بناء الاختبارات التحصيلية المقننة

تأليف

مدرس الدكتور محمد لوتي

أكتوبر 1, 2025

مقدمة

يمثل القياس والتقويم حجر الزاوية في أي نظام تعليمي يسعى إلى تحقيق الجودة والفعالية. فمن خلال أدوات القياس الموثوقة والصادقة، يمكن للمعلمين وصناع القرار التربوي الحصول على بيانات دقيقة حول مدى تحقق الأهداف التعليمية، ومستوى اكتساب الطلاب للمعرفة والمهارات، ونقاط القوة والضعف في المناهج وطرق التدريس. وتعد الاختبارات التحصيلية (Achievement Tests) من أبرز هذه الأدوات وأكثرها شيوعًا واستخدامًا في الميدان التربوي (علام، 2007). فهي تُصمم خصيصًا لقياس ما تعلمه الفرد أو ما اكتسبه من مهارات نتيجة لمروره بخبرة تعليمية معينة أو برنامج دراسي محدد.

ومع ذلك، لا يمكن الاعتماد على أي اختبار تحصيلي للحصول على نتائج دقيقة وقابلة للمقارنة ما لم يكن مقننًا (Standardized). فالاختبار المقنن هو ذلك الاختبار الذي يتم إعداده وتطبيقه وتصحيحه وتفسير نتائجه بطريقة موحدة ومنظمة لجميع الأفراد الذين يطبق عليهم، ويتمتع بخصائص سيكومترية عالية من حيث الصدق والثبات والموضوعية، وله معايير (Norms) مشتقة من أداء مجموعة مرجعية كبيرة وممثلة للمجتمع الذي سيطبق عليه الاختبار (Anastasi & Urbina, 1997). إن عملية التقنين هذه هي ما تمنح الاختبار قوته وتجعله أداة موثوقة لاتخاذ قرارات تربوية هامة، مثل تقييم فعالية البرامج التعليمية، أو تحديد مستوى الطلاب ووضعهم في المسارات المناسبة، أو تشخيص صعوبات التعلم، أو حتى لأغراض المساءلة والمقارنات بين المدارس أو المناطق التعليمية المختلفة.

إن بناء اختبار تحصيلي مقنن ليس بالمهمة السهلة أو العشوائية، بل هو عملية علمية منهجية تتطلب تخطيطًا دقيقًا، وخبرة فنية متخصصة، والتزامًا بمعايير صارمة في كل خطوة من خطواتها. تبدأ هذه العملية بتحديد الغرض من الاختبار والمحتوى الذي سيقيسه، مرورًا بإعداد جدول المواصفات وصياغة الفقرات وتحليلها وتجريبها، وصولًا إلى تحديد صدق الاختبار وثباته ووضع معاييره وإرشادات تطبيقه وتصحيحه. كل خطوة من هذه الخطوات تعتمد على سابقتها وتؤثر في لاحقتها، وأي خلل في أي مرحلة قد يؤثر سلبًا على جودة الاختبار النهائية وموثوقية نتائجه (Downing & Haladyna, 2006).

يهدف هذا الفصل إلى تقديم دليل شامل ومتكامل لعملية بناء الاختبارات التحصيلية المقننة وتطويرها. سنسير خطوة بخطوة عبر المراحل المختلفة لهذه العملية، موضحين الأسس النظرية والمبادئ العملية التي تحكم كل مرحلة. سنناقش بالتفصيل كيفية تحديد أهداف الاختبار ومحتواه، وأهمية بناء جدول المواصفات، وفن صياغة أنواع مختلفة من الفقرات الاختبارية مع مراعاة القواعد العلمية لضمان جودتها. كما سنتطرق إلى أهمية التجريب الاستطلاعي للاختبار وإجراء التحليلات الإحصائية اللازمة للفقرات (تحليل الصعوبة والتمييز وفعالية البدائل) لاتخاذ قرارات مستنيرة بشأنها. وسنولي اهتمامًا خاصًا لمفهومي الصدق (Validity) والثبات (Reliability) باعتبارهما من أهم الخصائص السيكومترية التي يجب توافرها في أي اختبار جيد، وسنستعرض أنواعهما المختلفة وطرق التحقق منهما. وأخيرًا، سنتناول كيفية إعداد الصورة النهائية للاختبار، ووضع تعليمات واضحة للتطبيق والتصحيح، وأهمية بناء المعايير لتفسير الدرجات، مع التأكيد على الاعتبارات الأخلاقية والقانونية التي يجب مراعاتها في جميع مراحل بناء الاختبار واستخدامه. نأمل أن يوفر هذا الفصل للمختصين والباحثين والمعلمين أساسًا متينًا وفهمًا عميقًا لهذه العملية الحيوية في مجال القياس والتقويم التربوي.

1. تحديد الغرض من الاختبار وتحديد المحتوى (Defining the Test Purpose and Content Domain)

تُعد هذه المرحلة نقطة الانطلاق الأساسية في عملية بناء أي اختبار تحصيلي مقنن، بل إن وضوحها ودقتها يمثلان الأساس

الذي تُبنى عليه جميع الخطوات اللاحقة. إن التسرع أو الإهمال في هذه المرحلة قد يؤدي إلى بناء اختبار غير ملائم للغرض الذي أُعد من أجله، أو غير ممثل للمحتوى المراد قياسه، مما يفقده قيمته وفائدته. لذلك، يجب أن تحظى هذه المرحلة بالعناية والتأمل الكافيين.

1.1. أهمية وضوح الهدف (Importance of a Clear Purpose)

قبل البدء في كتابة فقرة واحدة من فقرات الاختبار، يجب أن يكون هناك تصور واضح ودقيق للغرض الأساسي من بناء هذا الاختبار. ما الذي نريد قياسه بالضبط؟ ولماذا نريد قياسه؟ ما هي القرارات التي ستُتخذ بناءً على نتائج هذا الاختبار؟ إن الإجابة عن هذه الأسئلة تحدد مسار عملية البناء بأكملها. على سبيل المثال، هل الهدف من الاختبار هو:

قياس التحصيل النهائي (Summative Assessment): لتقييم مدى إتقان الطلاب للمادة الدراسية في نهاية وحدة أو فصل دراسي أو سنة دراسية، ومنحهم درجات أو شهادات.

التشخيص (Diagnostic Assessment): لتحديد نقاط القوة والضعف لدى الطلاب في مجال معين بهدف تقديم الدعم العلاجي أو الإثرائي المناسب.

التسكين أو التصنيف (Placement Assessment): لوضع الطلاب في المستوى التعليمي أو البرنامج الدراسي الذي يتناسب مع قدراتهم ومعرفتهم الحالية.

تقويم البرامج (Program Evaluation): للحكم على فعالية برنامج تعليمي معين أو منهج دراسي أو طريقة تدريس جديدة.

المساءلة (Accountability): لمقارنة أداء المدارس أو المعلمين أو المناطق التعليمية وتحميلهم مسؤولية نتائج الطلاب.

الاختيار أو القبول (Selection Assessment): لاختيار أفضل المتقدمين للالتحاق بكلية أو وظيفة معينة.

إن تحديد الغرض بدقة يساعد في توجيه القرارات المتعلقة بنوع المحتوى الذي يجب تضمينه، ومستوى صعوبة الفقرات، وأنواع الفقرات المناسبة، وطريقة تفسير الدرجات، وحتى طريقة تطبيق الاختبار. فاختبار يهدف إلى التشخيص سيختلف في طبيعته ومحتواه عن اختبار يهدف إلى الاختيار أو القبول. كما أن وضوح الهدف يضمن أن الاختبار سيقيس بالفعل ما يُفترض أن يقيسه، وهو ما يرتبط ارتباطًا وثيقًا بمفهوم الصدق الذي سنتناوله لاحقًا (Messick, 1989). بدون هدف واضح، يصبح الاختبار مجرد مجموعة من الأسئلة العشوائية التي تفتقر إلى المعنى والغاية.

1.2. تحديد المجتمع المستهدف (Defining the Target Population)

من هو الجمهور الذي سيطبق عليه هذا الاختبار؟ هل هم طلاب مرحلة ابتدائية، أم متوسطة، أم ثانوية، أم جامعية؟ هل هم طلاب في تخصص معين؟ هل هم متعلمون للغة كلغة ثانية؟ إن تحديد المجتمع المستهدف بدقة أمر بالغ الأهمية، لأنه يؤثر على عدة جوانب، منها:

مستوى صعوبة اللغة والمفردات المستخدمة: يجب أن تكون لغة الاختبار وتعليماته مناسبة للمستوى اللغوي والعمر العقلي للمفحوصين.

مستوى تعقيد المحتوى: يجب أن يتناسب المحتوى المعرفي والمهاري الذي يقيسه الاختبار مع ما هو متوقع من أفراد المجتمع المستهدف في تلك المرحلة الدراسية أو العمرية.

اختيار عينات التجريب والتقنين: يجب أن تكون العينات التي يُجرّب عليها الاختبار في مراحله الأولية (التجريب الاستطلاعي) والتي تُبنى عليها المعايير (عينة التقنين) ممثلة للمجتمع المستهدف الذي سيستخدم الاختبار معه في صورته النهائية (AERA, APA, & NCME, 2014).

صياغة الفقرات وشكل الإخراج: قد يتطلب الأمر استخدام صور أو رسوم توضيحية أكثر مع الأطفال الأصغر سنًا، أو مراعاة حجم الخط ونوع الطباعة للفئات العمرية المختلفة.

إن عدم مراعاة خصائص المجتمع المستهدف قد يؤدي إلى بناء اختبار غير عادل أو غير مناسب لهم، مما يؤثر على دقة النتائج وموثوقيتها. على سبيل المثال، استخدام لغة معقدة في اختبار موجه لطلاب المرحلة الابتدائية قد يقيس قدرتهم اللغوية بدلاً من تحصيلهم في المادة الدراسية المقصودة.

1.3. تحليل المحتوى الدراسي أو المهارات المستهدفة (Analyzing the Curriculum Content or Target Skills)

بعد تحديد الغرض والمجتمع المستهدف، تأتي خطوة تحديد وتوصيف المحتوى الذي سيغطيه الاختبار بدقة. يتضمن ذلك تحليلًا شاملاً للمادة الدراسية، أو المنهج، أو قائمة المهارات أو الكفايات التي يُفترض أن يقيسها الاختبار. يُعد هذا التحليل بمثابة "جرد" لكل المعارف والمفاهيم والمبادئ والمهارات التي يتوقع أن يكون الطلاب قد اكتسبوها بنهاية فترة التعلم.

يجب أن يكون هذا التحليل مفصلاً ومنظمًا، بحيث يغطي جميع جوانب المحتوى الهامة. ويمكن الاستعانة بمصادر متعددة في عملية التحليل، مثل:

الكتب المدرسية المقررة.

أدلة المعلمين.

وثائق المنهج الرسمية ومعايير التعلم.

آراء الخبراء في المادة الدراسية والمعلمين ذوي الخبرة.

الأهداف التعليمية المحددة للمادة أو البرنامج.

ينتج عن عملية تحليل المحتوى قائمة مفصلة بالموضوعات الرئيسية والفرعية، والمفاهيم الأساسية، والمهارات العملية أو الذهنية التي يجب أن يشملها الاختبار. يساعد هذا التحليل في ضمان "صدق المحتوى" (Content Validity) للاختبار، أي التأكد من أن فقرات الاختبار تمثل عينة ممثلة وشاملة للمجال الذي يقيسه (Crocker & Algina, 1986). يتطلب هذا التحليل جهدًا ووقتًا، وغالبًا ما يتم بواسطة فريق من المتخصصين في المادة الدراسية وفي القياس التربوي لضمان شموليته ودقته. فكلما كان تحليل المحتوى أكثر تفصيلاً وشمولية، زادت احتمالية أن يكون الاختبار ممثلاً جيدًا لهذا المحتوى.

1.4. تحديد نواتج التعلم المراد قياسها (Identifying Measurable Learning Outcomes)

لا يكفي مجرد تحديد موضوعات المحتوى، بل يجب أيضًا تحديد ما يُتوقع من الطلاب أن يكونوا قادرين على فعله بهذا المحتوى. أي، ما هي نواتج التعلم (Learning Outcomes) أو الأهداف السلوكية (Behavioral Objectives) التي يسعى

الاختبار لقياسها؟ هذه النواتج تصف الأداء المتوقع من الطالب والذي يمكن ملاحظته وقياسه.

يجب صياغة نواتج التعلم بعبارات واضحة ودقيقة، باستخدام أفعال سلوكية قابلة للقياس (مثل: يحدد، يصف، يقارن، يحلل، يطبق، يقيم، يصمم). وتساعد تصنيفات الأهداف التربوية، مثل تصنيف بلوم (Bloom et al., 1956) وتعديلاته اللاحقة (Anderson & Krathwohl, 2001)، في تحديد مستويات التفكير المختلفة التي يجب أن يقيسها الاختبار، بدءًا من المستويات الدنيا كالتذكر والفهم، وصولًا إلى المستويات العليا كالتحليل والتركيب والتقويم (أو الإبداع في التصنيف المعدل).

إن ربط محتوى المادة الدراسية بنواتج التعلم القابلة للقياس يضمن أن الاختبار لا يقيس مجرد حفظ المعلومات، بل يقيس أيضًا قدرة الطلاب على فهمها وتطبيقها وتحليلها وتقويمها. وهذه الخطوة ضرورية جدًا لترجمة المحتوى المجرد إلى سلوكيات قابلة للملاحظة والقياس من خلال فقرات الاختبار. على سبيل المثال، بدلاً من مجرد ذكر موضوع "الدورة الدموية"، يمكن تحديد نواتج تعلم مثل: "أن يصف الطالب مسار الدم في الدورة الدموية الكبرى والصغرى"، "أن يقارن الطالب بين الشرايين والأوردة من حيث التركيب والوظيفة"، "أن يحلل الطالب أثر ممارسة الرياضة على معدل ضربات القلب". هذه النواتج المحددة والقابلة للقياس هي التي سيتم بناء فقرات الاختبار لقياس مدى تحققها لدى الطلاب. إن تحديد الأهداف بدقة يُعد جانبًا جوهريًا في تصميم الاختبارات التحصيلية المقننة، حيث يعكف القائمون على بناء هذه الاختبارات على ترجمة الأهداف العامة للمادة إلى أهداف سلوكية محددة يمكن ملاحظتها وقياسها (كما ورد في مصدر).

2. إعداد جدول المواصفات (الخارطة الاختبارية) (Preparing the Table of Specifications - Test Blueprint)

بعد الانتهاء من تحديد الغرض من الاختبار، وتحليل المحتوى، وتحديد نواتج التعلم المراد قياسها، تأتي خطوة حاسمة ومفصلية في عملية بناء الاختبار، وهي إعداد "جدول المواصفات" (Table of Specifications)، والذي يُعرف أحيانًا بـ "الخارطة الاختبارية" (Test Blueprint). يُعد جدول المواصفات بمثابة المخطط التفصيلي الذي يربط بين محتوى المادة الدراسية والأهداف التعليمية (نواتج التعلم) ويحدد الوزن النسبي لكل منهما في الاختبار. إنه الأداة التي تضمن أن الاختبار النهائي سيكون متوازنًا وممثلاً للمجال الذي يقيسه، ويحقق الأهداف التي وُضع من أجلها.

2.1. مفهوم جدول المواصفات وأهميته (Concept and Importance of the Table of Specifications)

جدول المواصفات هو عبارة عن مخطط أو جدول ذي بعدين يوضح العلاقة بين مجالات المحتوى التي سيغطيها الاختبار، ومستويات الأهداف المعرفية أو المهارية التي ستُقاس في كل مجال من هذه المجالات. ويحدد هذا الجدول عدد الفقرات (الأسئلة) التي يجب تخصيصها لكل خلية ناتجة عن تقاطع مجال محتوى معين مع مستوى هدف معين، أو يحدد النسبة المئوية للفقرات لكل خلية (أبو لبدة، 1985).

تكمن أهمية جدول المواصفات في كونه:

أداة لتحقيق صدق المحتوى: يضمن أن فقرات الاختبار تغطي جميع جوانب المحتوى والأهداف الهامة بشكل متوازن وحسب الأهمية النسبية المحددة لها في التدريس، مما يعزز صدق المحتوى للاختبار (Linn & Gronlund, 2000).

مرشد لبناء الاختبار: يوفر خارطة طريق واضحة لمعدي الاختبار عند صياغة الفقرات، حيث يوجههم لتركيز جهودهم على

المجالات والأهداف ذات الأوزان النسبية الأعلى، ويضمن عدم إغفال أي جانب هام أو المبالغة في التركيز على جوانب أخرى أقل أهمية.

وسيلة لتحقيق التوازن: يساعد في تحقيق التوازن بين موضوعات المحتوى المختلفة، وبين المستويات المعرفية المختلفة (من التذكر إلى التقويم أو الإبداع)، مما يجعل الاختبار أكثر شمولاً ودقة في قياس تحصيل الطلاب.

أساس لبناء صور متكافئة للاختبار: يمكن استخدام جدول المواصفات كمرجع لبناء صور متعددة ومتكافئة (Parallel Forms) للاختبار، بحيث تقيس نفس المحتوى ونفس الأهداف وبنفس الأوزان النسبية، وهو أمر مهم لزيادة أمن الاختبار أو لقياس النمو في التحصيل على فترات زمنية مختلفة.

