مواصفات وشروط الاختبار المقنن الجيد

تأليف

مدرس الدكتور محمد لوتي

أكتوبر 1, 2025

مقدمة

في عالم يتزايد فيه الاعتماد على البيانات والمعلومات لاتخاذ قرارات حاسمة في مجالات متنوعة مثل التعليم، وعلم النفس، والتوظيف، والصحة النفسية، تبرز أهمية أدوات القياس الدقيقة والموثوقة. وتُعد الاختبارات المقننة (Standardized Tests) حجر الزاوية في هذا السياق، فهي أدوات مصممة بعناية فائقة لقياس سمات أو قدرات أو معارف محددة لدى الأفراد بطريقة موضوعية ومنظمة. لكن، ليست كل الاختبارات المقننة متساوية في جودتها أو فعاليتها. فالاختبار الجيد ليس مجرد مجموعة من الأسئلة، بل هو نتاج عملية تطوير دقيقة ومنهجية تضمن امتلاكه لمجموعة من الخصائص الأساسية التي تجعل نتائجه قابلة للثقة والتفسير والاستخدام الهادف (American Educational Research Association , American Psychological Association , & National Council on Measurement in Education , 2014).

إن فهم مواصفات وشروط الاختبار المقنن الجيد ليس مجرد مطلب أكاديمي للمتخصصين في القياس والتقويم، بل هو ضرورة عملية لكل من يستخدم نتائج هذه الاختبارات أو يتأثر بها، سواء كانوا معلمين، أو أخصائيين نفسيين، أو مديري موارد بشرية، أو حتى الطلاب وأولياء أمورهم. فالحكم على مدى كفاءة طالب، أو تشخيص صعوبة تعلم، أو اختيار موظف مناسب، يعتمد بشكل كبير على جودة الأداة المستخدمة في القياس. استخدام اختبار يفتقر إلى المواصفات المطلوبة قد يؤدي إلى استنتاجات خاطئة وقرارات غير عادلة أو غير فعالة، مما قد يكون له عواقب وخيمة على مسارات الأفراد وحياتهم (Urbina, 2014).

يهدف هذا الفصل إلى استعراض وتحليل المواصفات والشروط الأساسية التي يجب أن تتوافر في الاختبار المقنن لكي يُعتبر "جيدًا" أو "ذا جودة عالية". سنناقش بالتفصيل المفاهيم الجوهرية مثل الصدق (Validity) بأوجهه المتعددة، والثبات (Reliability) ومؤشراته المختلفة، والموضوعية (Objectivity) كشرط أساسي لضمان عدالة القياس. كما سنتطرق إلى أهمية وجود معايير (Norms) واضحة لتفسير الدرجات، بالإضافة إلى مراعاة الجوانب العملية (Practicality) التي تؤثر على قابلية استخدام الاختبار في الواقع. وأخيرًا، سنسلط الضوء على قضية العدالة (Fairness) وضرورة خلو الاختبار من التحيز (Bias) ضد فئات معينة. من خلال هذا النقاش، نسعى لتقديم إطار شامل يمكن الاسترشاد به في تقييم جودة الاختبارات المقننة واختيار الأنسب منها، وكذلك في فهم الأسس التي يقوم عليها تطوير مثل هذه الأدوات الحيوية.

مفهوم التقنين (Standardization): الأساس لعملية المقارنة

قبل الخوض في الخصائص السيكومترية التفصيلية، من الضروري فهم جوهر الاختبار "المقنن". التقنين يشير إلى عملية تطبيق الاختبار وإدارته وتصحيحه وتفسير نتائجه باستخدام إجراءات موحدة ومتسقة لجميع الأفراد الذين يخضعون للاختبار (Cohen & Swerdlik, 2018). هذا التوحيد في الإجراءات هو ما يميز الاختبار المقنن عن الاختبارات غير المقننة (مثل اختبارات الصف التي يعدها المعلمون بشكل فردي). لماذا هذا التوحيد مهم إلى هذا الحد؟ الإجابة تكمن في الهدف الأساسي من وراء استخدام العديد من هذه الاختبارات: إجراء مقارنات عادلة. سواء كنا نقارن أداء فرد بأداء مجموعة مرجعية (المعايير)، أو أداء فرد بنفسه في أوقات مختلفة، أو أداء مجموعات مختلفة من الأفراد، فإن هذه المقارنات لا تكون ذات معنى إلا إذا تم الحصول على الدرجات تحت ظروف متطابقة قدر الإمكان.

تتضمن عملية التقنين عدة جوانب رئيسية:

توحيد تعليمات الاختبار: يجب أن يتلقى جميع المختبرين نفس التعليمات الواضحة والدقيقة حول كيفية الإجابة على الأسئلة، والوقت المخصص، وما إذا كان التخمين مسموحًا به أم لا، وكيفية تسجيل الإجابات. أي اختلاف في التعليمات المقدمة يمكن أن يؤثر على أداء الأفراد بشكل غير متساوٍ، مما يخل بمبدأ المقارنة العادلة. غالبًا ما يتم توفير دليل دقيق لمطبق الاختبار (Test Administrator's Manual) يحدد النص الحرفي للتعليمات التي يجب قراءتها (Kaplan & Saccuzzo, 2017).

توحيد ظروف التطبيق: تشمل هذه الظروف البيئة الفيزيائية للاختبار (مثل الإضاءة، التهوية، مستوى الضوضاء، ترتيب المقاعد) والبيئة النفسية (مثل خلق جو مريح وغير مهدد). كما تشمل الالتزام الصارم بالوقت المحدد لكل قسم من أقسام الاختبار. الهدف هو ضمان أن تكون الظروف التي يؤدي فيها الأفراد الاختبار متشابهة قدر الإمكان، بحيث لا تؤثر عوامل خارجية على أدائهم بشكل متفاوت.

توحيد إجراءات التصحيح: سواء كان التصحيح يتم يدويًا أو آليًا، يجب أن تكون هناك قواعد واضحة ومحددة لكيفية منح الدرجات لكل إجابة. بالنسبة للأسئلة الموضوعية (مثل الاختيار من متعدد أو الصح والخطأ)، يتم استخدام مفتاح تصحيح (Scoring Key) موحد. أما بالنسبة للأسئلة التي تتطلب إجابات إنشائية أو أداءً معينًا (مثل اختبارات المقال أو تقييم المهارات)، فيجب تطوير قواعد تصحيح مفصلة (Scoring Rubrics) وتدريب المصححين عليها لضمان الاتساق والموضوعية في تقدير الدرجات (Weiner & Craighead, 2010).

