *Essentials of WISC-V assessment*. John Wiley & Sons. https://www.wiley.com/en-us/Essentials+of+WISC-V+Assessment-p-9781118980873

[2] Wechsler D, (2014). *Wechsler Intelligence Scale for Children–Fifth Edition: Technical and interpretive manual.* NCS Pearson.

[3] Fuchs D, Fuchs LS & Vaughn S, (2014). What is intensive instruction and why is it important?. *Teaching Exceptional Children*, *46*(4), 13-18.. https://doi.org/10.1177/0040059914522966

[4] Guralnick MJ, (2001). A developmental systems model for early intervention. *Infants & Young Children*, *14*(2), 1-18. https://doi.org/10.1097/00001163-200114020-00004

[5] Gottfredson LS, (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79-132. https://doi.org/10.1016/S0160-2896(97)90014-3

[6] Wechsler D, (1940). The measurement of adult intelligence. *The Journal of Nervous and Mental Disease*, *91*(4), 548. https://doi.org/10.1037/10020-000

[7] Canivez GL, Watkins MW & Dombrowski SC, (2017). Structural validity of the Wechsler Intelligence Scale for Children–Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment*, *29*(4), 458. https://doi.org/10.1037/pas0000358

[8] Lecerf T & Canivez GL, (2018). "Complementary exploratory and confirmatory factor analyses of the French WISC-V: Analyses based on the standardization sample": Correction to Lecerf and Canivez (2018). Psychological Assessment, 30(8), 1009. https://doi.org/10.1037/pas0000638

[9] Daniel MH, (2019). *Equivalence of Q-interactive and paper administrations of cognitive tasks: WISC-V.* Pearson. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/wisc-v/q-interactive-wisc-v.pdf

[10] Ryan JJ & Schnakenberg-Ott SD, (2003). Scoring reliability on the Wechsler adult intelligence scale-(WAIS-III). *Assessment*, *10*(2), 151-159. https://doi.org/10.1177/1073191103010002006

[11] Roid, G. H. (2003). *Stanford–Binet Intelligence Scales, Fifth Edition: Technical manual.* Riverside Publishing.

[12] Roid GH & Barram RA, (2004). *Essentials of Stanford-Binet intelligence scales (SB5) assessment*. John Wiley & Sons.

[13] Mei C, Sharon PE, Finch WH, David ME & Barbara RA, (2014). Joint confirmatory factor analysis of the Woodcock-Johnson Tests of Cognitive Abilities, Third Edition, and the Stanford-Binet Intelligence Scales, Fifth Edition, with a preschool population. *Psychology in the Schools, 51*(1), 32–57. https://doi.org/10.1002/pits.21734

[14] Raven J, Raven JC & Court JH, (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General overview.* Harcourt Assessment

[15] Court JH & Raven J, (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7: Research and references: Summaries of normative, reliability, and validity studies and references to all sections.* Oxford University Press; The Psychological Corporation.

[16] Kazlauskaite V & Lynn R, (2002). Two-year test-retest reliability of the colored progressive matrices. *Perceptual and Motor Skills*, *95*, 354-354. https://doi.org/10.2466/pms.2002.95.2.354

[17] Kaufman AS & Kaufman NL, (2004). *Kaufman Assessment Battery for Children–Second Edition (KABC-II): Technical manual.* AGS Publishing.

[18] Mitchell JM, Tomlinson M, Bland RM, Houle B, Stein A & Rochat TJ, (2018). Confirmatory factor analysis of the Kaufman assessment battery in a sample of primary school-aged children in rural South Africa. *South African Journal of Psychology*, *48*(4), 434-452. https://doi.org/10.1177/0081246317741822

[19] Reynolds CR, (2007). Review of the Kaufman Assessment Battery for Children, Second Edition. In Geisinger KF, Spies R, Plake BS, Carlson JF & Murphy LL, (2007). The seventeenth mental measurements yearbook. *(No Title)*. Buros Institute of Mental Measurements.

[20] Friedlander Moore A, McCallum RS, Bracken BA, (2017). The Universal Nonverbal Intelligence Test: Second Edition. In McCallum RS. (Ed.), (2003). *Handbook of nonverbal assessment* (Vol. 30). New York: Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-3-319-50604-3_7

[21] Richardson E, (1995). *Reliability and validity of the Universal Nonverbal Intelligence Test for children with hearing impairments.* Master's Thesis, Western Kentucky University. https://digitalcommons.wku.edu/theses/921

[22] McCallum RS, (2003). The universal nonverbal intelligence test. In *Handbook of nonverbal assessment* (pp. 87-111). Boston, MA: Springer US.

[23] Guo B, Aveyard P & Dai X, (2009). The Chinese Intelligence Scale for Young Children: Testing factor structure and measurement invariance using the framework of the Wechsler Intelligence Tests. *Educational and Psychological Measurement,* 69(3), 459–474. https://doi.org/10.1177/0013164409332209

[24] Watkins MW, Canivez GL, Dombrowski SC, McGill RJ, Pritchard AE, Holingue CB & Jacobson LA, (2022). Long-term stability of Wechsler Intelligence Scale for Children–Fifth Edition scores in a clinical sample. *Applied Neuropsychology: Child,* 11(3), 422–428. https://doi.org/10.1080/21622965.2021.1875827

[25] Brouwers SA, Van de Vijver FJ & Van Hemert DA, (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences,* 19(3), 330–338. https://doi.org/10.1016/j.lindif.2008.10.006

[26] Mackintosh NJ & Bennett ES, (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence,* 33(6), 663–674. https://doi.org/10.1016/j.intell.2005.03.004

[27] Raven J, (2003). Raven Progressive Matrices. In McCallum, R. S (Ed.), *Handbook of Nonverbal Assessment.* Springer. https://doi.org/10.1007/978-1-4615-0153-4_11

[28] Mayes SD & Calhoun SL, (2006). WISC-IV and WISC-III profiles in children with ADHD. *Journal of Attention Disorders*, *9*(3), 486-493. https://doi.org/10.1177/1087054705283616

[29] Lichtenberger EO & Kaufman AS, (2009). *Essentials of WAIS-IV assessment.* Wiley.

[30] Hu LT & Bentler PM, (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

[31] Friedlander Moore A, McCallum RS & Bracken BA, (2017). The Universal Nonverbal Intelligence Test: Second Edition. In McCallum RS, (Ed.), (2003). *Handbook of nonverbal assessment* (Vol. 30). New York: Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-3-319-50604-3_7

[32] Handargule AS, Taksande A, Meshram R, & Uke P, (2024). Applicability of Various Intelligence Scales Utilised in Paediatric Population: An Overview. *Journal of Clinical & Diagnostic Research*, *18*(8), SE01–SE06. https://doi.org/10.7860/JCDR/2024/67521.19752

[33] Skorupiński PM, (2015). American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Standards for educational and psychological testing. *Kwartalnik Pedagogiczny*, *238*(4), 201-203.