أداة للمراجعة والحكم: يُستخدم كمعيار لمراجعة الاختبار في صورته الأولية والحكم على مدى تمثيله للمحتوى والأهداف المحددة.

باختصار، يُعد جدول المواصفات العمود الفقري للاختبار الجيد، والضامن الأساسي لتمثيله للمادة الدراسية وأهدافها بشكل متوازن وشامل. وكما يشير أحد المصادر، فإن أفضل طريقة لتحديد الأهداف التربوية في منهج ما وصياغتها وتحليلها هي باستخدام جدول المواصفات، الذي يكون في الاختبارات التحصيلية المقننة مفصلاً ويتضمن تحليلاً دقيقاً للأنماط السلوكية ومجالات المحتوى (مصدر).

2.2. مكونات جدول المواصفات: المحتوى ومستويات الأهداف (Components: Content Areas and Cognitive Levels)

يتكون جدول المواصفات عادةً من بعدين رئيسيين:

بعد المحتوى (Content Dimension): يمثل هذا البعد الموضوعات أو الوحدات الدراسية أو مجالات المهارة الرئيسية التي تم تحديدها في مرحلة تحليل المحتوى. يتم إدراج هذه الموضوعات كصفوف (أو أعمدة) في الجدول.

بعد الأهداف (Objectives Dimension): يمثل هذا البعد نواتج التعلم أو الأهداف السلوكية المراد قياسها، مصنفة حسب مستوياتها المعرفية أو المهارية. غالبًا ما يُستخدم تصنيف بلوم للمجال المعرفي (Bloom et al., 1956) أو نسخته المعدلة (Anderson & Krathwohl, 2001) لهذا الغرض، والذي يتضمن مستويات مثل: التذكر (Remembering)، الفهم (Understanding)، التطبيق (Applying)، التحليل (Analyzing)، التقويم (Evaluating)، والتركيب/الإبداع (Creating). يتم إدراج هذه المستويات كأعمدة (أو صفوف) في الجدول.

ينتج عن تقاطع هذين البعدين مجموعة من الخلايا، كل خلية تمثل مستوى هدف معين ضمن مجال محتوى معين. يتم بعد ذلك تحديد الأهمية النسبية (Weight) لكل من مجالات المحتوى ولكل من مستويات الأهداف. تُحدد هذه الأهمية النسبية عادةً بناءً على الوقت المخصص لتدريس كل موضوع، أو الأهمية التي يوليها المعلمون أو الخبراء لكل موضوع أو هدف، أو تكرار التركيز عليها في المنهج. تُعبّر الأهمية النسبية عادةً كنسبة مئوية (%) من إجمالي وقت التدريس أو من إجمالي الأهمية.

2.3. خطوات بناء جدول المواصفات (Steps for Constructing the Table)

يمكن تلخيص خطوات بناء جدول المواصفات فيما يلي:

تحديد مجالات المحتوى الرئيسية: بناءً على تحليل المحتوى، يتم تحديد الموضوعات أو الوحدات الأساسية التي سيغطيها الاختبار.

تحديد مستويات الأهداف: اختيار تصنيف مناسب للأهداف (مثل تصنيف بلوم) وتحديد المستويات التي سيركز عليها الاختبار.

تحديد الأهمية النسبية لمجالات المحتوى: تقدير نسبة الأهمية لكل مجال محتوى بناءً على معايير محددة (مثل الوقت المخصص للتدريس). يجب أن يكون مجموع النسب المئوية لمجالات المحتوى 100%.

تحديد الأهمية النسبية لمستويات الأهداف: تقدير نسبة الأهمية لكل مستوى من مستويات الأهداف التي سيقيسها الاختبار. يجب أن يكون مجموع النسب المئوية لمستويات الأهداف 100%.

تحديد العدد الكلي لفقرات الاختبار: تقدير العدد المناسب من الفقرات للاختبار بناءً على عوامل مثل الوقت المتاح للاختبار، وعمر المفحوصين، وشمولية التغطية المطلوبة.

حساب عدد الفقرات لكل خلية: يتم حساب عدد الفقرات المخصصة لكل خلية في الجدول (تقاطع محتوى معين مع مستوى هدف معين) باستخدام المعادلة التالية:
عدد فقرات الخلية = العدد الكلي لفقرات الاختبار × الأهمية النسبية لمجال المحتوى × الأهمية النسبية لمستوى الهدف
على سبيل المثال، إذا كان العدد الكلي للفقرات 100 فقرة، والأهمية النسبية لموضوع "الجهاز الهضمي" هي 20%، والأهمية النسبية لمستوى "الفهم" هي 30%، فإن عدد الفقرات التي ستقيس مستوى الفهم في موضوع الجهاز الهضمي هو:
100 × 0.20 × 0.30 = 6 فقرات.

مراجعة وتدقيق الجدول: التأكد من أن مجموع عدد الفقرات في الصفوف يتوافق مع الأهمية النسبية للمحتوى، وأن مجموع عدد الفقرات في الأعمدة يتوافق مع الأهمية النسبية للأهداف، وأن المجموع الكلي للفقرات في الجدول يساوي العدد الكلي المحدد للاختبار. قد تحتاج الأعداد المحسوبة للفقرات في الخلايا إلى تقريب لأقرب عدد صحيح، مع التأكد من الحفاظ على التوازن العام قدر الإمكان.

2.4. مثال تطبيقي لجدول المواصفات (Practical Example)

لنفترض أننا نريد بناء اختبار تحصيلي في وحدة "الكهرباء والمغناطيسية" لطلاب الصف التاسع، وأن عدد فقرات الاختبار الكلي هو 50 فقرة.

مجالات المحتوى وأهميتها النسبية:

الكهرباء الساكنة (20%)

التيار الكهربائي والدوائر (40%)

المغناطيسية (20%)

الكهرومغناطيسية (20%)

المجموع = 100%

مستويات الأهداف (تصنيف بلوم المبسط) وأهميتها النسبية:

التذكر (20%)

الفهم (30%)

التطبيق (40%)

التحليل (10%)

المجموع = 100%

حساب عدد الفقرات لكل خلية: (العدد الكلي = 50)

المحتوى / الهدف	التذكر (20%)	الفهم (30%)	التطبيق (40%)	التحليل (10%)	المجموع (عدد الفقرات لكل محتوى)	الأهمية النسبية للمحتوى
الكهرباء الساكنة	2 (50x0.2x0.2)	3 (50x0.2x0.3)	4 (50x0.2x0.4)	1 (50x0.2x0.1)	10	20%
التيار والدوائر	4 (50x0.4x0.2)	6 (50x0.4x0.3)	8 (50x0.4x0.4)	2 (50x0.4x0.1)	20	40%
المغناطيسية	2 (50x0.2x0.2)	3 (50x0.2x0.3)	4 (50x0.2x0.4)	1 (50x0.2x0.1)	10	20%
الكهرومغناطيسية	2 (50x0.2x0.2)	3 (50x0.2x0.3)	4 (50x0.2x0.4)	1 (50x0.2x0.1)	10	20%
مجموع الفقرات لكل هدف	**10**	**15**	**20**	**5**	**50**	
الأهمية النسبية للهدف	**20%**	**30%**	**40%**	**10%**		**100%**

هذا الجدول يوفر الآن دليلاً واضحًا لكاتب الفقرات: عليه أن يكتب فقرتين تقيسان التذكر في الكهرباء الساكنة، وثلاث فقرات تقيس الفهم في نفس الموضوع، وهكذا لبقية الخلايا. يجب أن يتأكد من أن مجموع الفقرات في النهاية يساوي 50 فقرة، موزعة حسب الأوزان المحددة للمحتوى والأهداف.

إن إعداد جدول مواصفات دقيق وشامل يمثل استثمارًا ضروريًا للوقت والجهد، لأنه يضع الأساس المتين لاختبار تحصيلي مقنن عالي الجودة.

3. صياغة الفقرات الاختبارية (بنود الاختبار) (Writing Test Items)

تُعتبر مرحلة صياغة الفقرات الاختبارية (Test Items) أو بنود الاختبار قلب عملية بناء الاختبار. فجودة الاختبار ككل تعتمد بشكل كبير على جودة كل فقرة من فقراته. تتطلب صياغة الفقرات الجيدة مهارة فنية ومعرفة بالمبادئ العلمية

للقياس والتقويم، بالإضافة إلى فهم عميق للمادة الدراسية ولخصائص الجمهور المستهدف. إن فقرة اختبار سيئة الصياغة، حتى لو كانت ضمن اختبار مخطط له جيدًا، يمكن أن تؤدي إلى قياس غير دقيق أو مضلل لقدرات الطالب (Haladyna, Downing, & Rodriguez, 2002). لذلك، يجب أن يولى اهتمام كبير لهذه المرحلة لضمان أن تكون الفقرات واضحة، ومحددة، وتقيس الهدف المقصود بدقة، وخالية من العيوب الفنية التي قد تربك الطالب أو توجهه نحو إجابة معينة.

3.1. المبادئ العامة لصياغة الفقرات الجيدة (General Principles of Good Item Writing)

هناك مجموعة من المبادئ العامة التي يجب مراعاتها عند صياغة أي نوع من أنواع الفقرات الاختبارية لضمان جودتها (Osterlind, 1998; Popham, 2017):

قياس ناتج تعلم مهم: يجب أن تقيس كل فقرة ناتج تعلم محدد ومهم تم تحديده في جدول المواصفات، وأن تتجنب قياس معلومات تافهة أو غير ذات صلة.

الوضوح والدقة في الصياغة: يجب أن تكون لغة الفقرة واضحة ومباشرة ومفهومة للطالب، وأن تتجنب الغموض أو الازدواجية في المعنى. يجب أن يكون المطلوب من الطالب محددًا تمامًا.

الاستقلال بين الفقرات: يجب أن تكون كل فقرة مستقلة بذاتها، بحيث لا تعتمد الإجابة على فقرة معينة على معرفة الإجابة على فقرة أخرى، أو تقدم تلميحًا لإجابتها.

تجنب التلميحات غير المقصودة: يجب أن تخلو الفقرة (سواء في المقدمة أو البدائل في حالة الاختيار من متعدد) من أي تلميحات لفظية أو قواعدية قد ترشد الطالب الذكي إلى الإجابة الصحيحة دون معرفة فعلية بالمحتوى (مثل استخدام نفس الكلمة في المقدمة والإجابة الصحيحة، أو كون الإجابة الصحيحة أطول أو أكثر تفصيلاً من البدائل الأخرى بشكل ملحوظ).

ملائمة مستوى الصعوبة: يجب أن تكون صعوبة الفقرة مناسبة للمستوى العام للطلاب المستهدفين وللغرض من الاختبار. يجب تجنب الفقرات السهلة جدًا التي يجيب عليها الجميع بشكل صحيح، والفقرات الصعبة جدًا التي لا يجيب عليها أحد بشكل صحيح (إلا إذا كان الغرض من الاختبار يتطلب ذلك).

التركيز على محتوى واحد: يفضل أن تركز كل فقرة على قياس مفهوم أو مهارة واحدة محددة، وتجنب الفقرات المزدوجة التي تقيس شيئين في نفس الوقت.

تجنب العبارات المنقولة حرفيًا من الكتاب: يجب إعادة صياغة المعلومات الواردة في الكتاب المدرسي عند بناء الفقرة، لتجنب قياس مجرد الحفظ والاستظهار.

تجنب الأسئلة الخادعة أو المربكة: الهدف هو قياس فهم الطالب ومعرفته، وليس قدرته على اكتشاف الخدع أو الألغاز في صياغة السؤال.

المراجعة والتحرير: يجب مراجعة كل فقرة بعناية بعد صياغتها من قبل كاتب الفقرة نفسه ومن قبل متخصصين آخرين في المادة وفي القياس للتأكد من خلوها من الأخطاء اللغوية والعلمية والفنية.

3.2. أنواع الفقرات الشائعة ومزاياها وعيوبها (Common Item Types, Advantages, and

(Disadvantages

هناك أنواع متعددة من الفقرات الاختبارية، ويمكن تصنيفها بشكل عام إلى فئتين رئيسيتين: الفقرات الموضوعية (Objective Items) والفقرات المقالية أو الإنشائية (Essay/Constructed-Response Items).

الفئة الأولى: الفقرات الموضوعية **(Objective Items)**
تتميز هذه الفقرات بوجود إجابة صحيحة محددة أو أفضل إجابة، مما يجعل تصحيحها موضوعيًا وسريعًا ولا يتأثر بذاتية المصحح. تشمل هذه الفئة عدة أنواع، من أبرزها:

3.2.1. فقرات الاختيار من متعدد **(Multiple-Choice Items)**:

الوصف: تتكون من جزأين رئيسيين: المقدمة (Stem) التي تطرح المشكلة أو السؤال، وقائمة من البدائل (Alternatives/Options) تتضمن إجابة صحيحة واحدة (Key) ومجموعة من البدائل الخاطئة أو المشتتات (Distractors).

المزايا:

مرونة عالية في قياس مستويات معرفية متنوعة، من التذكر إلى التحليل والتقويم (إذا صيغت بعناية).

موضوعية عالية في التصحيح وسرعته، خاصة باستخدام الحاسوب أو الماسح الضوئي.

تغطية واسعة للمحتوى الدراسي في وقت قصير نسبيًا.

تقليل عامل التخمين مقارنة بفقرات الصواب والخطأ.

إمكانية التحليل الإحصائي للفقرات (صعوبة، تمييز، فعالية المشتتات) بسهولة.

العيوب:

صعوبة بناء فقرات جيدة، خاصة صياغة مشتتات فعالة ومنطقية.

قد تشجع على التعرف على الإجابة بدلاً من استدعائها أو إنشائها.

قد تكون عرضة للتخمين (وإن كان بدرجة أقل من أنواع أخرى).

تستغرق وقتًا طويلاً في الإعداد.

لا تقيس القدرة على التعبير الكتابي أو تنظيم الأفكار أو الإبداع بشكل مباشر.

3.2.2. فقرات الصواب والخطأ **(True/False Items)**:

الوصف: تتكون من عبارة خبرية يقرر الطالب ما إذا كانت صحيحة (صواب) أم خاطئة (خطأ).

المزايا:

سهولة وسرعة الصياغة نسبيًا.

سهولة وسرعة التصحيح وموضوعيته.

تسمح بتغطية كم كبير من المعلومات في وقت قصير.

مفيدة في قياس الحقائق والمفاهيم الأساسية أو التمييز بين المفاهيم المتشابهة.

العيوب:

ارتفاع احتمالية التخمين (50%).

غالبًا ما تقيس مستوى التذكر فقط، وصعوبة استخدامها لقياس مستويات عليا.

صعوبة صياغة عبارات تكون صحيحة تمامًا أو خاطئة تمامًا دون لبس أو تحفظ.

قد تشجع على الحفظ الحرفي للمعلومات.

لا توفر معلومات تشخيصية عن سبب خطأ الطالب.

3.2.3. فقرات المطابقة (Matching Items):

الوصف: تتكون من قائمتين من العبارات أو الكلمات أو الرموز (قائمة المقدمات وقائمة الإجابات)، ويطلب من الطالب ربط كل عنصر في القائمة الأولى بالعنصر المناسب له في القائمة الثانية بناءً على علاقة محددة.

المزايا:

كفاءة عالية في قياس القدرة على الربط بين المعلومات ذات العلاقة (مثل المصطلحات وتعريفاتها، الأحداث وتواريخها، العلماء وإسهاماتهم).

سهولة وسرعة التصحيح وموضوعيته.

تقلل من التخمين مقارنة بالصواب والخطأ إذا كانت قائمة الإجابات أطول من قائمة المقدمات.

اختصار للمساحة والوقت مقارنة بصياغة نفس المحتوى في شكل اختيار من متعدد.

العيوب:

غالبًا ما تقتصر على قياس مستوى التذكر والتعرف على العلاقات البسيطة.

تتطلب أن تكون جميع العناصر في كل قائمة متجانسة (تنتمي لنفس الفئة).

قد تصبح مربكة إذا كانت القوائم طويلة جدًا.

صعوبة إيجاد عدد كافٍ من العلاقات المتجانسة لبناء فقرة مطابقة جيدة.

3.2.4. فقرات الإكمال (Completion Items) أو ملء الفراغ (Fill-in-the-Blank):

الوصف: تتكون من عبارة أو جملة حُذف منها كلمة أو عبارة قصيرة مهمة، ويطلب من الطالب ملء الفراغ بالكلمة أو العبارة المناسبة.

المزايا:

تقيس القدرة على استدعاء المعلومة بدلاً من التعرف عليها.

سهولة الصياغة نسبيًا.

تقلل من التخمين بشكل كبير مقارنة بالأنواع الموضوعية الأخرى.

العيوب:

غالبًا ما تقيس مستوى التذكر للمعلومات والحقائق المحددة.

قد تحتمل أكثر من إجابة صحيحة أو شبه صحيحة، مما يؤثر على موضوعية التصحيح ويتطلب وضع قواعد واضحة له.

قد تشجع على الحفظ الحرفي إذا اقتصرت الفراغات على كلمات غير أساسية.

صعوبة استخدامها لقياس الفهم العميق أو المهارات المعقدة.

الفئة الثانية: الفقرات المقالية (Essay Items) أو الإنشائية (Constructed-Response Items)
تتطلب هذه الفقرات من الطالب إنتاج إجابته بنفسه كتابةً، بدلاً من اختيارها من بين بدائل. وتتراوح من أسئلة الإجابة القصيرة (Short-Answer Questions) التي تتطلب بضع كلمات أو جمل، إلى أسئلة الإجابة الممتدة (Extended-Response Questions) التي تتطلب تنظيمًا للأفكار وتحليلاً وتقويمًا وإبداعًا في فقرات متعددة.