توحيد تفسير الدرجات: لا تكتمل عملية التقنين دون وجود أساس موحد لتفسير معنى الدرجة التي يحصل عليها الفرد. يتم ذلك عادةً عن طريق مقارنة الدرجة الخام للفرد بأداء مجموعة مرجعية كبيرة وممثلة تُعرف باسم "عينة التقنين" (Norming Sample). هذه المقارنة تسمح بتحويل الدرجة الخام إلى درجة معيارية (مثل الدرجة المئينية أو الدرجة التائية)، مما يوضح موقع الفرد النسبي مقارنة بأقرانه (Urbina, 2014).

إن الالتزام الصارم بإجراءات التقنين هو الشرط المسبق الأساسي لضمان أن الفروق في الدرجات بين الأفراد تعكس بالفعل فروقًا حقيقية في السمة أو القدرة المقاسة، وليس مجرد اختلافات في كيفية تطبيق الاختبار أو تصحيحه. بدون تقنين، يفقد الاختبار أساس المقارنة الموضوعية، وتصبح خصائصه السيكومترية الأخرى (كالصدق والثبات) موضع شك كبير.

الخصائص السيكومترية الأساسية: جوهر جودة الاختبار

تعتبر الخصائص السيكومترية (Psychometric Properties) المعايير الفنية التي يُحكم من خلالها على جودة الاختبار المقنن. أهم هذه الخصائص وأكثرها جوهرية هي الصدق (Validity) والثبات (Reliability). هاتان الخاصيتان تمثلان العمود الفقري لأي أداة قياس جيدة، وعدم توافرهما بالمستوى المطلوب يجعل الاختبار عديم القيمة العلمية والعملية.

الصدق (Validity): هل يقيس الاختبار ما يُفترض أن يقيسه؟

يُعد الصدق أهم خاصية يجب توافرها في أي اختبار نفسي أو تربوي. ببساطة، يشير الصدق إلى الدرجة التي يقيس بها الاختبار بالفعل السمة أو المفهوم أو البناء النظري الذي صُمم لقياسه (AERA et al., 2014; Messick, 1989). من المهم التأكيد على أن الصدق ليس صفة مطلقة للاختبار نفسه (أي لا نقول "هذا اختبار صادق" بشكل عام)، بل هو يتعلق بمدى

صحة وملاءمة الاستدلالات (Inferences) التي نستخلصها من درجات الاختبار (Kane, 2013). بمعنى آخر، هل التفسيرات والاستخدامات المقترحة لنتائج الاختبار مدعومة بأدلة قوية ومنطقية؟

على سبيل المثال، إذا كان لدينا اختبار يهدف لقياس "القلق"، فالصدق هنا يتعلق بمدى دقة استنتاجنا بأن الدرجات المرتفعة على هذا الاختبار تشير فعلاً إلى مستويات عالية من القلق لدى الفرد، وليس شيئًا آخر مثل الاكتئاب أو مجرد صعوبة في فهم الأسئلة.

تاريخيًا، تم تصنيف الصدق إلى أنواع مختلفة (مثل صدق المحتوى، الصدق التلازمي، الصدق التنبؤي، صدق التكوين الفرضي). لكن النظرة الحديثة للصدق، كما تم تبنيها في معايير الاختبارات التربوية والنفسية (AERA et al., 2014)، تعتبر الصدق مفهومًا واحدًا موحدًا (Unitary Concept)، ولكن يمكن جمع أدلة عليه من مصادر متعددة. هذه المصادر المختلفة للأدلة تساعد في بناء حجة قوية تدعم التفسيرات والاستخدامات المقصودة لدرجات الاختبار. تشمل هذه المصادر ما يلي:

الأدلة المستندة إلى محتوى الاختبار (Evidence Based on Test Content):
يركز هذا النوع من الأدلة على مدى تمثيل بنود (فقرات) الاختبار للمجال أو المحتوى الذي يُفترض أن يقيسه الاختبار. هل تغطي الأسئلة جميع الجوانب الهامة للمفهوم أو المهارة المستهدفة بشكل متوازن؟ لجمع هذه الأدلة، يتم عادةً الاعتماد على حكم الخبراء المتخصصين في المجال (Subject Matter Experts). يقوم هؤلاء الخبراء بفحص بنود الاختبار ومقارنتها بتعريف دقيق ومفصل للمجال المراد قياسه (يُعرف أحيانًا بـ "جدول المواصفات" أو Test Blueprint). على سبيل المثال، في اختبار تحصيلي لمادة الرياضيات للصف السادس، يجب أن يتأكد الخبراء من أن الأسئلة تغطي الموضوعات المحددة في المنهج الدراسي لذلك الصف وبالأوزان النسبية المناسبة لكل موضوع (Crocker & Algina, 1986). ضعف تمثيل المحتوى قد يؤدي إلى استنتاجات غير دقيقة حول معرفة الطالب أو فهمه للمادة ككل.

الأدلة المستندة إلى العلاقة بمتغيرات أخرى (Evidence Based on Relations to Other Variables):
يتضمن هذا المصدر فحص العلاقات بين درجات الاختبار ودرجات مقاييس أخرى أو متغيرات خارجية يُتوقع نظريًا أن ترتبط بها أو لا ترتبط. يمكن أن يتخذ ذلك أشكالاً متعددة:

الأدلة التلازمية (Concurrent Evidence): تُجمع هذه الأدلة عن طريق مقارنة درجات الاختبار بدرجات مقياس آخر يقيس نفس السمة أو سمة مرتبطة بها، ويتم تطبيقهما في نفس الوقت تقريبًا. على سبيل المثال، يمكن حساب معامل الارتباط بين درجات اختبار جديد للذكاء ودرجات اختبار ذكاء آخر معروف وموثوق (مثل اختبار وكسلر). وجود ارتباط قوي وإيجابي يدعم صدق الاختبار الجديد (Anastasi & Urbina, 1997).

الأدلة التنبؤية (Predictive Evidence): هنا، يتم فحص قدرة درجات الاختبار على التنبؤ بأداء الفرد في المستقبل على متغير معين يُسمى "المحك" (Criterion). على سبيل المثال، مدى قدرة درجات اختبار الاستعداد الدراسي (مثل SAT أو ACT) على التنبؤ بمعدل الطالب التراكمي (GPA) في السنة الأولى الجامعية. وجود ارتباط دال إحصائيًا بين درجات الاختبار والمحك المستقبلي يعتبر دليلاً على الصدق التنبؤي (Cohen & Swerdlik, 2018). كلما كان الارتباط أقوى، كانت القدرة التنبؤية للاختبار أفضل.