الوصف: سؤال أو مشكلة يطلب من الطالب الإجابة عليها كتابةً بأسلوبه الخاص، موضحًا فهمه أو تحليله أو رأيه أو قدرته على حل المشكلات.

المزايا:

قدرة عالية على قياس مستويات التفكير العليا (التحليل، التركيب، التقويم، الإبداع).

تقييم القدرة على تنظيم الأفكار والتعبير عنها بوضوح وتسلسل منطقي.

تقييم مهارات الكتابة وحل المشكلات والتفكير الناقد.

سهولة وسرعة الإعداد نسبيًا مقارنة بفقرات الاختيار من متعدد الجيدة.

تقليل عامل التخمين إلى أدنى حد.

العيوب:

ذاتية التصحيح وصعوبة تحقيق الموضوعية والثبات بين المصححين (Inter-rater Reliability) وبين تصحيحات المصحح نفسه في أوقات مختلفة (Intra-rater Reliability). يتطلب الأمر وضع معايير تصحيح مفصلة (Scoring Rubrics) وتدريب المصححين لتقليل الذاتية.

استهلاك وقت طويل في التصحيح.

محدودية التغطية للمحتوى الدراسي، حيث يمكن تضمين عدد قليل نسبيًا من الأسئلة المقالية في الاختبار مقارنة بالأسئلة الموضوعية.

قد تتأثر درجة الطالب بعوامل غير مرتبطة بالتحصيل مثل جودة الخط وسرعة الكتابة والقدرة اللغوية العامة.

صعوبة التحليل الإحصائي المفصل للفقرات.

يعتمد اختيار نوع الفقرات المناسب على عدة عوامل، أهمها: نواتج التعلم المراد قياسها (هل هي تذكر حقائق أم تحليل مشكلات؟)، وطبيعة المادة الدراسية، وخصائص الطلاب، والوقت المتاح للاختبار وللتصحيح، والغرض من الاختبار (Downing, 2006; Welch, 2006). غالبًا ما يتضمن الاختبار التحصيلي المقنن الجيد مزيجًا من أنواع مختلفة من الفقرات لتحقيق التوازن والاستفادة من مزايا كل نوع وتلافي عيوبه قدر الإمكان.

3.3. قواعد محددة لصياغة كل نوع من الفقرات (Specific Rules for Writing Each Item Type)

بالإضافة إلى المبادئ العامة، هناك قواعد وإرشادات فنية محددة يجب مراعاتها عند صياغة كل نوع من أنواع الفقرات لزيادة فعاليتها وتقليل العيوب المحتملة (Haladyna et al., 2002; Burton, Sudweeks, Merrill, & Wood, 1991):

أ. فقرات الاختيار من متعدد:

وضوح المقدمة (Stem): يجب أن تطرح المقدمة مشكلة واضحة ومحددة بذاتها، بحيث يستطيع الطالب فهم المطلوب حتى قبل قراءة البدائل. يفضل أن تكون المقدمة في صيغة سؤال مباشر أو عبارة ناقصة تكتمل بأحد البدائل.

تضمين معظم الكلمات في المقدمة: لتجنب التكرار غير الضروري في البدائل وجعلها أقصر وأوضح.

إجابة صحيحة واحدة فقط (أو أفضل إجابة بوضوح): يجب التأكد تمامًا من أن هناك بديل واحد فقط هو الصحيح أو الأفضل بشكل لا يقبل الجدال بين الخبراء.

جاذبية المشتتات (Distractors): يجب أن تكون البدائل الخاطئة (المشتتات) правдоподобная ومعقولة للطلاب الذين لم يتقنوا المادة، وأن تستند إلى أخطاء شائعة أو مفاهيم خاطئة لديهم. يجب تجنب المشتتات الساذجة أو غير المنطقية التي يسهل استبعادها.

تجانس البدائل: يجب أن تكون جميع البدائل متجانسة في محتواها وشكلها وطولها وتركيبها اللغوي أو القواعدي قدر الإمكان، حتى لا يبرز البديل الصحيح عن غيره.

تجنب النفي المزدوج والنفي في المقدمة قدر الإمكان: عبارات النفي (مثل "ليس"، "لا"، "ما عدا") قد تربك الطلاب. إذا كان لا بد من استخدام النفي، فيجب إبرازه (بتغميق الخط أو وضع خط تحته).

تجنب عبارات مثل "كل ما سبق" أو "لا شيء مما سبق": هذه البدائل قد تسبب مشكلات فنية. إذا كانت إحدى البدائل صحيحة جزئيًا، فقد يختارها الطالب. وإذا اكتشف الطالب خطأ في بديل واحد، فسيستبعد "كل ما سبق". كما أن "لا شيء مما سبق" لا توفر معلومات حول فهم الطالب.

التوزيع العشوائي لموقع الإجابة الصحيحة: يجب توزيع موقع الإجابة الصحيحة (أ، ب، ج، د) بشكل عشوائي تقريبًا عبر فقرات الاختبار، وتجنب وجود نمط معين في ترتيبها.

الترتيب المنطقي للبدائل (إن وجد): إذا كانت البدائل عبارة عن أرقام أو تواريخ أو خطوات، فيجب ترتيبها تصاعديًا أو تنازليًا أو حسب تسلسل منطقي.

ب. فقرات الصواب والخطأ:

أن تكون العبارة صحيحة تمامًا أو خاطئة تمامًا: تجنب العبارات التي قد تكون صحيحة في بعض الظروف وخاطئة في ظروف أخرى، أو التي تتضمن كلمات توحي بعدم التأكد مثل "غالبًا"، "أحيانًا"، "قد".

التركيز على فكرة واحدة في العبارة: تجنب العبارات المركبة التي تتضمن جزأين، قد يكون أحدهما صحيحًا والآخر خاطئًا، مما يربك الطالب.

تجنب العبارات الطويلة والمعقدة: اجعل العبارة قصيرة ومباشرة قدر الإمكان.

تجنب النفي والنفي المزدوج: يزيد النفي من صعوبة فهم العبارة وقد يقيس القدرة اللغوية بدلاً من المعرفة بالمحتوى.

تجنب الكلمات المطلقة (المؤشرات النوعية): كلمات مثل "دائمًا"، "أبدًا"، "كل"، "جميع"، "فقط"، "مستحيل" غالبًا ما تجعل العبارة خاطئة (والعكس صحيح لكلمات مثل "بعض"، "أحيانًا"، "غالبًا")، وقد يتعلم الطلاب هذه القاعدة ويستخدمونها للتخمين.

توزيع عدد العبارات الصحيحة والخاطئة بالتساوي تقريبًا: لتجنب تحيز الطالب نحو اختيار معين.

ج. فقرات المطابقة:

تجانس القائمتين: يجب أن تكون جميع العناصر في قائمة المقدمات (Column A) من نفس النوع أو الفئة (مثل: علماء، تواريخ، مصطلحات)، وجميع العناصر في قائمة الإجابات (Column B) من نفس النوع (مثل: إسهامات، أحداث، تعريفات).

عدم تساوي عدد العناصر في القائمتين: يفضل أن تكون قائمة الإجابات أطول قليلاً من قائمة المقدمات (ببديل أو اثنين زيادة) لتقليل التخمين عن طريق الحذف.

وضوح العلاقة المطلوبة: يجب تحديد أساس المطابقة بوضوح في تعليمات الفقرة (مثلاً: "صل كل عالم في القائمة أ بأبرز إسهاماته في القائمة ب").

قصر العناصر: اجعل العناصر في كلتا القائمتين قصيرة وموجزة قدر الإمكان.

الترتيب المنطقي للقوائم: يفضل ترتيب عناصر قائمة الإجابات (التي سيختار منها الطالب) ترتيبًا منطقيًا (أبجديًا، زمنيًا، رقميًا) لتسهيل البحث.

وضع القائمة الأقصر كمقدمات (Column A): غالبًا ما تكون هذه هي العناصر التي يبدأ بها الطالب عملية البحث.

حصر الفقرة في صفحة واحدة: تجنب تقسيم فقرة المطابقة على صفحتين.

تحديد ما إذا كان يمكن استخدام الإجابة أكثر من مرة: يجب توضيح ذلك في التعليمات.

د. فقرات الإكمال:

أن يكون الفراغ لكلمة أو عبارة مهمة: يجب أن يمثل الجزء المحذوف معلومة أساسية أو مصطلحًا رئيسيًا، وليس كلمة غير هامة.

تجنب وجود عدد كبير من الفراغات في العبارة الواحدة: يفضل فراغ واحد أو اثنين على الأكثر، لأن كثرة الفراغات تجعل العبارة غامضة.

وضع الفراغ في نهاية العبارة أو قرب نهايتها: هذا يساعد الطالب على فهم سياق العبارة قبل الوصول إلى الفراغ.

أن تكون الإجابة المطلوبة محددة: يجب صياغة العبارة بحيث لا تحتمل إلا إجابة واحدة صحيحة أو عددًا محدودًا جدًا من الإجابات المرادفة التي يجب توقعها مسبقًا وتضمينها في مفتاح التصحيح.

جعل الفراغات متساوية في الطول: حتى لا يوحي طول الفراغ بطول الإجابة المطلوبة.

تجنب استخدام أدوات التعريف أو الإشارة قبل الفراغ مباشرة (أ، ال، an): لأنها قد توحي بالإجابة (مفرد/جمع، مذكر/مؤنث، تبدأ بحرف علة أم لا).

هـ. الفقرات المقالية:

تحديد المهمة المطلوبة بوضوح ودقة: يجب أن يعرف الطالب بالضبط ما هو مطلوب منه (قارن، حلل، اشرح، أعط رأيك مع التبرير، إلخ). تجنب الأسئلة العامة والغامضة مثل "تحدث عن...".

تحديد نطاق الإجابة وعمقها: هل المطلوب إجابة قصيرة محددة أم إجابة ممتدة تتضمن تفاصيل وأمثلة؟ يمكن تحديد عدد النقاط المطلوبة أو الوقت المخصص للإجابة للمساعدة في ذلك.

يفضل استخدام عدة أسئلة قصيرة بدلاً من سؤال واحد طويل: هذا يزيد من شمولية تغطية المحتوى ويقلل من أثر "حظ" الطالب في معرفة إجابة سؤال واحد فقط.

إعداد معايير تصحيح واضحة ومفصلة (Scoring Rubric) مسبقًا: يجب تحديد العناصر الأساسية التي يجب أن تتضمنها الإجابة المثالية، وتحديد الدرجات المخصصة لكل عنصر أو لكل مستوى من مستويات جودة الإجابة. هذا ضروري لزيادة موضوعية التصحيح.

تصحيح إجابات جميع الطلاب على سؤال واحد قبل الانتقال للسؤال التالي: هذا يساعد المصحح على الحفاظ على معيار ثابت للتقييم.

تصحيح إجابات الطلاب دون معرفة أسمائهم (إن أمكن): لتقليل التحيز الشخصي.

يفضل أن يقوم أكثر من مصحح واحد بتصحيح الإجابات (خاصة في الاختبارات الهامة): وحساب درجة الاتفاق بينهما لضمان ثبات التصحيح.

تدريب المصححين: يجب تدريب المصححين على استخدام معايير التصحيح بشكل متسق قبل البدء بعملية التصحيح الفعلية.

إن الالتزام بهذه القواعد الفنية يساعد بشكل كبير في إنتاج فقرات اختبارية ذات جودة عالية، تساهم في دقة وموثوقية القياس.

3.4. مراجعة وتحكيم الفقرات الأولية (Review and Expert Judgment of Draft Items)

بعد الانتهاء من الصياغة الأولية للفقرات وفقًا لجدول المواصفات والمبادئ والقواعد الفنية، لا يمكن اعتبار هذه الفقرات جاهزة للاستخدام مباشرة. بل يجب أن تخضع لعملية مراجعة وتحكيم دقيقة من قبل متخصصين آخرين. هذه الخطوة ضرورية للكشف عن أي أخطاء أو غموض أو تحيز قد يكون غفل عنه كاتب الفقرة (Downing & Haladyna, 2006; AERA, APA, & NCME, 2014).

تشمل عملية المراجعة والتحكيم عادةً الجوانب التالية:

المراجعة العلمية (Content Review): يقوم بها خبراء في المادة الدراسية للتأكد من دقة المحتوى العلمي للفقرة وصحة الإجابة المحددة لها.

المراجعة الفنية (Technical/Editorial Review): يقوم بها متخصصون في القياس والتقويم أو خبراء في صياغة الفقرات، وتركز على:

مدى وضوح الصياغة اللغوية وخلوها من الغموض.

الالتزام بقواعد صياغة نوع الفقرة المحدد (كما ذكر في القسم 3.3).

خلو الفقرة من التلميحات غير المقصودة للإجابة.

جاذبية المشتتات وفعاليتها (في الاختيار من متعدد).

عدم وجود تحيز ثقافي أو لغوي أو جنسي أو غيره في الفقرة.

مدى ملاءمة الفقرة للمستوى العمري والمعرفي للطلاب المستهدفين.

مطابقة الفقرة للهدف المحدد لها في جدول المواصفات.

مراجعة العدالة والتحيز (Fairness and Bias Review): يقوم بها متخصصون (قد يشملون ممثلين عن مجموعات مختلفة من المجتمع المستهدف) لمراجعة الفقرات والتأكد من أنها لا تميز بشكل غير عادل ضد أي مجموعة من الطلاب بسبب خلفيتهم الثقافية، أو العرقية، أو الاجتماعية والاقتصادية، أو نوع الجنس، أو وجود إعاقة (Zieky, 2006). يتم البحث عن أي صور نمطية، أو لغة مسيئة، أو محتوى قد يكون غير مألوف أو حساسًا لمجموعة معينة دون داعٍ.

مراجعة لغوية: التأكد من سلامة اللغة والإملاء والنحو وعلامات الترقيم.

يتم جمع ملاحظات المحكمين والمراجعين، وتستخدم هذه الملاحظات لتعديل الفقرات أو حذفها أو إعادة صياغتها. غالبًا ما يتم استخدام نماذج تحكيم معدة لهذا الغرض لتوجيه المراجعين وضمان تغطية جميع الجوانب الهامة. إن عملية المراجعة والتحكيم هذه تزيد بشكل كبير من جودة الفقرات قبل الانتقال إلى مرحلة التجريب الاستطلاعي، وتوفر الوقت والجهد لاحقًا. قد تبدو هذه الخطوة مرهقة، ولكنها استثمار ضروري لضمان جودة الاختبار النهائي.

4. التجريب الاستطلاعي للاختبار وتحليل الفقرات (Pilot Testing and Item Analysis)

بعد صياغة الفقرات ومراجعتها وتحكيمها، لا يمكن الاكتفاء بالحكم النظري على جودتها. فمهما بذل من جهد في الصياغة والمراجعة، قد تظل هناك مشكلات لا تظهر إلا عند تطبيق الفقرات فعليًا على عينة من الطلاب المشابهين للجمهور المستهدف للاختبار. هنا تأتي أهمية مرحلة التجريب الاستطلاعي (Pilot Testing) وما يتبعها من تحليل إحصائي لبيانات استجابات الطلاب على كل فقرة، وهو ما يعرف بـ "تحليل الفقرات" (Item Analysis). هذه المرحلة ضرورية للحصول على معلومات تجريبية حول كيفية أداء كل فقرة في الواقع، مما يساعد في اتخاذ قرارات مستنيرة بشأن الاحتفاظ بها أو تعديلها أو حذفها قبل تجميع الاختبار في صورته النهائية (Livingston, 2006).

4.1. أهمية التجريب الاستطلاعي (Importance of Pilot Testing)

التجريب الاستطلاعي هو تطبيق الصورة الأولية للاختبار (التي غالبًا ما تحتوي على عدد فقرات أكبر من المطلوب في النسخة النهائية) على عينة ممثلة للمجتمع الذي أُعد الاختبار له. ويهدف هذا التجريب إلى:

الحصول على بيانات إحصائية عن الفقرات: جمع استجابات الطلاب على كل فقرة لاستخدامها في حساب مؤشرات تحليل الفقرات (الصعوبة، التمييز، فعالية البدائل).

التحقق من وضوح تعليمات الاختبار والفقرات: ملاحظة ما إذا كانت التعليمات واضحة للطلاب، وما إذا كانت هناك فقرات معينة تسبب إرباكًا أو يساء فهمها بشكل متكرر.

تقدير الوقت اللازم للاختبار: معرفة متوسط الوقت الذي يستغرقه الطلاب للإجابة على الاختبار ككل وعلى الأجزاء المختلفة منه، للتأكد من أن الوقت المخصص للاختبار في صورته النهائية سيكون كافيًا.

الكشف عن أي مشكلات فنية أو إدارية: ملاحظة أي صعوبات في طريقة تطبيق الاختبار، أو ترتيب الفقرات، أو شكل الإخراج.

جمع ملاحظات الطلاب (اختياري): يمكن سؤال الطلاب بعد التجريب عن رأيهم في وضوح الفقرات، أو صعوبتها، أو أي مشكلات واجهوها.

باختصار، التجريب الاستطلاعي يوفر فرصة "لاختبار الاختبار" نفسه في ظروف شبه حقيقية قبل اعتماده النهائي، مما يسمح باكتشاف نقاط الضعف وتصحيحها.

4.2. اختيار عينة التجريب (Selecting the Pilot Sample)

للحصول على نتائج ذات معنى من التجريب الاستطلاعي وتحليل الفقرات، يجب أن تكون العينة التي يطبق عليها الاختبار ممثلة للمجتمع الأصلي للمفحوصين الذين سيستخدم معهم الاختبار لاحقًا (AERA, APA, & NCME, 2014). وهذا يعني أن العينة يجب أن تشبه المجتمع الأصلي من حيث الخصائص الديموغرافية الهامة (مثل العمر، المستوى الدراسي، التوزيع الجغرافي، الخلفية الاجتماعية والاقتصادية، إلخ) ومن حيث مستوى القدرة العام في المجال الذي يقيسه الاختبار.

يعتمد حجم العينة المطلوب للتجريب الاستطلاعي على عدة عوامل، منها الغرض من الاختبار، ونوع تحليل الفقرات المستخدم (النظرية الكلاسيكية للاختبار Classical Test Theory - CTT تتطلب عينات أصغر من نظرية الاستجابة للفقرة Item Response Theory - IRT)، ودرجة الدقة المطلوبة في تقدير مؤشرات الفقرات. كقاعدة عامة، يوصى بأن لا يقل حجم العينة عن 100 مفحوص في حالة استخدام النظرية الكلاسيكية، ويفضل أن يكون أكبر (200–500 أو أكثر) للحصول على تقديرات أكثر استقرارًا، خاصة إذا كان الاختبار سيستخدم لاتخاذ قرارات هامة (Crocker & Algina, 1986; Embretson & Reise, 2000). في حالة استخدام نماذج IRT، قد يتطلب الأمر عينات أكبر بكثير (500 أو 1000 مفحوص أو أكثر) حسب تعقيد النموذج المستخدم.

يجب اختيار العينة بطريقة تضمن التمثيل الجيد، ويفضل استخدام أساليب المعاينة العشوائية (مثل العشوائية البسيطة أو الطبقية أو العنقودية) إن أمكن.

4.3. إجراءات تطبيق الاختبار التجريبي (Procedures for Pilot Administration)

يجب أن تحاكي ظروف تطبيق الاختبار في مرحلة التجريب الاستطلاعي الظروف التي سيطبق فيها الاختبار في صورته النهائية قدر الإمكان. يتضمن ذلك:

توفير بيئة اختبار مناسبة: مكان هادئ، جيد الإضاءة والتهوية، ومريح للطلاب.

الالتزام بتعليمات موحدة: يجب أن يتلقى جميع المشاركين في التجريب نفس التعليمات الواضحة حول كيفية الإجابة، والوقت المتاح، وكيفية التعامل مع التخمين (إذا كان ذلك ذا صلة).

تحديد وقت كاف: يجب منح الطلاب وقتًا كافيًا للإجابة على جميع الفقرات دون استعجال مفرط، حتى لو كان هذا الوقت أطول قليلاً مما سيخصص للاختبار النهائي. الهدف هو الحصول على أفضل أداء ممكن من الطلاب على كل فقرة.

تحفيز الطلاب: يجب تشجيع الطلاب على بذل قصارى جهدهم والتعامل مع الاختبار بجدية، مع توضيح أن النتائج ستستخدم لتطوير الاختبار وليس لتقييمهم شخصيًا.

بعد جمع أوراق الإجابة، يتم تصحيحها بدقة وفقًا لمفتاح التصحيح المعد مسبقًا.

4.4. التحليل الإحصائي للفقرات (Statistical Item Analysis)

بعد تصحيح استجابات عينة التجريب، يتم إجراء تحليل إحصائي لكل فقرة من فقرات الاختبار بهدف تقييم جودتها وفعاليتها. يُعد تحليل الفقرات خطوة حيوية لاتخاذ قرارات مستنيرة حول أي الفقرات يجب الاحتفاظ بها، وأيها تحتاج إلى تعديل، وأيها يجب حذفها من النسخة النهائية للاختبار. يركز تحليل الفقرات عادةً، ضمن إطار النظرية الكلاسيكية للاختبار (CTT)، على ثلاثة مؤشرات رئيسية (Crocker & Algina, 1986; Osterlind, 1998; Livingston, 2006):

4.4.1. معامل الصعوبة (Difficulty Index - p):

المفهوم: يشير معامل الصعوبة لفقرة ما إلى نسبة الطلاب في عينة التجريب الذين أجابوا على الفقرة إجابة صحيحة. يُحسب بالمعادلة:

$p = R / N$

حيث R هو عدد الطلاب الذين أجابوا إجابة صحيحة، و N هو العدد الكلي للطلاب الذين حاولوا الإجابة على الفقرة.

التفسير: تتراوح قيمة معامل الصعوبة بين 0 (لم يجب أحد بشكل صحيح، فقرة صعبة جدًا) و 1 (أجاب الجميع بشكل صحيح، فقرة سهلة جدًا). على سبيل المثال، إذا كان $p = 0.75$ لفقرة ما، فهذا يعني أن 75% من الطلاب أجابوا عليها بشكل صحيح، مما يشير إلى أنها فقرة سهلة نسبيًا.

الاستخدام: يساعد معامل الصعوبة في الحكم على مدى ملاءمة صعوبة الفقرة للمستوى العام للطلاب. الفقرات ذات الصعوبة المتوسطة (تتراوح عادة بين 0.3 و 0.7) غالبًا ما تكون مفضلة في الاختبارات محكية المعيار (Norm-Referenced Tests) لأنها تزيد من قدرة الاختبار على التمييز بين مستويات الطلاب المختلفة. ومع ذلك، قد تكون هناك حاجة لفقرات سهلة جدًا (لتحفيز الطلاب في بداية الاختبار) أو فقرات صعبة جدًا (لقياس مستويات الأداء العليا) حسب الغرض من الاختبار. في الاختبارات محكية المرجع (Criterion-Referenced Tests)، قد يكون من المقبول وجود فقرات سهلة إذا كانت تقيس أهدافًا أساسية يُتوقع من معظم الطلاب إتقانها.

القرار: الفقرات التي تكون سهلة جدًا (p قريب من 1) أو صعبة جدًا (p قريب من 0) غالبًا ما يتم فحصها بعناية. قد تحتاج إلى تعديل أو حذف، خاصة إذا كانت لا تميز جيدًا بين الطلاب.

4.4.2. معامل التمييز (Discrimination Index - D أو r):

المفهوم: يشير معامل التمييز إلى قدرة الفقرة على التمييز بين الطلاب ذوي الأداء المرتفع (الذين حصلوا على درجات عالية في الاختبار ككل) والطلاب ذوي الأداء المنخفض (الذين حصلوا على درجات منخفضة في الاختبار ككل). يفترض أن الطالب ذا الأداء المرتفع يجب أن يكون أكثر احتمالاً للإجابة على الفقرة بشكل صحيح من الطالب ذي الأداء المنخفض.

الحساب (طريقة المجموعتين المتطرفتين):

ترتيب الطلاب تنازليًا حسب درجتهم الكلية في الاختبار.

تحديد مجموعة عليا (High Group - H)، غالبًا ما تكون أعلى 27% من الطلاب (أو نسبة أخرى مثل 25% أو 33%).

تحديد مجموعة دنيا (Low Group - L)، غالبًا ما تكون أدنى 27% من الطلاب (أو نفس النسبة المستخدمة للمجموعة العليا).

حساب نسبة الطلاب الذين أجابوا إجابة صحيحة في المجموعة العليا (P_H = R_H / N_H).

حساب نسبة الطلاب الذين أجابوا إجابة صحيحة في المجموعة الدنيا (P_L = R_L / N_L).

حساب معامل التمييز (D) بالفرق بين النسبتين: D = P_H - P_L

الحساب (طريقة معامل الارتباط): يمكن أيضًا حساب معامل التمييز عن طريق حساب معامل الارتباط (مثل معامل ارتباط بوينت بايسيريال Point-Biserial Correlation) بين درجة الطالب على الفقرة (0 للخطأ و 1 للصواب) ودرجته الكلية في الاختبار (<r_{pbis}).

التفسير: تتراوح قيمة معامل التمييز (D) عادة بين ـ1 و +1.

تمييز موجب مرتفع (مثل **D > 0.40**): يشير إلى أن الفقرة تميز بشكل جيد جدًا بين المجموعتين (الطلاب ذوو الأداء المرتفع يجيبون عليها بشكل صحيح أكثر بكثير من ذوي الأداء المنخفض). هذه هي الفقرات المرغوبة عادةً.

تمييز موجب مقبول (مثل **0.20 ≤ D ≤ 0.39**): الفقرة تميز بشكل مقبول، وقد تكون قابلة للاستخدام، ربما بعد بعض المراجعة.

تمييز موجب منخفض (مثل **0 < D < 0.20**): الفقرة ضعيفة التمييز، وتحتاج إلى مراجعة جوهرية أو حذف.

تمييز صفري (**D ≈ 0**): الفقرة لا تميز على الإطلاق بين المجموعتين. قد تكون سهلة جدًا أو صعبة جدًا أو غامضة أو تقيس شيئًا مختلفًا عن بقية الاختبار. غالبًا ما تُحذف.

تمييز سالب (**D < 0**): يشير إلى مشكلة خطيرة في الفقرة. الطلاب ذوو الأداء المنخفض يجيبون عليها بشكل صحيح أكثر من ذوي الأداء المرتفع! قد يكون هناك خطأ في مفتاح التصحيح، أو أن الفقرة مضللة جدًا، أو تقيس مفهومًا خاطئًا. يجب حذف هذه الفقرات أو إعادة صياغتها بالكامل بعد فحص دقيق لسبب التمييز السلبي.

القرار: يتم عادةً الاحتفاظ بالفقرات ذات التمييز الموجب المرتفع والمقبول. الفقرات ذات التمييز المنخفض أو الصفري أو السلبي تحتاج إلى فحص دقيق وقد يتم حذفها أو تعديلها جذريًا.

4.4.3. تحليل فعالية البدائل الخاطئة (**Distractor Analysis** - لفقرات الاختيار من متعدد):

المفهوم: يهدف هذا التحليل إلى تقييم مدى فعالية كل بديل من البدائل الخاطئة (المشتتات) في جذب الطلاب الذين لم يعرفوا الإجابة الصحيحة، وخاصة الطلاب ذوي الأداء المنخفض.

الإجراء: يتم فحص توزيع استجابات الطلاب (من المجموعتين العليا والدنيا، أو من جميع الطلاب) على كل بديل من بدائل الفقرة (بما في ذلك الإجابة الصحيحة).

التفسير:

المشتت الفعال: هو البديل الخاطئ الذي يختاره عدد من الطلاب (خاصة من المجموعة الدنيا) أكبر من أولئك الذين يختارون الإجابة الصحيحة بالصدفة، ولكن يختاره عدد قليل جدًا (أو لا أحد) من طلاب المجموعة العليا.

المشتت غير الفعال: هو البديل الذي لا يختاره أحد تقريبًا، أو يختاره عدد قليل جدًا من الطلاب. هذا المشتت لا يؤدي وظيفته في التضليل ويقلل فعليًا من عدد البدائل المعقولة، مما يزيد من فرصة التخمين.

المشتت ذو المشكلة: هو البديل الخاطئ الذي يختاره عدد كبير من طلاب المجموعة العليا، ربما أكثر من اختيارهم للإجابة الصحيحة. هذا يشير إلى أن المشتت قد يكون صحيحًا جزئيًا، أو أن هناك خطأ في مفتاح التصحيح، أو أن الفقرة نفسها مضللة.

القرار: يجب مراجعة الفقرات التي تحتوي على مشتتات غير فعالة أو مشتتات ذات مشكلات. قد تحتاج المشتتات غير الفعالة إلى إعادة صياغة لتصبح أكثر جاذبية. الفقرات التي بها مشتتات تجذب الطلاب المتفوقين تحتاج إلى فحص دقيق وربما تعديل أو حذف.

4.5. اتخاذ القرارات بشأن الفقرات (Making Decisions About Items - retain, revise, discard)

بناءً على نتائج تحليل الصعوبة والتمييز وفعالية المشتتات، بالإضافة إلى المراجعة النوعية لمحتوى الفقرة وصياغتها، يتم اتخاذ قرار بشأن كل فقرة من فقرات التجريب الاستطلاعي:

الاحتفاظ بالفقرة (Retain): الفقرات التي تظهر مؤشرات إحصائية جيدة (صعوبة مناسبة، تمييز مرتفع، مشتتات فعالة وخالية من العيوب النوعية يتم الاحتفاظ بها لتضمينها في النسخة النهائية للاختبار أو في بنك الفقرات (Item Bank).

تعديل الفقرة (Revise): الفقرات التي تظهر بعض المشكلات (مثل صعوبة غير مناسبة قليلاً، تمييز مقبول ولكنه ليس مرتفعًا، مشتت أو اثنان غير فعال) ولكنها تقيس هدفًا تعليميًا هامًا قد يتم تعديلها لتحسين جودتها. قد يشمل التعديل إعادة صياغة المقدمة، أو استبدال مشتت غير فعال، أو تبسيط اللغة. الفقرات المعدلة يجب أن تخضع للمراجعة وربما التجريب مرة أخرى.

حذف الفقرة (Discard): الفقرات التي تظهر مشكلات خطيرة (مثل تمييز صفري أو سلبي، صعوبة شديدة أو سهولة شديدة مع تمييز ضعيف، خطأ في المحتوى، غموض كبير، تحيز واضح) يتم حذفها نهائيًا من الاختبار.

تعتبر عملية تحليل الفقرات واتخاذ القرارات بناءً عليها عملية تكرارية قد تتطلب أكثر من جولة من التجريب والتعديل للوصول إلى مجموعة قوية من الفقرات ذات الخصائص السيكومترية الجيدة. إنها خطوة لا غنى عنها لضمان جودة الاختبار المقنن النهائي (Embretson & Reise, 2000). الهدف النهائي هو بناء اختبار يتكون من أفضل الفقرات الممكنة التي تم اختيارها بعناية بناءً على أدلة تجريبية ونوعية (مصدر ,).

5. تقويم الخصائص السيكومترية للاختبار (الصدق والثبات) (Evaluating Psychometric Properties: Validity and Reliability)

بعد اختيار الفقرات النهائية بناءً على التجريب الاستطلاعي وتحليل الفقرات، وقبل أن يصبح الاختبار جاهزًا للاستخدام الفعلي، لا بد من تقويم جودته الكلية من الناحية السيكومترية. يُعد التحقق من صدق (Validity) وثبات (Reliability) الاختبار من أهم الخطوات في هذه المرحلة، فهما يمثلان المعيارين الأساسيين للحكم على جودة أي أداة قياس تربوية أو نفسية (AERA, APA, & NCME, 2014; Cronbach, 1990). فاختبار غير صادق أو غير ثابت لا يمكن الوثوق بنتائجه أو الاعتماد عليها في اتخاذ أي قرارات هامة.

5.1. مفهوم الصدق وأنواعه وأساليب التحقق منه (Concept, Types, and Methods of Assessing Validity)

المفهوم: الصدق هو المفهوم الأكثر أهمية في القياس التربوي والنفسي. بشكل عام، يشير الصدق إلى الدرجة التي يقيس بها الاختبار ما يُفترض أن يقيسه، ومدى ملاءمة وصحة التفسيرات والاستخدامات المستندة إلى درجات الاختبار (Messick, 1989; Kane, 2013). الصدق ليس خاصية مطلقة للاختبار نفسه (أي لا نقول "الاختبار صادق" بشكل عام)،

بل هو يتعلق بدرجة صحة الاستدلالات التي تُبنى على درجات الاختبار لغرض معين وفي سياق معين. أي أننا نجمع أدلة لدعم تفسير معين لدرجات الاختبار.

أنواع أدلة الصدق: وفقًا للمعايير الحديثة (AERA, APA, & NCME, 2014)، يُنظر إلى الصدق كمفهوم موحد (Unitary Concept)، ولكن يمكن جمع أدلة عليه من مصادر متنوعة. تقليديًا، كان يتم الحديث عن "أنواع" الصدق، ولكن الآن يُفضل الحديث عن "مصادر الأدلة" التي تساهم في بناء حجة الصدق (Validity Argument). تشمل هذه المصادر ما كان يعرف سابقًا بأنواع الصدق الرئيسية:

5.1.1. أدلة الصدق المستندة إلى محتوى الاختبار (Evidence Based on Test Content - صدق المحتوى سابقًا):

المفهوم: تشير إلى مدى تمثيل فقرات الاختبار للمحتوى الدراسي أو مجال السلوك المراد قياسه تمثيلاً شاملاً ومتوازنًا. أي هل تغطي الفقرات جميع الجوانب الهامة للمجال وفقًا للأوزان المحددة في جدول المواصفات؟

أساليب التحقق:

الاعتماد على جدول المواصفات: يُعد بناء الاختبار وفقًا لجدول مواصفات مفصل ودقيق هو الإجراء الأساسي لضمان هذا النوع من الأدلة.

تحكيم الخبراء: عرض فقرات الاختبار على مجموعة من الخبراء في المادة الدراسية وفي القياس للحكم على مدى انتماء كل فقرة للمحتوى المحدد ومدى تغطية الفقرات ككل للمجال المستهدف وملاءمتها للأهداف. يتم جمع تقديراتهم وتحليلها (مثل استخدام نسبة الاتفاق أو معامل لوكهارت).

الأهمية: هذا النوع من الأدلة ضروري جدًا للاختبارات التحصيلية واختبارات الكفاءة المهنية.