الأدلة التمييزية والتقاربية (Convergent and Discriminant Evidence): تشير الأدلة التقاربية إلى وجود ارتباطات قوية

بين درجات الاختبار ومقاييس أخرى يُفترض نظريًا أنها تقيس نفس البناء أو بناءً مشابهًا. بينما تشير الأدلة التمييزية إلى وجود ارتباطات ضعيفة أو معدومة بين درجات الاختبار ومقاييس أخرى يُفترض نظريًا أنها تقيس أبنية مختلفة أو غير مرتبطة. على سبيل المثال، يُتوقع أن يرتبط اختبار جديد للاكتئاب بشكل قوي باختبارات اكتئاب أخرى (تقاربي)، وبشكل ضعيف باختبارات تقيس القلق الاجتماعي أو الرضا عن الحياة (تمييزي). تحقيق هذا النمط المتوقع من الارتباطات يدعم ما يسمى بـ "صدق التكوين الفرضي" (Campbell & Fiske, 1959) (Construct Validity).

الأدلة المستندة إلى البنية الداخلية للاختبار (Evidence Based on Internal Structure):

يدرس هذا النوع من الأدلة مدى تطابق البنية الداخلية للاختبار (العلاقات بين البنود أو الأقسام الفرعية) مع البناء النظري الذي يُفترض أن يقيسه الاختبار. أحد الأساليب الشائعة هنا هو التحليل العاملي (Factor Analysis)، الذي يهدف إلى تحديد الأبعاد أو العوامل الأساسية التي تكمن وراء استجابات الأفراد على بنود الاختبار. إذا كان الاختبار مصممًا لقياس بناء متعدد الأبعاد (مثل الشخصية بخمسة عوامل كبرى)، فيجب أن تُظهر نتائج التحليل العاملي أن البنود تتجمع فعلاً حول هذه الأبعاد المفترضة نظريًا (Thompson, 2004). الاتساق الداخلي العالي بين بنود يُفترض أنها تقيس نفس البعد يعتبر أيضًا جزءًا من هذه الأدلة.

الأدلة المستندة إلى عمليات الاستجابة (Evidence Based on Response Processes):

تركز هذه الأدلة على فهم العمليات الذهنية أو السلوكية التي يستخدمها الأفراد عند الإجابة على بنود الاختبار. هل العمليات التي يستخدمها المختبرون تتوافق مع ما يفترض الاختبار أنه يقيسه؟ يمكن جمع هذه الأدلة من خلال ملاحظة الأفراد أثناء حل الاختبار، أو سؤالهم عن استراتيجيات تفكيرهم (مثل التفكير بصوت عالٍ "Think-aloud protocols")، أو تحليل أنماط الاستجابة. على سبيل المثال، في اختبار للفهم القرائي، يمكن التحقق مما إذا كان الأفراد يستخدمون بالفعل مهارات الفهم المطلوبة (مثل استنتاج المعنى، تحديد الفكرة الرئيسية) بدلاً من مجرد الاعتماد على الذاكرة السطحية (AERA et al., 2014).

الأدلة المستندة إلى عواقب الاختبار (Evidence Based on Consequences of Testing):

هذا المصدر للأدلة هو الأكثر إثارة للجدل، ويتعلق بالآثار الاجتماعية المقصودة وغير المقصودة لاستخدام الاختبار. هل يؤدي استخدام الاختبار إلى تحقيق الفوائد المرجوة (مثل تحسين التعلم، اختيار أفضل المرشحين)؟ وهل هناك أي عواقب سلبية غير متوقعة (مثل تضييق المنهج الدراسي، زيادة قلق الطلاب، التمييز ضد مجموعات معينة)؟ يرى البعض (مثل Messick, 1989) أن تقييم هذه العواقب جزء لا يتجزأ من عملية التحقق من الصدق، خاصة فيما يتعلق بالقيمة الاجتماعية لاستخدام الاختبار. بينما يرى آخرون أن هذه القضايا تتعلق بسياسات استخدام الاختبار أكثر من كونها تتعلق بصدق القياس نفسه. ومع ذلك، تتفق المعايير (AERA et al., 2014) على أهمية توثيق وتقييم العواقب المحتملة كجزء من عملية التحقق الشاملة.

من المهم إدراك أن عملية جمع أدلة الصدق هي عملية مستمرة وديناميكية، تبدأ مع تصميم الاختبار وتستمر طوال فترة استخدامه. لا يوجد اختبار "صادق" بشكل مطلق أو نهائي، بل تتراكم الأدلة لدعم أو دحض صلاحية استخدامه لأغراض معينة وفي سياقات محددة. كلما كانت الأدلة من مصادر متعددة أقوى وأكثر اتساقًا، زادت ثقتنا في الاستدلالات المستخلصة من درجات الاختبار.

الثبات (Reliability): هل يمكن الاعتماد على نتائج الاختبار؟

إذا كان الصدق يتعلق بما يقيسه الاختبار، فإن الثبات يتعلق بمدى اتساق (Consistency) ودقة القياس. يشير الثبات إلى

الدرجة التي تكون فيها درجات الاختبار خالية من أخطاء القياس العشوائية (Random Measurement Error). بمعنى آخر، إذا أعدنا تطبيق نفس الاختبار على نفس الفرد تحت نفس الظروف (أو استخدمنا صيغة مكافئة منه)، فإلى أي مدى سنحصل على نفس الدرجة أو درجة قريبة منها؟ (Cohen & Swerdlik, 2018). الاختبار الثابت هو الذي يعطي نتائج متسقة ومستقرة عبر مرات التطبيق المختلفة أو عبر أجزائه المختلفة.

من المهم فهم العلاقة بين الثبات والصدق. الثبات شرط ضروري ولكنه غير كافٍ للصدق (Urbina, 2014). يمكن أن يكون الاختبار ثابتًا (يعطي نتائج متسقة) ولكنه غير صادق (لا يقيس ما يُفترض أن يقيسه). تخيل ميزانًا يعطي دائمًا قراءة تزيد 5 كيلوجرامات عن الوزن الفعلي. هذا الميزان ثابت (يعطي نفس القراءة الزائدة باستمرار)، ولكنه غير صادق (لا يقيس الوزن الفعلي بدقة). بالمقابل، لا يمكن أن يكون الاختبار صادقًا إذا لم يكن ثابتًا. إذا كانت درجات الاختبار تتقلب بشكل عشوائي كبير في كل مرة يتم تطبيقه، فلا يمكننا الوثوق بها كقياس دقيق للسمة المستهدفة، وبالتالي لا يمكن أن تكون الاستدلالات المستندة إليها صادقة.

تنشأ أخطاء القياس العشوائية من مصادر متعددة، مثل:

تقلبات في أداء الفرد: الحالة الصحية، المزاج، مستوى القلق، التعب، الحظ في التخمين.