5.1.2. أدلة الصدق المستندة إلى علاقة الاختبار بمتغيرات أخرى (Evidence Based on Relations to Other Variables - الصدق المرتبط بمحك سابقًا):

المفهوم: تشير إلى مدى ارتباط درجات الاختبار بدرجات مقياس آخر (يسمى المحك Criterion) يُفترض أن يقيس نفس السمة أو سمة ذات علاقة بها، أو يقيس أداءً مستقبليًا.

الأنواع الفرعية وأساليب التحقق:

الأدلة التلازمية (Concurrent Evidence - الصدق التلازمي سابقًا): يتم جمع بيانات الاختبار وبيانات المحك في نفس الوقت تقريبًا، ثم يحسب معامل الارتباط بينهما. المحك هنا قد يكون اختبارًا آخر صادقًا يقيس نفس السمة، أو تقديرات المعلمين، أو مقاييس أداء حالية. معامل ارتباط مرتفع وإيجابي يشير إلى وجود دليل على الصدق التلازمي. مثال: حساب الارتباط بين درجات اختبار تحصيلي جديد في الرياضيات ودرجات الطلاب في اختبار رياضيات آخر معروف بصداقته أو تقديرات معلميهم الحالية.

الأدلة التنبؤية (Predictive Evidence - الصدق التنبؤي سابقًا): يتم تطبيق الاختبار أولاً، ثم بعد فترة زمنية معينة (قد تكون شهورًا أو سنوات)، يتم جمع بيانات عن أداء الأفراد على المحك المستقبلي. يحسب معامل الارتباط بين درجات الاختبار ودرجات المحك. معامل ارتباط مرتفع وإيجابي يشير إلى قدرة الاختبار على التنبؤ بالأداء المستقبلي. مثال: حساب الارتباط بين درجات اختبار القبول الجامعي (الاختبار) ومعدل الطالب التراكمي في السنة الأولى بالجامعة (المحك).

الأهمية: هذا النوع من الأدلة مهم جدًا للاختبارات المستخدمة في الاختيار والتصنيف والتنبؤ بالأداء المستقبلي. يتوقف نجاحه على جودة المحك المختار ومدى موثوقيته وملاءمته.

5.1.3. أدلة الصدق المستندة إلى البنية الداخلية للاختبار (Evidence Based on Internal Structure - صدق البناء سابقًا):

المفهوم: تشير إلى مدى تطابق البنية الداخلية للاختبار (العلاقات بين الفقرات والأبعاد التي تقيسها) مع البنية النظرية للمفهوم أو السمة (Construct) التي يُفترض أن يقيسها الاختبار. هل تقيس الفقرات بالفعل السمة المقصودة؟ هل الأبعاد الفرعية المفترضة للاختبار تظهر تجريبيًا في استجابات الطلاب؟

أساليب التحقق:

تحليل الاتساق الداخلي: حساب معاملات الثبات التي تقيس تجانس الفقرات (مثل ألفا كرونباخ). التجانس العالي قد يشير إلى أن الفقرات تقيس سمة واحدة مشتركة.

التحليل العاملي (Factor Analysis): أسلوب إحصائي متقدم يستخدم لاستكشاف الأبعاد الكامنة التي تقيسها مجموعة من الفقرات. يتم تطبيقه على مصفوفة الارتباطات بين الفقرات لتحديد ما إذا كانت الفقرات تتجمع معًا في عوامل (أبعاد) تتفق مع البنية النظرية المفترضة للمفهوم.

تحليل الفروق بين المجموعات: مقارنة أداء مجموعات مختلفة يُتوقع نظريًا أن تختلف في السمة المقاسة (مثل مقارنة أداء الخبراء والمبتدئين، أو مجموعات عمرية مختلفة). إذا أظهر الاختبار الفروق المتوقعة، فهذا يدعم صدق بنائه.

مصفوفة السمات المتعددة والطرق المتعددة (Multitrait-Multimethod Matrix - MTMM): أسلوب يقارن الارتباطات بين مقاييس مختلفة تقيس سمات مختلفة باستخدام طرق مختلفة لتقييم الصدق التقاربي (Convergent Validity - الارتباط العالي بين مقاييس السمة نفسها بطرق مختلفة) والصدق التمايزي (Discriminant Validity - الارتباط المنخفض بين مقاييس سمات مختلفة حتى لو استخدمت نفس الطريقة).

الأهمية: هذا النوع من الأدلة أساسي لفهم ماذا يقيس الاختبار بالفعل، وهو مهم لجميع أنواع الاختبارات، خاصة تلك التي تقيس سمات نفسية أو قدرات عقلية مجردة.

5.1.4. أدلة الصدق المستندة إلى عمليات الاستجابة (Evidence Based on Response Processes):

المفهوم: تحليل العمليات العقلية أو الاستراتيجيات التي يستخدمها المفحوصون عند الإجابة على فقرات الاختبار للتأكد من أنها تتطابق مع العمليات التي يفترض أن تستثيرها الفقرة. هل يفكر الطلاب بالطريقة التي يتوقعها مصمم الاختبار عند حلهم لمسألة رياضية مثلاً؟

أساليب التحقق:

مقابلات "التفكير بصوت عال" (Think-aloud protocols): الطلب من المفحوصين التحدث عن أفكارهم وعملياتهم الذهنية أثناء الإجابة على الفقرات.

تحليل أخطاء الطلاب: فحص أنواع الأخطاء الشائعة التي يقع فيها الطلاب لفهم الصعوبات التي يواجهونها أو المفاهيم الخاطئة لديهم.

تحليل زمن الاستجابة: دراسة الوقت الذي يستغرقه الطلاب للإجابة على أنواع مختلفة من الفقرات.

5.1.5. أدلة الصدق المستندة إلى عواقب الاختبار (Evidence Based on Consequences of Testing):

المفهوم: دراسة الآثار الإيجابية والسلبية المقصودة وغير المقصودة المترتبة على استخدام الاختبار على الأفراد والنظام التعليمي والمجتمع. هل يحقق الاختبار الفوائد المرجوة منه؟ وهل يؤدي إلى عواقب سلبية غير متوقعة (مثل تضييق المنهج، أو زيادة قلق الطلاب، أو التمييز ضد فئات معينة)؟

أساليب التحقق:

دراسات تقييم الأثر: تحليل تأثير استخدام الاختبار على الممارسات التعليمية، أو دافعية الطلاب، أو القرارات المتخذة بناءً عليه.

تحليل العدالة: دراسة ما إذا كان الاختبار يؤدي إلى نتائج مختلفة بشكل منهجي لمجموعات فرعية مختلفة (مثل المجموعات العرقية أو الجنسية) لا يمكن تفسيرها بفروق حقيقية في السمة المقاسة (مفهوم الأداء التفاضلي للفقرة Differential Item Functioning - DIF).

الأهمية: هذا الجانب من الصدق يثير جدلاً حول ما إذا كان ينبغي اعتباره جزءًا من الصدق الفني للاختبار أم جانبًا يتعلق بسياسات استخدام الاختبار. ومع ذلك، تؤكد المعايير الحديثة (AERA, APA, & NCME, 2014) على أهمية النظر في العواقب كجزء من عملية تقييم الصدق الكلية، خاصة عند استخدام الاختبار لاتخاذ قرارات عالية المخاطر.

ملاحظة حول الصدق الظاهري (Face Validity): يشير إلى مدى ظهور الاختبار بمظهر مناسب ومعقول للمفحوصين وغير المتخصصين. أي هل يبدو الاختبار "على وجهه" أنه يقيس ما يفترض أن يقيسه؟ على الرغم من أنه لا يعتبر دليلاً علميًا قويًا على الصدق بالمعنى الفني، إلا أن الصدق الظاهري مهم لقبول المفحوصين للاختبار وتعاونهم وزيادة دافعيتهم. يجب أن تبدو الفقرات والتعليمات مناسبة وذات صلة.

إن عملية جمع أدلة الصدق هي عملية مستمرة تبدأ مع تصميم الاختبار وتستمر حتى بعد نشره واستخدامه. وكلما توفرت أدلة قوية ومتنوعة من مصادر مختلفة تدعم التفسيرات والاستخدامات المقصودة لدرجات الاختبار، زادت الثقة في صدق هذه الدرجات (Kane, 2013).

5.2. مفهوم الثبات وأنواعه وطرق حسابه (Concept, Types, and Methods of Estimating Reliability)

المفهوم: يشير الثبات إلى مدى اتساق أو استقرار درجات الاختبار عبر تطبيقات مختلفة أو ظروف قياس مختلفة. أي إلى أي مدى يمكن الاعتماد على درجة الاختبار كتقدير دقيق لمستوى الأداء الحقيقي للفرد؟ اختبار ثابت هو الذي يعطي نتائج متقاربة إذا طبق على نفس الأفراد في مناسبات مختلفة (بافتراض عدم تغير السمة المقاسة)، أو إذا استخدمت صور متكافئة منه، أو إذا تم تصحيحه بواسطة مصححين مختلفين (في حالة الفقرات المقالية) (Crocker & Algina, 1986; Traub, 1994).

يرتبط الثبات بخلو القياس من الخطأ العشوائي (Random Error). كل قياس يتضمن درجة حقيقية (True Score) ودرجة خطأ (Error Score). الثبات هو نسبة التباين في الدرجات الحقيقية إلى التباين الكلي في الدرجات الملاحظة. كلما قل

خطأ القياس العشوائي، زاد ثبات الاختبار.

أنواع (طرق تقدير) الثبات: هناك عدة طرق لتقدير معامل الثبات (Reliability Coefficient)، الذي تتراوح قيمته عادة بين 0 (انعدام الثبات) و 1 (ثبات تام). اختيار الطريقة المناسبة يعتمد على طبيعة الاختبار ومصادر الخطأ المحتملة التي يراد تقييمها:

5.2.1. طريقة إعادة الاختبار (Test-Retest Reliability):

الإجراء: تطبيق نفس الاختبار على نفس المجموعة من الأفراد مرتين تفصل بينهما فترة زمنية مناسبة (ليست قصيرة جدًا فيتذكر الطلاب الإجابات، وليست طويلة جدًا فتتغير السمة المقاسة).

الحساب: حساب معامل الارتباط (مثل ارتباط بيرسون) بين درجات الأفراد في التطبيق الأول ودرجاتهم في التطبيق الثاني.

ماذا يقيس: مدى استقرار الدرجات عبر الزمن (Stability over time). يقيم أثر التغيرات العشوائية في الأفراد أو ظروف الاختبار بين المرتين.

العيوب: يتأثر بطول الفترة الزمنية بين التطبيقين. قد يحدث تعلم أو نسيان بين المرتين. غير مناسب للاختبارات التي تتأثر بالسرعة أو التي تتغير السمة المقاسة فيها بسرعة.

5.2.2. طريقة الصور المتكافئة (Parallel/Alternate Forms Reliability):

الإجراء: بناء صورتين متكافئتين تمامًا من الاختبار (نفس المحتوى، نفس الأهداف، نفس عدد الفقرات، نفس مستوى الصعوبة والتمييز). يتم تطبيق الصورتين على نفس المجموعة من الأفراد، إما في نفس الجلسة (لتجنب أثر الزمن) أو في جلستين متقاربتين.

الحساب: حساب معامل الارتباط بين درجات الأفراد على الصورة الأولى ودرجاتهم على الصورة الثانية.

ماذا يقيس: مدى التكافؤ بين الصورتين (Equivalence). يقيم أثر الاختلاف العشوائي في محتوى الفقرات بين الصورتين (خطأ العينة من الفقرات). إذا طبقتا في زمنين مختلفين، فإنه يقيس الاستقرار والتكافؤ معًا.

العيوب: صعوبة وتكلفة بناء صورتين متكافئتين تمامًا. قد يظل هناك أثر للإرهاق أو التدريب إذا طبقتا في نفس الجلسة.

5.2.3. طريقة التجزئة النصفية (Split-Half Reliability):

الإجراء: تطبيق الاختبار مرة واحدة فقط. بعد التصحيح، يتم تقسيم فقرات الاختبار إلى نصفين متكافئين قدر الإمكان (مثل تقسيم الفقرات الفردية مقابل الزوجية، أو النصف الأول مقابل النصف الثاني، أو تقسيم عشوائي).

الحساب: حساب معامل الارتباط بين درجات الأفراد على النصف الأول ودرجاتهم على النصف الثاني. نظرًا لأن هذا الارتباط يمثل ثبات نصف الاختبار فقط، يتم تعديله باستخدام معادلة سبيرمان-براون للتصحيح (Spearman-Brown Prophecy Formula) للحصول على تقدير لثبات الاختبار كاملاً:

(<r₁₂) / (1 + r₁₂ * 2) = (الثبات الكلي) <r_{xx'}

حيث <r₁₂ هو معامل الارتباط بين النصفين.

ماذا يقيس: مدى الاتساق الداخلي (Internal Consistency) بين نصفي الاختبار. يقيم أثر الاختلاف العشوائي في محتوى الفقرات بين النصفين.

العيوب: يعتمد تقدير الثبات على طريقة تقسيم الاختبار إلى نصفين. لا يقيم استقرار الدرجات عبر الزمن. غير مناسب لاختبارات السرعة (Speed Tests).

5.2.4. طرق الاتساق الداخلي الأخرى (Other Internal Consistency Methods):

المفهوم: هذه الطرق تقيس أيضًا مدى تجانس فقرات الاختبار واتساقها في قياس نفس السمة أو المفهوم، وتتطلب تطبيق الاختبار مرة واحدة فقط. تعتبر امتدادًا لفكرة التجزئة النصفية، حيث تعامل كل فقرة كأنها اختبار صغير بحد ذاته.

أشهر المعاملات:

معامل ألفا كرونباخ (Cronbach's Alpha - α): هو المعامل الأكثر شيوعًا لتقدير الاتساق الداخلي. يمكن اعتباره متوسط جميع معاملات التجزئة النصفية الممكنة. يستخدم للاختبارات التي تكون فقراتها ذات استجابات متدرجة (مثل مقاييس ليكرت) أو يمكن تسجيلها كـ 0 و 1. يعتمد على عدد الفقرات ومتوسط الارتباط بينها.

معادلات كودر-ريتشاردسون (Kuder-Richardson Formulas - KR-20 and KR-21): تستخدم خصيصًا لتقدير الاتساق الداخلي للاختبارات التي تكون فقراتها ثنائية التصحيح (dichotomous scoring - أي إجابة صحيحة أو خاطئة، تسجل كـ 1 أو 0). KR-20 هي حالة خاصة من ألفا كرونباخ للبيانات الثنائية وتعتمد على تباين كل فقرة. KR-21 هي صيغة أبسط تفترض أن جميع الفقرات متساوية الصعوبة (وهو افتراض نادرًا ما يتحقق بدقة).

ماذا تقيس: درجة التجانس بين فقرات الاختبار. أي إلى أي مدى تقيس جميع الفقرات نفس البناء أو السمة الأساسية.

العيوب: تتأثر بطول الاختبار (تزداد بزيادة عدد الفقرات المتجانسة). قد تعطي تقديرًا منخفضًا للثبات إذا كان الاختبار يقيس سمات متعددة الأبعاد. غير مناسبة لاختبارات السرعة. لا تقيم استقرار الدرجات عبر الزمن.

ثبات المصححين (Scorer Reliability / Inter-rater Reliability):

المفهوم: هذا النوع من الثبات خاص بالاختبارات التي تتضمن فقرات ذاتية التصحيح (مثل الفقرات المقالية أو تقييم الأداء)، ويهدف إلى قياس مدى الاتفاق بين مصححين مختلفين عند تقييمهم لنفس الإجابات.

الإجراء: يقوم مصححان أو أكثر بتقييم نفس العينة من إجابات الطلاب بشكل مستقل، باستخدام نفس معايير التصحيح.

الحساب: حساب معامل الاتفاق بين تقديرات المصححين (مثل معامل ارتباط بيرسون، أو معامل كابا لكوهين للبيانات الفئوية، أو معامل الارتباط داخل الفئة Intraclass Correlation - ICC).

ماذا يقيس: درجة الموضوعية في عملية التصحيح.

الأهمية: ضروري لضمان عدم تأثر درجات الطلاب بذاتية المصحح. يمكن تحسينه عن طريق استخدام معايير تصحيح واضحة وتدريب المصححين.

اختيار طريقة الثبات المناسبة: يعتمد على مصدر الخطأ الرئيسي الذي يثير القلق. إذا كان القلق هو استقرار الدرجات عبر

الزمن، تستخدم إعادة الاختبار. إذا كان القلق هو الاختلاف بين صور الاختبار، تستخدم الصور المتكافئة. إذا كان القلق هو عدم تجانس محتوى الفقرات، تستخدم طرق الاتساق الداخلي. وإذا كان القلق هو ذاتية التصحيح، يقاس ثبات المصححين. غالبًا ما يُطلب تقديم تقديرات للثبات من أكثر من طريقة للحصول على صورة أكمل عن موثوقية الاختبار.

مستوى الثبات المقبول: يعتمد على الغرض من استخدام الاختبار. للاختبارات التي تستخدم لاتخاذ قرارات هامة حول الأفراد (مثل القبول أو التشخيص)، يُفضل أن يكون معامل الثبات 0.90 فأكثر. للاختبارات المستخدمة لتقييم أداء المجموعات أو لأغراض البحث، قد يكون معامل الثبات 0.70 أو 0.80 مقبولاً. لا توجد قاعدة صارمة، ولكن بشكل عام، كلما زاد الثبات كان أفضل (Nunnally & Bernstein, 1994).

5.3. العوامل المؤثرة في الصدق والثبات (Factors Affecting Validity and Reliability)

هناك عدة عوامل يمكن أن تؤثر على صدق وثبات الاختبار، ويجب على مطوري الاختبار ومستخدميه أن يكونوا على دراية بها:

عوامل تؤثر في الصدق:

عوامل متعلقة بالاختبار نفسه:

وضوح التعليمات والفقرات: الغموض يؤدي إلى سوء فهم ويقلل الصدق.

مستوى صعوبة الفقرات: إذا كانت الفقرات سهلة جدًا أو صعبة جدًا، فلن تميز بين الطلاب وبالتالي لن تكون مرتبطة جيدًا بالمحكات أو السمات الأخرى.

جودة صياغة الفقرات: الفقرات السيئة الصياغة أو التي بها تلميحات تقلل الصدق.

قصور تمثيل المحتوى: عدم تغطية جوانب هامة من المحتوى (ضعف صدق المحتوى).

طول الاختبار: الاختبار القصير جدًا قد لا يمثل المحتوى جيدًا.

ترتيب الفقرات: وضع الفقرات الصعبة جدًا في البداية قد يثبط الطلاب.

عوامل متعلقة بتطبيق الاختبار وتصحيحه:

ظروف التطبيق غير الموحدة: اختلاف الوقت المسموح به، أو مستوى الضوضاء، أو تعليمات المشرفين.

ذاتية التصحيح (للمقالي): عدم وجود معايير واضحة أو عدم التزام المصححين بها.

الغش أو المساعدة غير المشروعة.

عوامل متعلقة بالمفحوصين:

قلق الاختبار: قد يؤثر سلبًا على أداء بعض الطلاب.

الدافعية والجهد المبذول.

الحالة الصحية أو الانفعالية للطالب وقت الاختبار.

التخمين.

عوامل متعلقة بالمحك (في الصدق المرتبط بمحك):

عدم موثوقية المحك: لا يمكن أن يرتبط اختبار بمتغير غير موثوق به ارتباطًا عاليًا.

تلوث المحك (Criterion Contamination): معرفة المصحح أو المقيم لدرجات الأفراد على الاختبار قد يؤثر على تقييمهم لأدائهم على المحك.

عوامل تؤثر في الثبات:

طول الاختبار: كلما زاد عدد الفقرات المتجانسة في الاختبار، زاد ثباته (مع تساوي العوامل الأخرى). هذا هو أساس معادلة سبيرمان–براون.

تجانس الفقرات: الاختبار الذي تقيس فقراته نفس السمة أو المهارة بشكل متسق يكون أكثر ثباتًا (خاصة في طرق الاتساق الداخلي).

صعوبة الفقرات: الاختبارات التي تحتوي على فقرات متوسطة الصعوبة (p حوالي 0.5) تميل إلى أن تكون أكثر ثباتًا لأنها تزيد التباين في الدرجات. الفقرات السهلة جدًا أو الصعبة جدًا تقلل التباين وبالتالي الثبات.

مدى تباين درجات المجموعة: كلما زاد تباين (انتشار) درجات الأفراد في المجموعة التي حسب عليها الثبات، زاد معامل الثبات. يجب توخي الحذر عند مقارنة معاملات الثبات المحسوبة على مجموعات مختلفة في التجانس.

الزمن المخصص للاختبار: في اختبارات السرعة، يؤدي تقصير الوقت إلى تقليل عدد الفقرات التي يجيب عليها الطلاب، مما قد يقلل الثبات الظاهري إذا استخدمت طرق تعتمد على التباين. طرق الاتساق الداخلي غير مناسبة لهذه الاختبارات.

موضوعية التصحيح: كما ذكر سابقًا، ذاتية التصحيح تقلل من ثبات الاختبار (خاصة ثبات المصححين).

الفترة الزمنية بين التطبيقين (في إعادة الاختبار): فترة قصيرة جدًا قد تزيد الثبات بشكل مصطنع بسبب الذاكرة، وفترة طويلة جدًا قد تقلله بسبب التغير الحقيقي في السمة.

5.4. العلاقة بين الصدق والثبات (The Relationship Between Validity and Reliability)

الصدق والثبات مرتبطان ولكنهما مفهومان مختلفان. يمكن تلخيص العلاقة بينهما كالتالي:

الثبات شرط ضروري للصدق، ولكنه ليس كافيًا: لكي يكون الاختبار صادقًا (يقيس ما يفترض أن يقيسه)، يجب أن يكون ثابتًا (يقيسه باتساق). لا يمكن لاختبار غير ثابت (يعطي نتائج عشوائية) أن يكون صادقًا. إذا كانت الدرجات غير مستقرة، فلا يمكن أن تمثل السمة المقصودة بدقة أو ترتبط بمحكات أخرى بشكل مفيد.

الاختبار الثابت قد لا يكون صادقًا: يمكن أن يكون الاختبار ثابتًا جدًا (يعطي نفس النتائج باستمرار) ولكنه لا يقيس السمة المقصودة. مثال: ميزان يعطي دائمًا قراءة تزيد 5 كيلوجرامات عن الوزن الفعلي. هو ثابت (يعطي نفس القراءة الزائدة كل مرة)، ولكنه ليس صادقًا (لا يعطي الوزن الحقيقي). مثال آخر: اختبار يقيس مهارات حسابية بسيطة بثبات عالٍ، ولكنه

يستخدم للتنبؤ بالنجاح في الهندسة المعمارية (الذي يتطلب قدرات مكانية وتصميمية). قد يكون الاختبار ثابتًا، ولكنه ليس صادقًا لهذا الغرض التنبؤي.

إذًا، الثبات يضع سقفًا للصدق. بمعنى أن أقصى معامل صدق ممكن لاختبار ما لا يمكن أن يتجاوز الجذر التربيعي لمعامل ثباته (<r_{xx'}√). كلما زاد الثبات، زادت الإمكانية لتحقيق صدق أعلى، ولكن الصدق يعتمد أيضًا على عوامل أخرى كثيرة (مثل تمثيل المحتوى، الارتباط بالمحكات، البنية النظرية).

في عملية بناء الاختبار، يجب السعي لتحقيق كل من الثبات المرتفع والصدق القوي من خلال التخطيط الدقيق، والصياغة الجيدة للفقرات، والمراجعة، والتجريب، والتحليل، وجمع الأدلة المتنوعة.

6. إعداد الصورة النهائية للاختبار وتحديد المعايير (Preparing the Final Test Form and Establishing Norms)

بعد رحلة طويلة من التخطيط وتحديد الأهداف وتحليل المحتوى وبناء جدول المواصفات وصياغة الفقرات ومراجعتها وتجريبها وتحليلها وتقويم صدقها وثباتها، نصل إلى مرحلة تجميع الاختبار في صورته النهائية وإعداده للاستخدام الفعلي. تتضمن هذه المرحلة تنظيم الفقرات المختارة، وصياغة تعليمات واضحة، وتصميم شكل الإخراج، وإعداد مفتاح التصحيح. وبالنسبة للاختبارات المقننة محكية المعيار (Norm-Referenced)، تتضمن هذه المرحلة أيضًا خطوة حيوية أخرى وهي عملية التقنين (Standardization) وبناء المعايير (Norms) التي ستستخدم لتفسير درجات الأفراد ومقارنتها بأداء مجموعة مرجعية.

6.1. تجميع الفقرات المختارة وترتيبها (Assembling and Arranging Selected Items)

بناءً على نتائج تحليل الفقرات وتقويم الخصائص السيكومترية، يتم اختيار أفضل الفقرات التي تم الاحتفاظ بها أو تعديلها لتكوين النسخة النهائية للاختبار. يجب أن يتوافق عدد الفقرات وتوزيعها مع ما هو محدد في جدول المواصفات الأصلي قدر الإمكان لضمان الحفاظ على صدق المحتوى والتوازن.

عند تجميع هذه الفقرات، يجب مراعاة ترتيبها داخل الاختبار. هناك عدة طرق لترتيب الفقرات، ولكل منها مزايا وعيوب (Osterlind, 1998):

الترتيب حسب نوع الفقرة: تجميع كل نوع من الفقرات المتشابهة معًا (مثل كل فقرات الاختيار من متعدد معًا، ثم كل فقرات الصواب والخطأ، وهكذا). هذا يسهل على الطالب فهم التعليمات المطلوبة لكل نوع والانتقال بينها.

الترتيب حسب موضوع المحتوى: تجميع الفقرات التي تنتمي لنفس الموضوع أو الوحدة الدراسية معًا. هذا قد يساعد الطالب على التركيز على مجال محتوى معين في كل مرة.

الترتيب حسب مستوى الصعوبة: البدء بالفقرات الأسهل ثم التدرج نحو الفقرات الأكثر صعوبة. هذا يساعد على بناء ثقة الطالب وتقليل القلق في بداية الاختبار، ويضمن أن يحاول معظم الطلاب الإجابة على أكبر عدد ممكن من الفقرات التي يستطيعون حلها قبل نفاد الوقت. غالبًا ما يُفضل هذا الترتيب في الاختبارات التحصيلية.

الترتيب العشوائي: ترتيب الفقرات بشكل عشوائي داخل الاختبار أو داخل كل قسم.

الترتيب الأكثر شيوعًا وتوصية به للاختبارات التحصيلية هو الجمع بين الترتيب حسب نوع الفقرة والترتيب حسب الصعوبة داخل كل نوع. أي، يتم تجميع فقرات كل نوع معًا، وداخل كل مجموعة، يتم ترتيب الفقرات من الأسهل إلى الأصعب بناءً على معاملات الصعوبة التي تم الحصول عليها في مرحلة التجريب.

6.2. صياغة تعليمات الاختبار (Writing Test Instructions)

تُعد تعليمات الاختبار الواضحة والموجزة جزءًا لا يتجزأ من الاختبار المقنن. يجب أن توفر التعليمات للمفحوصين جميع المعلومات اللازمة لأداء الاختبار بشكل صحيح وموحد. يجب أن تكون التعليمات مكتوبة بلغة بسيطة ومناسبة للمستوى العمري للطلاب، ويجب أن تغطي الجوانب التالية (AERA, APA, & NCME, 2014):

الغرض العام من الاختبار: بشكل موجز ومناسب.

الوقت الكلي المخصص للاختبار: أو الوقت المخصص لكل قسم (إن وجد).

عدد الأقسام (إن وجدت) وعدد الفقرات في كل قسم.

كيفية تسجيل الإجابات: هل تكون على نفس ورقة الأسئلة أم في ورقة إجابة منفصلة؟ هل تستخدم قلم رصاص أم حبر؟ كيفية تظليل الدوائر (إذا كانت الإجابة آلية).

تعليمات محددة لكل نوع من أنواع الفقرات: شرح واضح للمطلوب في كل قسم (مثل: "اختر أفضل إجابة واحدة"، "ضع علامة صح أو خطأ"، "صل العمود أ بما يناسبه من العمود ب").

معلومات عن كيفية حساب الدرجة: هل هناك درجات مختلفة للفقرات؟ هل هناك خصم للإجابة الخاطئة (معادلة التخمين)؟ إذا كان الأمر كذلك، يجب توضيح سياسة التعامل مع التخمين بوضوح (هل ينصح الطالب بتخمين الإجابات التي لا يعرفها أم بتركها؟).

مثال توضيحي (إن لزم الأمر): خاصة للأنواع غير المألوفة من الفقرات أو لطرق التسجيل المعقدة.

تذكير بمراجعة الإجابات قبل تسليم الورقة (إذا سمح الوقت).

يجب أن تكون التعليمات موحدة تمامًا لجميع المفحوصين، وغالبًا ما يقوم المشرف على الاختبار بقراءتها بصوت عالٍ قبل البدء للتأكد من فهم الجميع.

6.3. تصميم صفحة الغلاف وشكل الإخراج (Designing the Cover Page and Layout)

يجب أن يكون شكل إخراج الاختبار وتصميمه جذابًا وواضحًا وسهل القراءة. يتضمن ذلك:

صفحة الغلاف: يجب أن تحتوي على معلومات أساسية مثل:

اسم الاختبار بوضوح.

اسم المؤسسة التي أعدت الاختبار.

المستوى الدراسي أو الجمهور المستهدف.

المساحة المخصصة لكتابة اسم الطالب ومعلومات التعريف الأخرى المطلوبة.

تعليمات عامة موجزة (مثل عدم فتح الورقة قبل الإذن، والوقت المخصص).

تاريخ التطبيق (إن لزم الأمر).

التنسيق الداخلي:

استخدام حجم خط ونوع خط واضح ومقروء.

ترك هوامش ومسافات كافية بين الفقرات والأسطر لتجنب الازدحام.

ترقيم الفقرات بشكل واضح ومتسلسل.

وضع كل فقرة اختيار من متعدد مع بدائلها كوحدة واحدة (تجنب فصل المقدمة عن البدائل بين صفحتين).

محاذاة البدائل رأسيًا (أ، ب، ج، د) لتسهيل القراءة.

وضوح الرسوم البيانية أو الصور أو الجداول المستخدمة وجودتها.

جودة الطباعة والورق.

شكل الإخراج الجيد يساهم في الصدق الظاهري للاختبار ويقلل من الأخطاء الناتجة عن صعوبة القراءة أو الإرباك.

6.4. إعداد مفتاح التصحيح وقواعده (Preparing the Scoring Key and Rules)

يجب إعداد مفتاح تصحيح دقيق (Scoring Key) يتضمن الإجابات الصحيحة لجميع الفقرات الموضوعية. وبالنسبة للفقرات المقالية، يجب إعداد قواعد تصحيح مفصلة أو معايير تقييم (Scoring Rubrics) تحدد مستويات الأداء المختلفة والدرجات المقابلة لكل مستوى، مع أمثلة توضيحية إن أمكن.

يجب أن تكون قواعد التصحيح واضحة وموضوعية قدر الإمكان لضمان ثبات التصحيح، سواء قام به نفس المصحح في أوقات مختلفة أو قام به مصححون مختلفون. يجب تدريب المصححين على استخدام هذه القواعد قبل البدء بالتصحيح الفعلي، وإجراء فحوصات دورية لضمان الاتساق بينهم.

6.5. التقنين وتحديد المعايير (Standardization and Norming)

هذه الخطوة خاصة بالاختبارات المقننة التي تهدف إلى مقارنة أداء الفرد بأداء مجموعة مرجعية (Norm-Referenced Tests).

6.5.1. مفهوم التقنين (Concept of Standardization):

التقنين هو عملية تطبيق الاختبار على عينة كبيرة وممثلة للمجتمع الذي صُمم الاختبار له، وذلك تحت ظروف موحدة تمامًا (نفس التعليمات، نفس الوقت، نفس طريقة التطبيق والتصحيح). الهدف من التقنين هو الحصول على بيانات أداء هذه

العينة المرجعية لاستخدامها في بناء المعايير.

6.5.2. اختيار عينة التقنين (Selecting the Norming Sample):
تُعد جودة عينة التقنين أمرًا حاسمًا لجودة المعايير وصحة تفسير الدرجات. يجب أن تكون هذه العينة:

كبيرة الحجم: لضمان استقرار المعايير وتقليل الخطأ المعياري. قد يتراوح حجمها من مئات إلى آلاف الأفراد حسب حجم المجتمع الأصلي ومدى تنوعه.

ممثلة للمجتمع المستهدف: يجب أن تعكس العينة بدقة خصائص المجتمع الأصلي من حيث المتغيرات الديموغرافية الهامة (مثل العمر، الجنس، المستوى التعليمي، المنطقة الجغرافية، الوضع الاجتماعي والاقتصادي، العرق/الإثنية، إلخ). غالبًا ما تستخدم طرق المعاينة الطبقية العشوائية (Stratified Random Sampling) لضمان تمثيل المجموعات الفرعية المختلفة بنسب تتوافق مع وجودها في المجتمع.

حديثة: يجب تحديث عينات التقنين والمعايير بشكل دوري (كل بضع سنوات) لتعكس التغيرات التي قد تطرأ على المجتمع أو على مستويات الأداء (مثل تأثير التغيرات في المناهج أو ظاهرة فلين Flynn Effect في اختبارات الذكاء).

6.5.3. أنواع المعايير (Types of Norms):
المعايير هي مجموعة من الدرجات المشتقة من أداء عينة التقنين، وتستخدم كنقطة مرجعية لتفسير معنى الدرجة الخام (Raw Score) التي يحصل عليها الفرد في الاختبار. الدرجة الخام بمفردها (مثل عدد الإجابات الصحيحة) غالبًا ما تكون عديمة المعنى ما لم تتم مقارنتها بأداء الآخرين. أشهر أنواع المعايير المستخدمة في الاختبارات التحصيلية هي:

المئينيات (Percentiles / Percentile Ranks - PR): 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 خام 45 هي 70 (PR=70)، فهذا يعني أن الفرد الذي حصل على 45 درجة تفوق على 70% من أفراد عينة التقنين. المئينيات سهلة الفهم والتفسير، لكنها لا تعكس المسافات المتساوية بين الدرجات (فالفرق بين المئين 90 و 95 قد يمثل فرقًا أكبر في الدرجات الخام من الفرق بين المئين 50 و 55).