عوامل مرتبطة بتطبيق الاختبار: اختلافات طفيفة في التعليمات، التوقيت، ظروف البيئة الفيزيائية.

عوامل مرتبطة بتصحيح الاختبار: عدم الدقة أو الذاتية في تقدير الدرجات (خاصة في الأسئلة غير الموضوعية).

عوامل مرتبطة بمحتوى الاختبار نفسه: اختيار عينة غير ممثلة من البنود لقياس السمة.

يقاس الثبات عادةً باستخدام معامل الثبات (Reliability Coefficient)، وهو قيمة تتراوح بين 0 و 1. القيمة 1 تمثل ثباتًا تامًا (لا يوجد خطأ قياس)، بينما القيمة 0 تمثل انعدام الثبات (الدرجات مجرد خطأ عشوائي). كلما اقتربت قيمة معامل الثبات من 1، كان الاختبار أكثر ثباتًا وأقل تأثرًا بأخطاء القياس العشوائية. لا يوجد حد فاصل سحري لمستوى الثبات المقبول، حيث يعتمد ذلك على نوع الاختبار والغرض من استخدامه. بشكل عام، تتطلب القرارات الهامة التي تؤثر على الأفراد (مثل التشخيص السريري أو قرارات التوظيف) معاملات ثبات عالية جدًا (غالبًا 0.90 أو أعلى)، بينما قد تكون معاملات الثبات الأقل (مثل 0.70 أو 0.80) مقبولة للاختبارات المستخدمة لأغراض بحثية أو لقرارات أقل حساسية (Kaplan & Saccuzzo, 2017; Nunnally & Bernstein, 1994).

هناك عدة طرق لتقدير معامل الثبات، وكل طريقة تركز على مصادر مختلفة لأخطاء القياس:

طريقة إعادة الاختبار (Test-Retest Reliability):
تتضمن هذه الطريقة تطبيق نفس الاختبار على نفس المجموعة من الأفراد في فترتين زمنيتين مختلفتين (بفاصل زمني مناسب، قد يكون أسابيع أو أشهر قليلة). ثم يتم حساب معامل الارتباط بين درجات الأفراد في المرتين. معامل الارتباط الناتج يعتبر مؤشرًا على "استقرار" الدرجات عبر الزمن. هذه الطريقة حساسة للتغيرات الحقيقية في السمة المقاسة بين التطبيقين (خاصة إذا كانت السمة غير مستقرة بطبيعتها مثل المزاج)، وكذلك لعامل "تذكر" الإجابات من المرة الأولى إذا كان الفاصل الزمني قصيرًا جدًا (Anastasi & Urbina, 1997).

طريقة الصور المتكافئة (Parallel/Alternate Forms Reliability):
لتجنب مشكلة التذكر في طريقة إعادة الاختبار، يتم بناء صيغتين متكافئتين تمامًا من الاختبار (نفس المحتوى، نفس مستوى الصعوبة، نفس عدد البنود، نفس التباين). يتم تطبيق الصيغتين على نفس المجموعة من الأفراد (إما في نفس الجلسة أو بفترة زمنية قصيرة). معامل الارتباط بين درجات الأفراد على الصيغتين يعتبر مؤشرًا على الثبات. هذه الطريقة تقيس كلاً من استقرار الدرجات عبر الزمن (إذا كان هناك فاصل زمني) واتساق الاستجابات عبر عينات مختلفة من البنود (Crocker & Algina, 1986). التحدي الرئيسي هنا هو صعوبة بناء صيغتين متكافئتين بالفعل.

طرق الاتساق الداخلي (Internal Consistency Reliability):
تقيس هذه الطرق مدى اتساق أداء الأفراد عبر البنود المختلفة داخل الاختبار نفسه، وذلك من خلال تطبيق الاختبار مرة واحدة فقط. إنها تفترض أن جميع بنود الاختبار تقيس نفس البناء الأساسي. هناك عدة أساليب شائعة ضمن هذه الفئة:

طريقة التجزئة النصفية (Split-Half Reliability): يتم تقسيم الاختبار إلى نصفين متكافئين قدر الإمكان (عادةً البنود الفردية مقابل الزوجية). تُحسب درجة كل فرد على كل نصف، ثم يُحسب معامل الارتباط بين درجات النصفين. نظرًا لأن هذا الارتباط يعتمد على نصف طول الاختبار فقط، يتم تعديله باستخدام معادلة سبيرمان–براون للتصحيح (Spearman-Brown prophecy formula) لتقدير ثبات الاختبار الكلي (Cohen & Swerdlik, 2018).

معامل ألفا كرونباخ (Cronbach's Alpha - α): يعتبر هذا المعامل مقياسًا لمتوسط جميع معاملات التجزئة النصفية الممكنة للاختبار. يُستخدم على نطاق واسع لتقدير الاتساق الداخلي، خاصة للاختبارات التي تكون فيها الإجابات متدرجة (مثل مقاييس ليكرت). قيمة ألفا تتأثر بعدد بنود الاختبار ومدى الارتباط بينها (Cronbach, 1951).

معادلات كودر–ريتشاردسون (Kuder-Richardson Formulas - KR-20 and KR-21): هي حالات خاصة من معامل ألفا تُستخدم عندما تكون الإجابات على البنود ثنائية (Dichotomous)، أي إما صحيحة أو خاطئة (مثل أسئلة الاختيار من متعدد). KR-20 أكثر دقة ولكن يتطلب معرفة نسبة المجيبين بشكل صحيح على كل بند، بينما KR-21 أسهل حسابيًا ولكنه يفترض أن جميع البنود لها نفس مستوى الصعوبة (وهو افتراض نادرًا ما يتحقق بدقة) (Kuder & Richardson, 1937).

الثبات بين المصححين (Inter-Rater Reliability):
هذه الطريقة ضرورية للاختبارات التي تتضمن حكمًا ذاتيًا في التصحيح، مثل اختبارات المقال، أو تقييمات الأداء، أو تفسير الاختبارات الإسقاطية. يتم تقديرها عن طريق جعل اثنين أو أكثر من المصححين المستقلين يقومون بتصحيح نفس العينة من الإجابات، ثم حساب درجة الاتفاق أو الارتباط بين تقييماتهم. يمكن استخدام معاملات ارتباط بيرسون، أو معامل كابا لكوهين (Cohen's Kappa) للبيانات الفئوية، أو معامل الارتباط داخل الفئة (Intraclass Correlation Coefficient - ICC) (Shrout & Fleiss, 1979). ارتفاع الثبات بين المصححين يشير إلى أن قواعد التصحيح واضحة وموضوعية ويتم تطبيقها باتساق.

يرتبط بمفهوم الثبات مفهوم آخر مهم وهو الخطأ المعياري للقياس (Standard Error of Measurement - SEM). يمثل SEM مقدار الخطأ المتوقع في درجة الفرد الواحد. يمكن تقديره باستخدام معامل الثبات و الانحراف المعياري لدرجات عينة التقنين (SEM = SD * √(1 - Reliability Coefficient)). كلما انخفض معامل الثبات، زاد الخطأ المعياري للقياس. يُستخدم SEM لتكوين "نطاق ثقة" (Confidence Interval) حول الدرجة التي حصل عليها الفرد، مما يعطي فكرة عن المدى الذي يُحتمل أن تقع فيه "الدرجة الحقيقية" للفرد (وهي الدرجة التي كان سيحصل عليها لو كان القياس خاليًا تمامًا من الخطأ) (Harvill, 1991). على سبيل المثال، إذا حصل طالب على درجة 100 في اختبار ذكاء وكان SEM يساوي 5،

فإننا نستطيع أن نقول بدرجة ثقة معينة (عادة 95%) أن درجته الحقيقية تقع بين 90 و 110 (تقريبًا +/- مرتين قيمة SEM). هذا التقدير يساعد في تفسير الدرجات بحذر وعدم التعامل معها كقيم مطلقة ودقيقة تمامًا.

الموضوعية (Objectivity): التحرر من الذاتية في القياس

الموضوعية هي خاصية أساسية أخرى للاختبار المقنن الجيد، وتعني أن تكون عملية تطبيق الاختبار وتصحيحه وتفسير نتائجه مستقلة عن الأحكام الذاتية أو التحيزات الشخصية للقائمين على الاختبار (Urbina, 2014). ترتبط الموضوعية ارتباطًا وثيقًا بعملية التقنين، حيث أن الإجراءات الموحدة هي التي تضمن تحقيق درجة عالية من الموضوعية.

تتجلى الموضوعية في عدة جوانب:

موضوعية التطبيق: تتحقق من خلال الالتزام الصارم بالتعليمات الموحدة وظروف التطبيق المحددة في دليل الاختبار. يجب على مطبق الاختبار أن يتجنب أي إشارات أو تلميحات أو مساعدات قد تؤثر على أداء بعض المختبرين دون غيرهم.

موضوعية التصحيح: هذا هو الجانب الأكثر شيوعًا عند الحديث عن الموضوعية. تتحقق الموضوعية التامة في التصحيح عندما يتوصل أي مصححين مؤهلين يستخدمان نفس مفتاح أو قواعد التصحيح إلى نفس الدرجة بالضبط لنفس الإجابة. هذا سهل التحقيق في الاختبارات ذات البنود الموضوعية (مثل الاختيار من متعدد) التي لها إجابة صحيحة واحدة ومحددة. التحدي الأكبر يكمن في الاختبارات ذات الإجابات المفتوحة أو الإنتاجية (مثل المقالات، حل المشكلات، تقييم الأداء). في هذه الحالات، يجب تطوير قواعد تصحيح مفصلة وواضحة (Rubrics) تتضمن معايير محددة لمنح الدرجات، مع تدريب المصححين على استخدامها بدقة واتساق. قياس الثبات بين المصححين (كما ذكر سابقًا) هو مؤشر كمي على مدى موضوعية عملية التصحيح في هذه الحالات (Weiner & Craighead, 2010).

موضوعية التفسير: حتى لو تم تطبيق الاختبار وتصحيحه بموضوعية، قد تدخل الذاتية في تفسير معنى الدرجات. الاختبار المقنن الجيد يسعى لتقليل هذه الذاتية من خلال توفير معايير (Norms) واضحة ومحددة لمقارنة أداء الفرد بأداء مجموعة مرجعية، مما يعطي معنى أكثر موضوعية للدرجة. سنناقش المعايير بالتفصيل في القسم التالي.

غياب الموضوعية يقوض الثقة في نتائج الاختبار ويفتح الباب أمام التحيز وعدم العدالة. إذا كانت درجة الفرد تعتمد على هوية المصحح أو مزاجه أو انطباعاته الشخصية، فإن الاختبار يفقد قيمته كأداة قياس علمية وموثوقة.

المعايير وتفسير الدرجات (Norms and Score Interpretation): إعطاء معنى للدرجات

الدرجة الخام (Raw Score) التي يحصل عليها الفرد في اختبار مقنن (مثل عدد الإجابات الصحيحة) غالبًا ما تكون قليلة المعنى بمفردها. هل درجة 40 على اختبار للقلق تعتبر مرتفعة أم منخفضة؟ هل إجابة 25 سؤالًا بشكل صحيح في اختبار رياضيات يعني أن الطالب متمكن أم ضعيف؟ للإجابة على هذه الأسئلة، نحتاج إلى إطار مرجعي (Frame of Reference) لمقارنة الدرجة الخام به. توفر "المعايير" (Norms) هذا الإطار المرجعي الأساسي في الاختبارات المقننة التي تهدف إلى المقارنة بين الأفراد (Cohen & Swerdlik, 2018) (Norm-Referenced Tests).

المعايير هي بيانات تصف أداء مجموعة محددة وممثلة من الأفراد، تُعرف بـ "عينة التقنين" (Norming Sample or Standardization Sample)، على الاختبار. يتم اختيار هذه العينة بعناية لتكون ممثلة للمجتمع الأوسع الذي يُفترض أن يُستخدم الاختبار معه (من حيث العمر، الجنس، المستوى التعليمي، الخلفية الثقافية، المنطقة الجغرافية، إلخ). من خلال تحليل أداء عينة التقنين، يمكن تحويل الدرجات الخام إلى "درجات مشتقة" أو "درجات معيارية" (Derived Scores or Standard Scores) تحمل معنى نسبيًا يوضح موقع الفرد مقارنة بأقرانه في عينة التقنين (Urbina, 2014).

أشهر أنواع الدرجات المشتقة المستخدمة كمعايير هي:

الرتب المئينية (Percentile Ranks): تشير الرتبة المئينية لدرجة خام معينة إلى النسبة المئوية للأفراد في عينة التقنين الذين حصلوا على درجة أقل من هذه الدرجة. على سبيل المثال، إذا كانت الرتبة المئينية لدرجة 40 هي 85 (تُكتب P85)، فهذا يعني أن 85% من أفراد عينة التقنين حصلوا على درجة أقل من 40. المئينيات سهلة الفهم والتفسير لغير المتخصصين، لكنها لا تعكس المسافات المتساوية بين الدرجات (فالفرق بين المئين 90 و 99 قد يكون أكبر بكثير من الفرق بين المئين 50 و 59 من حيث الدرجات الخام) (Kaplan & Saccuzzo, 2017).