الدرجات المعيارية (Standard Scores): هي درجات مشتقة تحول الدرجات الخام إلى مقياس جديد له متوسط حسابي وانحراف معياري محدد مسبقًا. تتميز بأنها تحافظ على المسافات المتساوية بين الدرجات. أشهر أنواعها:

الدرجة الزائية (Z-score): أبسط الدرجات المعيارية. تعبر عن بعد الدرجة الخام عن المتوسط الحسابي لعينة التقنين بوحدات الانحراف المعياري. Z = (X - M) / SD، حيث X هي الدرجة الخام، M هو المتوسط، و SD هو الانحراف المعياري. متوسط الدرجات الزائية دائمًا 0 وانحرافها المعياري 1. قد تكون سالبة أو تحتوي على كسور عشرية.

الدرجة التائية (T-score): تحويل خطي للدرجة الزائية لتجنب القيم السالبة والكسور. T = (Z * 10) + 50. متوسط الدرجات التائية 50 وانحرافها المعياري 10.

درجات أخرى مشتقة: مثل درجات اختبارات الذكاء (المتوسط 100 والانحراف المعياري 15)، أو درجات CEEB (المستخدمة في اختبار SAT، المتوسط 500 والانحراف المعياري 100).

المكافئات العمرية والصفية (Age and Grade Equivalents): تستخدم أحيانًا في الاختبارات التحصيلية للأطفال. المكافئ العمري لدرجة خام معينة هو متوسط العمر الزمني للأفراد في عينة التقنين الذين حصلوا على هذه الدرجة. المكافئ الصفي

هو متوسط الصف الدراسي للأفراد الذين حصلوا على هذه الدرجة. على الرغم من سهولة فهمها ظاهريًا، إلا أن هذه المعايير تعاني من مشكلات سيكومترية عديدة (مثل عدم تساوي الوحدات، والمبالغة في تفسير الفروق الصغيرة، وعدم ملاءمتها للمستويات العليا والدنيا)، ويوصى باستخدامها بحذر شديد أو تجنبها لصالح المئينيات أو الدرجات المعيارية (AERA, APA, & NCME, 2014).

6.5.4. إعداد جداول المعايير (Developing Norm Tables):

بعد حساب المعايير المختلفة (مئينيات، درجات معيارية) لكل درجة خام ممكنة بناءً على بيانات عينة التقنين، يتم تنظيم هذه المعلومات في جداول واضحة في دليل الاختبار (Test Manual). تتيح هذه الجداول لمستخدم الاختبار تحويل الدرجة الخام التي يحصل عليها أي فرد إلى الدرجة المعيارية المقابلة لها، مما يسهل تفسير أدائه ومقارنته بأداء أقرانه في عينة التقنين. قد يتم توفير جداول معايير منفصلة لمجموعات فرعية مختلفة (مثل فئات عمرية أو صفية مختلفة) إذا كانت هناك فروق ذات دلالة إحصائية في الأداء بين هذه المجموعات.

إن إعداد الصورة النهائية للاختبار وتحديد المعايير يمثل تتويجًا لعملية البناء الطويلة والمعقدة، ويجعل الاختبار أداة قابلة للاستخدام والتفسير بشكل موحد وموضوعي.

7. تطبيق الاختبار وتفسير النتائج (Test Administration and Interpretation of Results)

لا تكتمل عملية بناء الاختبار التحصيلي المقنن بمجرد إعداد صورته النهائية ومعاييره، بل تمتد لتشمل كيفية تطبيقه بشكل موحد وصحيح، وكيفية تصحيح الإجابات بدقة، والأهم من ذلك، كيفية تفسير النتائج التي يتم الحصول عليها بطريقة سليمة ومفيدة لاتخاذ القرارات التربوية. إن أي خلل في هذه المراحل النهائية قد يقلل من قيمة كل الجهود التي بذلت في بناء الاختبار.

7.1. إجراءات التطبيق الموحدة (Standardized Administration Procedures)

لضمان أن تكون درجات الاختبار قابلة للمقارنة بين الأفراد المختلفين أو المجموعات المختلفة أو عبر الزمن، يجب أن يتم تطبيق الاختبار في ظروف موحدة تمامًا لجميع المفحوصين. هذا هو جوهر التقنين في مرحلة التطبيق. يجب أن يتضمن دليل الاختبار (Test Manual) وصفًا تفصيليًا ودقيقًا لإجراءات التطبيق التي يجب على المشرفين أو المطبقين الالتزام بها حرفيًا. تشمل هذه الإجراءات:

التعليمات التي يجب قراءتها للمفحوصين: يجب قراءة التعليمات كلمة بكلمة كما هي مكتوبة في الدليل، دون زيادة أو نقصان أو إعادة صياغة.

الإجابة عن استفسارات المفحوصين: يجب تحديد نوعية الاستفسارات التي يمكن الإجابة عليها (غالبًا ما تقتصر على توضيح التعليمات فقط) وتلك التي لا يمكن الإجابة عليها (مثل شرح معنى كلمة في الفقرة أو تقديم تلميح).

تحديد الوقت المسموح به بدقة: يجب الالتزام بالوقت المحدد للاختبار ككل أو لكل قسم، واستخدام ساعة توقيت دقيقة. يجب إخبار المفحوصين بالوقت المتبقي في فترات محددة (إذا نصت التعليمات على ذلك).

توزيع مواد الاختبار وجمعها: تحديد الطريقة الموحدة لتوزيع كراسات الأسئلة وأوراق الإجابة ومتى يتم ذلك، وكيفية جمعها والتأكد من استلام جميع المواد في نهاية الوقت.

المراقبة أثناء الاختبار: كيفية التعامل مع حالات الغش أو الإزعاج، وكيفية التحرك في قاعة الاختبار لضمان عدم تشتيت انتباه المفحوصين.

المواد المسموح بها أو الممنوعة: تحديد ما إذا كان مسموحًا باستخدام الآلات الحاسبة، أو القواميس، أو أوراق خارجية، وما هي المواد الممنوع إدخالها إلى قاعة الاختبار.

إن أي انحراف عن هذه الإجراءات الموحدة قد يؤدي إلى إدخال خطأ في القياس ويجعل مقارنة الدرجات غير عادلة أو غير دقيقة. لذلك، يجب تدريب مطبقي الاختبار (Test Administrators) جيدًا على هذه الإجراءات والتأكد من التزامهم بها.

7.2. تهيئة بيئة الاختبار (Preparing the Testing Environment)

تلعب بيئة الاختبار دورًا هامًا في أداء المفحوصين. يجب أن تكون البيئة مريحة ومناسبة وتساعد على التركيز قدر الإمكان. تشمل شروط البيئة المناسبة:

مكان هادئ: خالٍ من الضوضاء والمقاطعات الخارجية.

إضاءة جيدة: كافية ومريحة للقراءة دون إجهاد.

تهوية مناسبة ودرجة حرارة معتدلة.

مقاعد مريحة ومساحة كافية: لضمان راحة المفحوصين ومنع الغش.

توفير جميع المواد اللازمة: أقلام رصاص، ممحاة، أوراق إجابة، آلات حاسبة (إذا كانت مسموحة)، ساعة واضحة لرؤية الوقت.

تقليل عوامل التشتيت: إغلاق الهواتف المحمولة، منع الأحاديث الجانبية.

يجب على مطبق الاختبار التأكد من توفر هذه الشروط قبل بدء الاختبار.

7.3. دور المطبق والمراقب (Role of the Test Administrator and Proctor)

يقع على عاتق مطبق الاختبار (أو المراقب Proctor) مسؤولية كبيرة في ضمان سير عملية الاختبار بسلاسة ونظام ووفقًا للإجراءات الموحدة. يجب أن يكون المطبق:

مدربًا جيدًا: على دراية تامة بتعليمات الاختبار وإجراءات تطبيقه.

محايدًا وموضوعيًا: يتعامل مع جميع المفحوصين بنفس الطريقة دون محاباة أو تحيز.

واضحًا في تواصله: عند قراءة التعليمات أو الإجابة على الاستفسارات المسموح بها.

يقظًا وحازمًا: في مراقبة القاعة ومنع أي محاولات للغش أو الإخلال بالنظام، ولكن دون إثارة قلق أو توتر لا داعي له.

منظمًا: في توزيع المواد وجمعها والالتزام بالوقت.

قادرًا على التعامل مع الحالات الطارئة: مثل مرض أحد المفحوصين أو انقطاع الكهرباء، وفقًا للإرشادات المعدة مسبقًا.

إن سلوك المطبق وتعامله مع المفحوصين يمكن أن يؤثر على دافعيتهم وأدائهم، وبالتالي على نتائج الاختبار.

7.4. تصحيح الاختبار (Scoring the Test)

بعد الانتهاء من تطبيق الاختبار وجمع أوراق الإجابة، تأتي مرحلة التصحيح. يجب أن تتم عملية التصحيح بدقة وموضوعية لضمان عدم إضافة خطأ جديد في هذه المرحلة.

للفقرات الموضوعية: غالبًا ما يتم التصحيح باستخدام مفتاح مثقوب (Scoring Stencil) يوضع فوق ورقة الإجابة، أو باستخدام التصحيح الآلي (Machine Scoring) عن طريق الماسح الضوئي (Optical Scanner) وبرامج الكمبيوتر المخصصة. التصحيح الآلي هو الأكثر دقة وسرعة وموضوعية، ولكنه يتطلب استخدام أوراق إجابة خاصة وتظليل دقيق من قبل المفحوصين. يجب التأكد من دقة برمجة مفتاح التصحيح في الكمبيوتر ومن معايرة جهاز المسح بشكل دوري.

للفقرات المقالية (أو بنود الأداء): كما ذكر سابقًا، يتطلب التصحيح استخدام معايير تقييم مفصلة (Rubrics) وتدريب المصححين لضمان الاتساق (ثبات المصححين). يفضل أن يقوم أكثر من مصحح واحد بتصحيح نفس الإجابة (خاصة في الاختبارات عالية المخاطر) وحساب متوسط الدرجات أو حل الخلافات بينهم. يجب أيضًا اتخاذ إجراءات لتقليل التحيز، مثل التصحيح دون معرفة اسم الطالب، وتصحيح جميع إجابات سؤال واحد قبل الانتقال للسؤال التالي.

بعد الحصول على الدرجة الخام (Raw Score) لكل مفحوص (غالبًا ما تكون مجموع الدرجات على الفقرات الصحيحة، مع أو بدون تطبيق معادلة التخمين)، يتم تحويل هذه الدرجة الخام إلى درجة معيارية باستخدام جداول المعايير الموجودة في دليل الاختبار.

7.5. تفسير الدرجات في ضوء المعايير (Interpreting Scores Using Norms)

الخطوة الأهم هي فهم معنى الدرجات التي حصل عليها الأفراد. في الاختبارات المقننة محكية المعيار، يتم تفسير الدرجة من خلال مقارنتها بأداء عينة التقنين، وذلك باستخدام المعايير (المئينيات أو الدرجات المعيارية) التي تم اشتقاقها.

تفسير المئينيات: يشير المئين إلى الموقع النسبي للفرد مقارنة بأقرانه. مئين 80 يعني أن أداء الفرد أفضل من 80% من أفراد عينة التقنين.

تفسير الدرجات المعيارية: تشير الدرجة المعيارية (مثل Z أو T) إلى مدى انحراف أداء الفرد عن متوسط أداء عينة التقنين بوحدات الانحراف المعياري. درجة T=60 (التي تقابل Z=+1) تعني أن أداء الفرد يزيد بمقدار انحراف معياري واحد فوق المتوسط.

نقاط هامة عند تفسير الدرجات:

الخطأ المعياري للقياس (Standard Error of Measurement - SEM): لا توجد درجة اختبار خالية تمامًا من الخطأ. الخطأ المعياري للقياس هو مقياس لمدى الدقة أو الثبات في درجة الاختبار، ويعكس مقدار التذبذب المتوقع في درجة الفرد لو أُعيد اختباره مرات عديدة. يُحسب SEM باستخدام معامل ثبات الاختبار والانحراف المعياري للدرجات: SEM = SD * √(1 - r_{xx'}). يُستخدم SEM لإنشاء "نطاق ثقة" (Confidence Interval) حول الدرجة الملاحظة للفرد،

والذي يمثل النطاق الذي تقع فيه الدرجة الحقيقية للفرد بدرجة معينة من الاحتمالية (مثل 95%). على سبيل المثال، إذا كانت درجة الطالب 55 وكان SEM = 3، فإن نطاق الثقة 95% يكون تقريبًا 55 ± (1.96 * 3)، أي بين 49 و 61. هذا التفسير أكثر واقعية من التعامل مع الدرجة الملاحظة كنقطة دقيقة تمامًا. يجب دائمًا أخذ SEM في الاعتبار عند مقارنة درجات الأفراد أو اتخاذ قرارات بناءً عليها (AERA, APA, & NCME, 2014).

ملاءمة المعايير: يجب التأكد من أن المعايير المستخدمة للتفسير مناسبة للفرد الذي يتم تفسير درجته (من حيث العمر، الصف الدراسي، الخلفية، إلخ). استخدام معايير غير مناسبة يؤدي إلى تفسير خاطئ.

عدم الاعتماد على درجة اختبار واحد فقط: يجب عدم اتخاذ قرارات هامة بناءً على درجة اختبار واحد فقط. يجب دائمًا دمج معلومات الاختبار مع مصادر أخرى للمعلومات (مثل ملاحظات المعلم، الأداء الصفي، اختبارات أخرى، مقابلات) للحصول على صورة شاملة عن الفرد.

فهم حدود الاختبار: يجب أن يكون المفسر على دراية بما يقيسه الاختبار وما لا يقيسه، والغرض الذي صُمم من أجله، وحدود صدقه وثباته.

7.6. كتابة التقارير (Reporting Results)

يجب توصيل نتائج الاختبار إلى المعنيين (الطلاب، أولياء الأمور، المعلمين، الإداريين) بطريقة واضحة ومفهومة ومفيدة. يجب أن تتضمن التقارير عادةً:

معلومات تعريفية: اسم الطالب، اسم الاختبار، تاريخ التطبيق.

الدرجات: الدرجة الخام (إن كانت مفيدة)، والدرجات المعيارية (مئينيات، درجات T، إلخ) لكل مجال تم قياسه.

تفسير مبسط للمعايير: شرح لمعنى المئينيات أو الدرجات المعيارية بلغة سهلة.

عرض نطاق الثقة (SEM): للتأكيد على عدم دقة الدرجة المطلقة.

وصف لنقاط القوة والضعف النسبية: بناءً على الأداء في المجالات المختلفة للاختبار.

توصيات (إن أمكن): اقتراحات عملية بناءً على النتائج، مع التأكيد على ضرورة دمجها مع معلومات أخرى.

معلومات حول الاختبار نفسه: وصف موجز لما يقيسه الاختبار وخصائصه السيكومترية الأساسية (صدق، ثبات).

يجب أن تكون التقارير مصممة لتكون مفيدة في اتخاذ القرارات التعليمية وتحسين تعلم الطلاب، مع تجنب استخدام لغة فنية مفرطة أو إصدار أحكام قاطعة بناءً على الاختبار وحده.

إن التطبيق الدقيق والتفسير الحذر للنتائج هما خطوتان أساسيتان لضمان الاستفادة القصوى من الاختبارات التحصيلية المقننة وتحقيق الأهداف التي بنيت من أجلها.

8. الاعتبارات الأخلاقية والقانونية في بناء واستخدام الاختبارات (Ethical and Legal

(Considerations in Test Construction and Use

لا تقتصر عملية بناء واستخدام الاختبارات التحصيلية المقننة على الجوانب الفنية والسيكومترية فقط، بل تمتد لتشمل مجموعة هامة من الاعتبارات الأخلاقية والقانونية التي يجب على جميع المعنيين بهذه العملية ـ من مطوري الاختبارات إلى مستخدميها والمفحوصين أنفسهم ـ الالتزام بها. تهدف هذه الاعتبارات إلى ضمان استخدام الاختبارات بطريقة عادلة ومسؤولة، وحماية حقوق جميع الأطراف، وتحقيق الأهداف التربوية المنشودة دون إلحاق ضرر غير مبرر بالأفراد أو المجموعات (AERA, APA, & NCME, 2014; APA, 2017; Code of Fair Testing Practices in Education, 2004).

8.1. مسؤوليات مطور الاختبار (Responsibilities of the Test Developer)

يقع على عاتق الجهة أو الأفراد الذين يقومون بتطوير الاختبار المقنن مسؤوليات أخلاقية وقانونية كبيرة، تشمل:

بناء اختبار عالي الجودة: الالتزام بالمعايير المهنية والفنية في جميع مراحل البناء (التخطيط، الصياغة، التحليل، التقنين، التحقق من الصدق والثبات) لإنتاج أداة قياس دقيقة وموثوقة قدر الإمكان.

تحديد الغرض والاستخدام المناسب للاختبار بوضوح: توضيح ما يقيسه الاختبار وما لا يقيسه، والأغراض التي صُمم من أجلها، والمجتمع المستهدف، والتحذير من الاستخدامات غير المناسبة أو التفسيرات الخاطئة المحتملة.

توفير معلومات شاملة في دليل الاختبار: يجب أن يتضمن دليل الاختبار (Test Manual) معلومات مفصلة حول عملية تطوير الاختبار، وخصائصه السيكومترية (أدلة الصدق، معاملات الثبات، الخطأ المعياري للقياس)، وعينة التقنين وخصائصها، وإجراءات التطبيق والتصحيح الموحدة، وجداول المعايير، وإرشادات التفسير، والتحذيرات المتعلقة بالاستخدام.

ضمان عدالة الاختبار: اتخاذ خطوات لتقليل التحيز غير المقصود في محتوى الفقرات أو صياغتها ضد مجموعات فرعية معينة (على أساس العرق، الجنس، الخلفية الثقافية، اللغة، الإعاقة، إلخ). يتضمن ذلك مراجعة الفقرات للكشف عن التحيز (Fairness Review) وإجراء تحليلات إحصائية للأداء التفاضلي للفقرة (DIF).

تحديث الاختبار والمعايير بانتظام: مراجعة الاختبار ومعاييره بشكل دوري للتأكد من أنها لا تزال ملائمة وحديثة وتعكس التغيرات في المناهج أو في أداء المجتمع المستهدف.

تأمين الاختبار: اتخاذ إجراءات لحماية محتوى الاختبار من التسريب أو الوصول غير المصرح به للحفاظ على صدقه وفائدته.

تحديد المؤهلات المطلوبة لمستخدمي الاختبار: قد تتطلب بعض الاختبارات (خاصة تلك المستخدمة للتشخيص أو القرارات عالية المخاطر) مؤهلات أو تدريبًا معينًا لدى المستخدمين لضمان تطبيقها وتفسيرها بشكل صحيح.

8.2. مسؤوليات مستخدم الاختبار (Responsibilities of the Test User)

لا تقل مسؤوليات مستخدمي الاختبار (مثل المعلمين، المرشدين، الإداريين، أخصائيي القياس) أهمية عن مسؤوليات المطورين. فسوء استخدام اختبار جيد قد يؤدي إلى نتائج ضارة. تشمل مسؤوليات المستخدمين:

اختيار الاختبار المناسب: التأكد من أن الاختبار المختار ملائم للغرض المحدد وللجمهور المستهدف، وأن خصائصه

السيكومترية (الصدق والثبات) مقبولة لهذا الغرض. يتطلب ذلك قراءة دليل الاختبار بعناية.

امتلاك الكفاءة اللازمة: التأكد من أن لديهم المعرفة والمهارات والتدريب الكافي لتطبيق الاختبار وتصحيحه وتفسير نتائجه بشكل صحيح. وإذا لم تتوفر لديهم الكفاءة، يجب عليهم طلب المساعدة من متخصصين مؤهلين أو الامتناع عن استخدام الاختبار.

الالتزام بإجراءات التطبيق والتصحيح الموحدة: تطبيق الاختبار وتصحيحه بدقة وفقًا للتعليمات الواردة في الدليل لضمان الحصول على نتائج قابلة للمقارنة.

التفسير الصحيح والحذر للنتائج: تفسير الدرجات في ضوء المعايير المناسبة، مع الأخذ في الاعتبار الخطأ المعياري للقياس، وعدم المبالغة في تعميم النتائج أو استخدامها لاتخاذ قرارات تتجاوز ما يسمح به صدق الاختبار.

عدم الاعتماد على نتائج الاختبار وحدها: دمج نتائج الاختبار مع مصادر أخرى للمعلومات عند اتخاذ قرارات هامة حول الأفراد.

الحفاظ على سرية النتائج: حماية خصوصية المفحوصين وعدم الكشف عن نتائجهم إلا للأشخاص المصرح لهم بذلك (مثل الطالب نفسه، أولياء الأمور، المعنيين التربويين) وللأغراض المشروعة.

توصيل النتائج بفعالية: شرح النتائج للمفحوصين أو أولياء أمورهم بطريقة واضحة ومفهومة، مع التركيز على الاستخدام البناء للمعلومات.

الحفاظ على أمن مواد الاختبار: عدم نسخ الاختبار أو كشف محتواه أو تدريب الطلاب على فقراته بشكل مباشر للحفاظ على صلاحيته.

8.3. حقوق المفحوصين (Rights of Test Takers)

للأفراد الذين يخضعون للاختبارات حقوق أساسية يجب احترامها وحمايتها، منها:

الحق في المعاملة العادلة والمنصفة: يجب ألا يتعرض المفحوصون للتمييز على أي أساس غير مبرر، وأن تتاح لهم فرص متكافئة لإظهار قدراتهم.

الحق في اختبار ملائم وموثوق: يجب أن يكون الاختبار المستخدم ذا جودة فنية مقبولة ويقيس ما يفترض أن يقيسه بدقة معقولة.

الحق في المعرفة بالغرض من الاختبار وكيفية استخدام نتائجه: يجب إعلام المفحوصين (أو أولياء أمورهم) لماذا يتم اختبارهم وكيف ستستخدم النتائج ومن سيطلع عليها.

الحق في الحصول على تفسير للنتائج: يجب أن تتاح للمفحوصين فرصة لمعرفة نتائجهم وفهم معناها، ويفضل أن يكون ذلك بلغة واضحة ومفهومة.

الحق في الخصوصية والسرية: يجب الحفاظ على سرية نتائج الاختبار وعدم الكشف عنها لأطراف غير مصرح لها دون موافقة المفحوص (أو ولي أمره).

الحق في ظروف تطبيق مناسبة: يجب أن يتم تطبيق الاختبار في بيئة مريحة وخالية من المشتتات قدر الإمكان.

الحق في طلب إعادة النظر (في بعض الحالات): قد يكون للمفحوصين الحق في طلب مراجعة عملية التصحيح أو إعادة الاختبار إذا كان هناك دليل على وجود خطأ أو ظروف غير عادلة أثرت على أدائهم (تعتمد الإجراءات على سياسات الجهة المنظمة للاختبار).

الحق في الحصول على تسهيلات معقولة (للأشخاص ذوي الإعاقة): يجب توفير ترتيبات خاصة (Accommodations) للأشخاص ذوي الإعاقة (مثل تمديد الوقت، أو توفير صيغ مكبرة أو مسموعة للاختبار، أو استخدام قارئ أو كاتب) لضمان أن الاختبار يقيس قدراتهم الحقيقية وليس إعاقتهم، ما لم تكن الإعاقة نفسها هي ما يقاس. يجب أن تكون هذه التسهيلات مبررة وموثقة ولا تغير من طبيعة السمة المقاسة.

8.4. التحيز في الاختبارات (Test Bias)

التحيز في الاختبار هو مشكلة فنية وأخلاقية خطيرة تحدث عندما يؤدي الاختبار إلى نتائج غير عادلة بشكل منهجي لمجموعة فرعية معينة من المفحوصين مقارنة بمجموعة أخرى، على الرغم من أن المجموعتين متساويتان في القدرة الحقيقية التي يقيسها الاختبار. يمكن أن ينشأ التحيز من مصادر مختلفة:

تحيز المحتوى (Content Bias): احتواء الفقرات على معلومات أو سياقات أو صور نمطية تكون مألوفة لمجموعة أكثر من أخرى دون أن تكون ضرورية لقياس السمة المقصودة.

تحيز الصياغة (Wording Bias): استخدام لغة أو مفردات تكون مفهومة لمجموعة أكثر من أخرى.

تحيز البنية الداخلية: إذا كانت البنية العاملية للاختبار تختلف بين المجموعات، فقد يعني ذلك أن الاختبار يقيس أشياء مختلفة لهم.

التحيز التنبؤي (Predictive Bias): إذا كان الاختبار يتنبأ بأداء مستقبلي (مثل النجاح الجامعي) بشكل مختلف للمجموعات المختلفة (أي أن خط الانحدار للعلاقة بين الاختبار والمحك يختلف بين المجموعات).

يجب على مطوري الاختبارات اتخاذ خطوات جادة للكشف عن التحيز المحتمل وتقليله من خلال المراجعة المتأنية للفقرات من قبل ممثلين عن المجموعات المختلفة، واستخدام التحليلات الإحصائية مثل تحليل الأداء التفاضلي للفقرة (DIF)، الذي يحدد الفقرات التي يجيب عليها أفراد من مجموعتين (متساويتين في القدرة الكلية) بشكل مختلف. الهدف هو بناء اختبار "عادل" (Fair) قدر الإمكان، بمعنى أنه يوفر لجميع المفحوصين فرصة متساوية لإظهار معرفتهم ومهاراتهم بغض النظر عن خلفياتهم غير المرتبطة بالسمة المقاسة.

8.5. أمن الاختبار (Test Security)

الحفاظ على سرية محتوى الاختبار قبل وأثناء وبعد التطبيق أمر ضروري للحفاظ على صدقه وفائدته. إذا تسربت فقرات الاختبار أو الإجابات الصحيحة، فإن الدرجات لن تعكس القدرة الحقيقية للأفراد. تشمل إجراءات الأمن:

تخزين مواد الاختبار (الكراسات، مفاتيح التصحيح، المعايير) في مكان آمن.

تقييد الوصول إلى مواد الاختبار للأشخاص المصرح لهم فقط.

اتخاذ إجراءات لمنع الغش أثناء التطبيق (مراقبة جيدة، ترتيب جلوس مناسب).

عدم استخدام نفس الصورة من الاختبار بشكل متكرر دون تغيير (خاصة في الاختبارات عالية المخاطر).

وضع سياسات واضحة للتعامل مع حالات انتهاك أمن الاختبار.

استخدام صور متكافئة للاختبار إذا كان سيطبق في أوقات مختلفة أو أماكن مختلفة.

8.6. المعايير المهنية والأخلاقية (Professional and Ethical Standards)

هناك العديد من المنظمات المهنية التي أصدرت معايير ومبادئ توجيهية لتطوير واستخدام الاختبارات التربوية والنفسية. من أبرز هذه المعايير:

معايير الاختبارات التربوية والنفسية (Standards for Educational and Psychological Testing): التي تصدرها بالاشتراك الجمعية الأمريكية للبحوث التربوية (AERA)، والجمعية الأمريكية لعلم النفس (APA)، والمجلس الوطني للقياس في التعليم (NCME). تعتبر هذه المعايير المرجع الأساسي للممارسات السليمة في مجال الاختبارات (AERA, APA, & NCME, 2014).

المبادئ الأخلاقية لعلماء النفس ومدونة قواعد السلوك (Ethical Principles of Psychologists and Code of Conduct): الصادرة عن الجمعية الأمريكية لعلم النفس (APA)، وتتضمن قسمًا خاصًا بالتقييم (APA, 2017).

مدونة ممارسات الاختبار العادل في التعليم (Code of Fair Testing Practices in Education): التي أعدها اللجنة المشتركة للاختبارات (Joint Committee on Testing Practices - JCTP)، وتقدم إرشادات لمطوري الاختبارات ومستخدميها لضمان العدالة (JCTP, 2004).

يجب على جميع العاملين في مجال القياس والتقويم التربوي أن يكونوا على دراية بهذه المعايير وأن يلتزموا بها في ممارساتهم المهنية لضمان استخدام الاختبارات بطريقة مسؤولة وأخلاقية تخدم مصلحة التعليم والمجتمع.

إن الوعي والالتزام بالاعتبارات الأخلاقية والقانونية ليسا مجرد إضافة شكلية لعملية بناء الاختبارات، بل هما جزء لا يتجزأ من الممارسة المهنية المسؤولة التي تضمن أن تكون هذه الأدوات القوية في خدمة الأهداف التربوية النبيلة وتحقيق العدالة لجميع المتعلمين.

خاتمة

لقد استعرضنا في هذا الفصل الرحلة التفصيلية والمعقدة لبناء الاختبارات التحصيلية المقننة، بدءًا من الشرارة الأولى لتحديد الغرض والمحتوى، ومرورًا بالمراحل الدقيقة لصياغة الفقرات وتجريبها وتحليلها، وصولًا إلى تقويم الخصائص السيكومترية الأساسية من صدق وثبات، وانتهاءً بإعداد الصورة النهائية وتحديد المعايير ووضع الإجراءات الموحدة للتطبيق والتصحيح والتفسير. وكما اتضح، فإن كل خطوة في هذه العملية تتطلب تخطيطًا منهجيًا، ومعرفة علمية، ومهارة فنية، والتزامًا بمعايير الجودة والدقة.

إن بناء اختبار تحصيلي مقنن ليس مجرد تجميع عشوائي للأسئلة، بل هو عملية علمية تهدف إلى إنتاج أداة قياس قادرة على توفير معلومات موثوقة وصادقة حول ما تعلمه الطلاب وما أتقنوه من مهارات. يتطلب الأمر بناء جسر متين بين محتوى

المنهج وأهدافه التعليمية وبين الفقرات التي تمثل هذا المحتوى وتقيس تلك الأهداف، وهذا الجسر هو جدول المواصفات. ويتطلب الأمر صياغة فنية دقيقة للفقرات، تتجنب الغموض والتلميحات والتحيز، وتناسب طبيعة المحتوى والمستوى المعرفي المستهدف.

ولا يمكن الاكتفاء بالصياغة والمراجعة النظرية، بل لا بد من النزول إلى الميدان عبر التجريب الاستطلاعي، واستinterrogating البيانات الناتجة من خلال تحليل الفقرات للكشف عن مواطن القوة والضعف في كل فقرة قبل اعتمادها. إن مؤشرات الصعوبة والتمييز وفعالية البدائل هي بمثابة "الأشعة السينية" التي تكشف عن صحة الفقرة وقدرتها على أداء وظيفتها.

ويظل الهدف الأسمى هو التأكد من جودة الأداة ككل من خلال التحقق من صدقها وثباتها. فالصدق، بمصادر أدلته المتعددة، يضمن أننا نقيس بالفعل ما نزعم أننا نقيسه، وأن تفسيراتنا واستخداماتنا للدرجات لها ما يبررها. والثبات يضمن أن قياساتنا متسقة ومستقرة وخالية من الخطأ العشوائي قدر الإمكان. إنهما وجهان لعملة الجودة السيكومترية التي لا غنى عنها لأي اختبار يطمح إلى الموضوعية والموثوقية.

وأخيرًا، حتى أفضل الاختبارات بناءً قد تفقد قيمتها إذا لم يتم تطبيقها وتصحيحها وتفسير نتائجها بشكل موحد وصحيح. إن الالتزام بإجراءات التقنين في التطبيق، والدقة في التصحيح، والحذر والموضوعية في التفسير باستخدام المعايير المناسبة وأخذ الخطأ المعياري للقياس في الاعتبار، كلها خطوات ضرورية لضمان الاستخدام العادل والفعال لنتائج الاختبار.

وفوق كل ذلك، يجب أن تحكم الاعتبارات الأخلاقية والقانونية كل خطوة من خطوات هذه العملية، بدءًا من مسؤولية المطور في بناء أداة عادلة وعالية الجودة، مرورًا بمسؤولية المستخدم في اختيار الاختبار المناسب وتطبيقه وتفسيره بحكمة، وانتهاءً بحماية حقوق المفحوصين وضمان خصوصيتهم ومعاملتهم بإنصاف.

إن بناء الاختبارات التحصيلية المقننة هو علم وفن يتطلب جهدًا دؤوبًا وتعاونًا بين المتخصصين في المادة الدراسية والمتخصصين في القياس والتقويم. وعلى الرغم من التحديات والصعوبات، فإن العائد يستحق هذا الجهد، فالاختبارات الجيدة هي أدوات لا غنى عنها لتوجيه العملية التعليمية، وتقويم فعاليتها، واتخاذ قرارات مستنيرة تخدم مصلحة الطلاب والمجتمع، وتسهم في نهاية المطاف في رفع جودة التعليم وتحقيق أهدافه المنشودة. ومع التطورات المستمرة في نظريات القياس (مثل نظرية الاستجابة للفقرة IRT) والتكنولوجيا (مثل الاختبارات المحوسبة التكيفية CAT)، تظل المبادئ الأساسية لبناء الاختبارات الجيدة التي ناقشناها في هذا الفصل هي الأساس المتين الذي يجب أن تبنى عليه هذه التطورات.

المراجع (References)

أبو لبدة، سبع محمد. (1985). مبادئ القياس النفسي والتقييم التربوي. جمعية عمال المطابع التعاونية.

علام، صلاح الدين محمود. (2007). القياس والتقويم التربوي في العملية التدريسية. دار المسيرة للنشر والتوزيع.

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.

American Psychological Association (APA). (2017). Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code/

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

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A .revision of Bloom's Taxonomy of Educational Objectives. Longman

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of .educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Brigham Young University Testing Services and The Department .of Instructional Science

Code of Fair Testing Practices in Education. (2004). Joint Committee on Testing Practices. (Available from the American Psychological Association, Washington, DC). https://www.apa.org/science/programs/testing/fair-testing.pdf

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

.Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). Harper & Row

Downing, S. M. (2006). Selected-response item development basics. In S. M. Downing & T. M. Haladyna .(Eds.), Handbook of test development (pp. 311-334). Lawrence Erlbaum Associates

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Lawrence Erlbaum .Associates

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum .Associates

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Joint Committee on Testing Practices (JCTP). (2004). Code of Fair Testing Practices in Education. .American Psychological Association

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

.Linn, R. L., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8th ed.). Prentice-Hall

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test

.development (pp. 421-442). Lawrence Erlbaum Associates

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

Osterlind, S. J. (1998). Constructing test items: Multiple-choice, constructed-response, performance, and .other formats (2nd ed.). Kluwer Academic Publishers

.Popham, W. J. (2017). Classroom assessment: What teachers need to know (8th ed.). Pearson

.Traub, R. E. (1994). Reliability for the social sciences: Theory and applications. Sage Publications

Welch, C. J. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. .Haladyna (Eds.), Handbook of test development (pp. 335-364). Lawrence Erlbaum Associates

Zieky, M. J. (2006). Fairness review in testing. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of .test development (pp. 365-384). Lawrence Erlbaum Associates