الدرجات المعيارية (Standard Scores): تعبر هذه الدرجات عن بعد درجة الفرد عن متوسط أداء عينة التقنين، وذلك بوحدات الانحراف المعياري. أشهرها:

الدرجة الزائية (Z-score): هي أبسط أشكال الدرجات المعيارية. تُحسب بطرح متوسط درجات عينة التقنين (M) من الدرجة الخام للفرد (X) وقسمة الناتج على الانحراف المعياري (SD) لعينة التقنين ($Z = (X - M) / SD$). الدرجة الزائية صفر تمثل المتوسط، والدرجات الموجبة فوق المتوسط، والسالبة تحت المتوسط. ميزتها أنها تحافظ على شكل توزيع الدرجات الأصلي وتسمح بمقارنة أداء الفرد على اختبارات مختلفة لها متوسطات وانحرافات معيارية مختلفة (Cohen & Swerdlik, 2018).

الدرجة التائية (T-score): هي تحويل خطي للدرجة الزائية لتجنب القيم السالبة والكسور العشرية. لها متوسط ثابت (عادة 50) وانحراف معياري ثابت (عادة 10). تُحسب بالمعادلة: $T = 10Z + 50$. تُستخدم بكثرة في اختبارات الشخصية والاهتمامات المهنية.

درجات أخرى: هناك تحويلات معيارية أخرى شائعة مثل درجات اختبارات الذكاء (IQ scores) التي غالبًا ما يكون متوسطها 100 وانحرافها المعياري 15 (مثل اختبارات وكسلر وستانفورد-بينيه)، والدرجات الستانينية (Stanines) التي تقسم التوزيع إلى تسع فئات بمتوسط 5 وانحراف معياري 2 تقريبًا.

المعايير العمرية والصفية (Age and Grade Norms): تُستخدم هذه المعايير بشكل خاص في اختبارات القدرات والتحصيل للأطفال والمراهقين. تشير إلى متوسط أداء الأفراد في عمر زمني معين أو صف دراسي معين. على سبيل المثال، إذا كان أداء طفل في اختبار للقراءة يعادل أداء متوسط الأطفال في الصف الرابع، يقال إن لديه "مكافئًا صفيًا" (Grade Equivalent) يساوي 4.0. يجب تفسير هذه المعايير بحذر شديد، لأنها لا تعني أن الطفل يمتلك كل مهارات الصف الرابع، وقد تكون الفروق في الأداء بين الصفوف غير متساوية (Anastasi & Urbina, 1997).

جودة المعايير تعتمد بشكل حاسم على جودة عينة التقنين التي اشتقت منها. يجب أن تتوفر في عينة التقنين عدة شروط:

التمثيل (Representativeness): يجب أن تعكس خصائص المجتمع المستهدف الذي سيطبق عليه الاختبار بدقة.

الحجم (Size): يجب أن تكون كبيرة بما يكفي لضمان استقرار التقديرات الإحصائية (المتوسطات والانحرافات المعيارية).

الملاءمة (Relevance): يجب أن تكون العينة مناسبة للمجموعة التي نقارن الفرد بها (فلا نقارن أداء طفل بمعايير مشتقة من راشدين).

الحداثة (Recency): يجب تحديث المعايير بشكل دوري (كل 10–15 سنة تقريبًا أو أقل)، لأن أداء المجتمعات يتغير بمرور الوقت (وهو ما يُعرف أحيانًا بـ "تأثير فلين" (Flynn effect) في اختبارات الذكاء) (Flynn, 1987).

من المهم أيضًا التمييز بين الاختبارات معيارية المرجع (Norm-Referenced Tests - NRTs) والاختبارات محكية المرجع (Criterion-Referenced Tests - CRTs). الاختبارات معيارية المرجع (التي ناقشناها للتو) تهدف إلى مقارنة أداء الفرد بأداء الآخرين. أما الاختبارات محكية المرجع، فتهدف إلى تقييم مدى إتقان الفرد لمعرفة أو مهارة محددة، بغض النظر عن أداء الآخرين (Popham, 1978). يتم تفسير الدرجات في الاختبارات المحكية بالرجوع إلى "محك" أو مستوى أداء محدد مسبقًا (مثل "إتقان 80% من أهداف التعلم"). مثال عليها اختبارات رخصة القيادة، أو اختبارات إتقان مهارات معينة في برنامج تدريبي. هذه الاختبارات لا تحتاج بالضرورة إلى معايير بالمعنى التقليدي، لكنها تحتاج إلى تحديد دقيق لمستويات الأداء المطلوبة (Cut Scores) بطريقة مبررة ومنطقية (Cizek & Bunch, 2007). يجب أن يكون الاختبار الجيد واضحًا بشأن ما إذا كان مصممًا للتفسير المعياري أم المحكي، وأن يوفر الأدوات اللازمة للتفسير الصحيح.

الاعتبارات العملية (Practicality): قابلية الاستخدام في الواقع

بالإضافة إلى الخصائص السيكومترية الأساسية، يجب أن يأخذ مطورو ومستخدمو الاختبارات المقننة في الاعتبار مجموعة من الجوانب العملية التي تؤثر على سهولة استخدام الاختبار وفعاليته في المواقف الحقيقية. قد يكون الاختبار مثاليًا من الناحية النظرية (صادق وثابت وموضوعي وله معايير جيدة)، ولكنه غير قابل للتطبيق عمليًا إذا لم تراعَ هذه الجوانب (Cohen & Swerdlik, 2018; Kaplan & Saccuzzo, 2017):

سهولة التطبيق (Ease of Administration): هل تعليمات تطبيق الاختبار واضحة وسهلة المتابعة لمطبق الاختبار؟ هل يتطلب تطبيقه تدريبًا متخصصًا ومكثفًا؟ هل يمكن تطبيقه بشكل فردي أم جماعي أم كليهما؟ كلما كان تطبيق الاختبار أبسط وأقل تعقيدًا (مع الحفاظ على التقنين)، كان استخدامه أسهل وأقل عرضة للأخطاء.

الوقت المستغرق (Time Required): كم من الوقت يحتاجه الفرد لإكمال الاختبار؟ هل يتناسب هذا الوقت مع الجدول الزمني المتاح (مثل مدة الحصة الدراسية أو جلسة التقييم)؟ الاختبارات الطويلة جدًا قد تكون مرهقة للمختبرين وتزيد من احتمالية تأثير التعب أو الملل على الأداء. يجب الموازنة بين الحاجة لتغطية المحتوى بشكل كافٍ وبين جعل مدة الاختبار معقولة.

سهولة التصحيح (Ease of Scoring): هل عملية تصحيح الاختبار سريعة ودقيقة؟ هل يمكن تصحيحه آليًا (مثل استخدام الماسح الضوئي للاختبارات الموضوعية) أم يتطلب تصحيحًا يدويًا؟ إذا كان يتطلب تصحيحًا يدويًا، فهل قواعد التصحيح واضحة وسهلة التطبيق؟ هل يحتاج المصححون إلى تدريب مكثف؟ الاختبارات التي يصعب تصحيحها أو تستغرق وقتًا طويلاً قد تكون غير عملية في البيئات التي تتطلب نتائج سريعة أو تقييم أعداد كبيرة من الأفراد.

سهولة التفسير (Ease of Interpretation): هل نتائج الاختبار مقدمة بطريقة واضحة ومفهومة للمستخدمين المقصودين (سواء كانوا متخصصين أو غير متخصصين)؟ هل دليل الاختبار يوفر شرحًا كافيًا لكيفية تفسير الدرجات المختلفة (الخام والمشتقة)؟ هل يقدم معلومات حول الخطأ المعياري للقياس ونطاقات الثقة؟ الاختبار الذي يقدم نتائج غامضة أو صعبة التفسير يفقد الكثير من قيمته العملية.

التكلفة (Cost): ما هي تكلفة شراء مواد الاختبار (كتيبات الأسئلة، أوراق الإجابة، دليل الاختبار، مفاتيح التصحيح)؟ هل هناك رسوم لكل استخدام أو لكل تقرير؟ هل تكلفة تطبيق وتصحيح الاختبار (بما في ذلك وقت الموظفين أو الحاجة لخبراء) معقولة في حدود الميزانية المتاحة؟ يجب موازنة تكلفة الاختبار مع الفوائد المتوقعة من استخدامه.

جودة مواد الاختبار (Quality of Test Materials): هل مواد الاختبار (الكتيبات، الأوراق، الصور، الأدوات إن وجدت) ذات جودة عالية من حيث الطباعة والوضوح والمتانة؟ هل التصميم جذاب وسهل القراءة للمختبرين؟ المواد ذات الجودة الرديئة قد تؤثر سلبًا على تجربة المختبر وتزيد من أخطاء القياس.

التوفر (Availability): هل الاختبار متاح بسهولة للشراء والاستخدام من قبل المؤهلين؟ هل هناك قيود على توزيعه أو استخدامه؟

لا يمكن تحقيق كل هذه الجوانب العملية بشكل مثالي دائمًا، وغالبًا ما يكون هناك مفاضلة بينها وبين الخصائص السيكومترية. على سبيل المثال، قد يكون الاختبار الأكثر صدقًا وثباتًا هو الأكثر تكلفة أو الأصعب تطبيقًا. يجب على مستخدم الاختبار أن يوازن بين هذه العوامل المختلفة عند اختيار الاختبار الأنسب لاحتياجاته وظروفه المحددة.

العدالة وغياب التحيز (Fairness and Absence of Bias): شرط أخلاقي وعلمي

أخيرًا، ولكن ليس آخرًا، يجب أن يكون الاختبار المقنن الجيد عادلًا (Fair) لجميع الأفراد الذين يخضعون له، وخاليًا من التحيز (Bias) ضد أي مجموعة فرعية من المختبرين (AERA et al., 2014). العدالة مفهوم واسع ومعقد، ولكن في سياق الاختبارات، يشير بشكل عام إلى أن الاختبار يجب أن يوفر لجميع الأفراد فرصة متساوية لإظهار ما يعرفونه أو يستطيعون فعله فيما يتعلق بالبناء المقاس، بغض النظر عن خصائصهم الديموغرافية (مثل الجنس، العرق، الخلفية الثقافية، اللغة، الإعاقة) التي لا علاقة لها بهذا البناء (Camilli, 2006).

التحيز في الاختبار (Test Bias) هو مفهوم فني أكثر تحديدًا، ويشير إلى وجود خطأ منهجي (Systematic Error) في درجات الاختبار يؤثر بشكل مختلف على أداء مجموعات فرعية معينة، على الرغم من أن هذه المجموعات قد تكون متساوية في القدرة أو السمة الحقيقية التي يقيسها الاختبار. بعبارة أخرى، التحيز يعني أن الاختبار ليس صادقًا بنفس الدرجة لجميع المجموعات (Cole & Moss, 1989).

يمكن أن ينشأ التحيز من مصادر مختلفة في عملية تطوير الاختبار وتطبيقه:

تحيز المحتوى (Content Bias): قد تتضمن بعض بنود الاختبار معلومات أو مواقف أو مفردات تكون مألوفة لمجموعة ثقافية أو اجتماعية معينة أكثر من غيرها، مما يعطي أفراد تلك المجموعة ميزة غير عادلة. أو قد يعكس المحتوى صورًا نمطية سلبية عن مجموعات معينة.

تحيز الصياغة أو اللغة (Wording/Language Bias): قد تستخدم بعض البنود لغة معقدة أو تراكيب جمل غير مألوفة لغير الناطقين الأصليين باللغة أو لأفراد من خلفيات لغوية معينة، مما يعيق فهمهم للسؤال حتى لو كانوا يعرفون الإجابة.

تحيز النسق أو الشكل (Format Bias): قد يكون نسق الاختبار أو نوع الأسئلة (مثل الاعتماد الكبير على سرعة الأداء أو استخدام أنواع معينة من الرسوم البيانية) أسهل لبعض المجموعات مقارنة بغيرها بناءً على خبراتهم التعليمية أو الثقافية.

تحيز التطبيق (Administration Bias): قد تختلف طريقة تفاعل مطبق الاختبار أو توقعاته تجاه مجموعات مختلفة من المختبرين، مما يؤثر على أدائهم.

تحيز التصحيح (Scoring Bias): قد يتأثر المصححون (بوعي أو بغير وعي) بأحكام مسبقة عند تصحيح الإجابات المفتوحة لأفراد من مجموعات مختلفة.

تحيز المحك (Criterion Bias): إذا كان المحك المستخدم للتحقق من صدق الاختبار (مثل تقييمات الأداء في العمل أو الدرجات الجامعية) متحيزًا هو نفسه ضد مجموعة معينة، فقد يبدو الاختبار متحيزًا بشكل خاطئ (أو قد يخفي تحيزًا حقيقيًا) (Reynolds & Brown, 1984).

يجب على مطوري الاختبارات بذل جهود حثيثة للكشف عن التحيز المحتمل وإزالته أو تقليله إلى أدنى حد ممكن. تشمل هذه الجهود:

مراجعة البنود من قبل خبراء متنوعين: يقوم خبراء من خلفيات ثقافية ولغوية وجنسانية مختلفة بمراجعة بنود الاختبار للكشف عن أي محتوى أو صياغة قد تكون متحيزة.

تحليل الأداء التفاضلي للبنود (Differential Item Functioning - DIF): هو مجموعة من الأساليب الإحصائية التي تقارن أداء مجموعات فرعية مختلفة (مثل الذكور والإناث، أو مجموعات عرقية مختلفة) على كل بند من بنود الاختبار، بعد ضبط الفروق الإجمالية في القدرة بين المجموعات. إذا أظهر بند معين أداءً مختلفًا بشكل كبير بين مجموعتين متساويتين في القدرة الكلية، فقد يكون هذا البند متحيزًا ويحتاج إلى مراجعة أو حذف (Holland & Wainer, 1993; Zumbo, 1999).

دراسات الصدق التفاضلي (Differential Validity Studies): فحص ما إذا كان الاختبار يتنبأ بالمحك بنفس الدقة لجميع المجموعات الفرعية.

توفير تسهيلات معقولة (Accommodations): بالنسبة للأفراد ذوي الإعاقة أو الذين يتعلمون اللغة، قد يكون من الضروري توفير تعديلات في طريقة تطبيق الاختبار (مثل وقت إضافي، قراءة الأسئلة بصوت عالٍ، استخدام لغة الإشارة، طباعة بحروف كبيرة) لضمان أن يقيس الاختبار قدرتهم الحقيقية وليس إعاقتهم أو حاجز اللغة. يجب أن تكون هذه التسهيلات مدعومة بأدلة على أنها لا تغير من طبيعة البناء المقاس (AERA et al., 2014).

ضمان العدالة يتجاوز مجرد غياب التحيز الإحصائي. إنه يتطلب أيضًا التأكد من أن الاختبار يُستخدم بطريقة مناسبة وأخلاقية، وأن التفسيرات والاستنتاجات المستخلصة من نتائجه عادلة لجميع الأفراد. وهذا يشمل توفير معلومات كافية للمختبرين عن الغرض من الاختبار وكيفية استخدام نتائجه، وحماية سرية بياناتهم، والتأكد من أن القرارات المبنية على الاختبار لا تؤدي إلى تمييز غير مبرر.

الخلاصة

إن تطوير واستخدام اختبار مقنن جيد هو عملية معقدة تتطلب توازنًا دقيقًا بين الدقة العلمية والاعتبارات العملية والأخلاقية. كما ناقشنا في هذا الفصل، لا يكفي أن يكون الاختبار مقننًا، بل يجب أن يفي بمجموعة من المواصفات والشروط الأساسية ليحقق الغرض منه بفعالية وعدالة. تشمل هذه الشروط الأساسية:

الصدق (Validity): يجب أن تتوفر أدلة قوية ومتنوعة تدعم صحة التفسيرات والاستخدامات المقصودة لدرجات الاختبار، أي أنه يقيس بالفعل ما يُفترض أن يقيسه.

الثبات (Reliability): يجب أن تكون درجات الاختبار متسقة ومستقرة وخالية نسبيًا من أخطاء القياس العشوائية، مما يسمح بالثقة في النتائج.

الموضوعية (Objectivity): يجب أن تكون إجراءات تطبيق وتصحيح وتفسير الاختبار خالية من الذاتية والتحيزات الشخصية.

المعايير (Norms): يجب أن تتوفر معايير مناسبة وحديثة وممثلة (في حالة الاختبارات معيارية المرجع) تسمح بتفسير معنى الدرجات بشكل واضح ومقارنة أداء الفرد بأداء مجموعة مرجعية مناسبة.

الاعتبارات العملية (Practicality): يجب أن يكون الاختبار قابلاً للاستخدام في الواقع من حيث سهولة التطبيق والتصحيح والتفسير، والوقت المستغرق، والتكلفة، وجودة المواد.

العدالة وغياب التحيز (Fairness and Absence of Bias): يجب أن يوفر الاختبار فرصة متساوية لجميع المختبرين لإظهار قدراتهم، وأن يكون خاليًا من أي تحيز منهجي ضد مجموعات فرعية معينة، وأن يُستخدم بطريقة أخلاقية وعادلة.

هذه الخصائص ليست منفصلة تمامًا، بل هي مترابطة وتؤثر على بعضها البعض. فالاختبار غير الثابت لا يمكن أن يكون صادقًا، والاختبار غير الموضوعي يهدد كلاً من الثبات والصدق والعدالة. يتطلب بناء اختبار جيد تلبية هذه الشروط مجتمعة من خلال عملية تطوير دقيقة ومراجعة مستمرة.

على مستخدمي الاختبارات المقننة، سواء كانوا معلمين، أو أخصائيين، أو باحثين، أو صانعي قرار، تقع مسؤولية كبيرة في فهم هذه المواصفات والشروط. يجب عليهم تقييم جودة الاختبارات التي يستخدمونها بناءً على هذه المعايير، واختيار الأدوات الأكثر ملاءمة لأغراضهم، وتفسير النتائج بحذر ومسؤولية، مع الوعي الدائم بحدود القياس وإمكانية الخطأ. إن الاستخدام المستنير والأخلاقي للاختبارات المقننة الجيدة يمكن أن يساهم بشكل كبير في تحسين القرارات المتعلقة بالأفراد والمجتمع، بينما الاستخدام غير النقدي أو غير المسؤول لأدوات قياس ضعيفة يمكن أن يؤدي إلى أضرار بالغة.

المراجع (References)

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

.Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. .221-256). Praeger

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. https://doi.org/10.1037/h0046016

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide for educational and psychological testing*. .Sage Publications

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and .measurement* (9th ed.). McGraw-Hill Education

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., .pp. 201-219). American Council on Education/Macmillan

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and .Winston

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191. https://doi.org/10.1037/0033-2909.101.2.171

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33-41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

.Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th

.ed.). Cengage Learning

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). .American Council on Education/Macmillan

.Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill

.Popham, W. J. (1978). *Criterion-referenced measurement*. Prentice-Hall

.Reynolds, C. R., & Brown, R. T. (Eds.). (1984). *Perspectives on bias in mental testing*. Plenum Press

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association. https://doi.org/10.1037/10694-000

.Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). John Wiley & Sons

Weiner, I. B., & Craighead, W. E. (Eds.). (2010). *The Corsini encyclopedia of psychology* (4th ed., Vol. (3). John Wiley & Sons. (See entry on Objectivity

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. .Directorate of Human Resources Research and Evaluation, Department of National Defense