وزارة التعليم العالي والبحث العلمي

جامعة محمد لمين دباغين سطيف 2

كلية العلوم الإنسانية والاجتماعية

قسم علم النفس وعلوم التربية والأرطفونيا

مطبوعة الدعم البيداغوجي في مقياس:

القيـاس النفسـي

موجهة لطلبة السنة الثانية علم النفس

السداسي الأول

إعداد الدكتور: طباع فاروق

السنة الجامعية: 2022-2023

فهرس المحتويات

مقدمة:

يعد قياس القدرات العقلية والسمات النفسية المختلفة دعامة أساسية للممارسة النفسية والتربوية، حيث يقيس علماء النفس مزاج العميل وشخصيته من خلال مقاييس، مثل: قائمة "بيك" للاكتئاب، وقائمة "مينيسوتا" للشخصية، ويقيس المعلمون والمدارس أداء الطلاب من خلال الاختبارات وتقييمات الأداء وتقييمات السلوك، وتستخدم الكليات والجامعات قياسات الكفاءة الدراسية والأداء من خلال "اختبار القدرة الدراسية"، و"امتحانات تسجيل الخريجين" في اتخاذ قرارات القبول حول الطلبة على مستوى الليسانس والدراسات العليا.

وفي جميع هذه الحالات، يتم تقييم وتحديد بعض جوانب السلوك البشري أو الأداء المعرفي أو الحالة النفسية، وتلعب هذه التقييمات دوراً حاسماً في اتخاذ القرار بشأن الأفراد والجماعات، بما في ذلك في سياق تشخيص الأمراض النفسية، وتحديد النجاح في الاختبارات التعليمية، ومنح التراخيص المهنية، ونظراً لأهميتها الكبيرة في العديد من جوانب الحياة الحديثة، يجب فهم أداء هذه القياسات جيداً للتأكد من أنها توفر أفضل المعلومات الممكنة.

على مدى قرن تقريباً، تم تطوير تخصص فرعي يجمع بين الإحصاء وعلم النفس من أجل دراسة مثل هذه المقاييس، والذي يُعرف باسم **القياس النفسي**، يركز هذا المجال على تطوير وفحص التقييمات النفسية والتربوية باستخدام مجموعة متنوعة من الأدوات والأساليب المنهجية، والتي يحاول المتخصصون تغطية العديد منها، مع التركيز على الجوانب النظرية والتطبيقية.

يعد القياس مكوناً رئيسياً في معظم جوانب الحياة الحديثة، من التقدم إلى المدرسة إلى الحصول على عمل، إلى تشخيص وعلاج الأمراض العقلية الخطيرة، وفي محور هذه العملية يوجد متخصص القياس الذي يعمل على التأكد من أن هذه المقاييس تقوم بما صُممت من أجله وأن النتائج مدعومة بالأدلة المناسبة لتبرير الاستخدام.

تتمثل وظيفة أخصائي القياس النفسي في المساعدة على تزويد المعنيين، بما في ذلك الطلاب والمعلمين وأولياء الأمور والمرضى والمتقدمين للوظائف، من بين آخرين بأعلى معايير الجودة الممكنة. ويقومون بذلك من خلال تحليل دقيق وحذر لجميع جوانب عملية القياس، بما في ذلك تحديد المجال المستهدف، وتقييم البنود وجودة المقياس، وتقدير مدى اتفاق المقدّرين على

درجاتهم، وتقديم دليل بشأن ما إذا كان المقياس يقيس البناء المقصود منه، ووضع معايير الأداء، وتحديد المهارات المعرفية المستخدمة في أداء مهمة معينة.

وعلى اعتبار أن القياس النفسي من العلوم المتطوّرة يستند إلى أطر نظرية ومنهجية وإحصائية وتقنيات متقدمة يحتاج إليها الطالب في إعداد اختبارات ومقاييس جديدة أو تقنين اختبارات أعدت من قبل. ويستخدم في المجالات المذكورة أعلاه العديد من أدوات القياس لتقييم مختلف جوانب الشخصية؛ المعرفية، والمهارية، والوجدانية. لذلك فانه من المهم معرفة طرق وفنيات إعداد وبناء مختلف هذه الأدوات على أسس علمية سليمة واستخدامها في مجالات الاهتمام استخداما مناسبا.

إن دراسة القياس النفسي من قبل الطلبة في مقررات الدراسات الجامعية في مجالات علم النفس ذات أهمية بالغة من الناحية العلمية والمهنية، وذلك نظرا للحاجة الملحة إلى القياس الموضوعي من خلال اكتساب معارف ومهارات أساسية، تسمح بتطوير قدراتهم على دمج أفكار جديدة إلى هيكل المعارف وتطبيقها في الحياة المهنية. لذلك فانه من المهم معرفة كيفية بناء مختلف هذه الأدوات على أسس علمية سليم واستخدامها في مجالات الاهتمام استخداما مناسبا. وقد وضعت هذه الوحدة ليساعد الطالب على اكتساب كفاءات جديدة في استخدام مفاهيم وتقنيات تضاف على رصيده المعرفي والمهاري في التعامل مع وضعيات القياس المختلفة.

فالدارسين في العلوم النفسية والتربوية يحتاجون بدرجة أساسية إلى إتقان المفاهيم والأسس والمبادئ المتعلقة بالقياس النفسي وتطبيقاته المتنوعة في المجالات الإكلينيكية، والمهنية، والتربوية...وغيرها، لأن المتخصص النفسي يحتاج لأداء مهامه المهنية إلى إعداد أدوات قياس (اختبارات، مقاييس، قوائم ملاحظات، استبيانات...) أو تكييف المتوفرة منها بهدف الاختيار والتصنيف، والتشخيص والعلاج، والارشاد والتوجيه، وتقويم التحصيل، وتقويم المناهج والبرامج والمشاريع، وتقويم المعلم والمدارس.

وقد وُضعت هذه المادة لتساعد الطالب على اكتساب كفاءة جديدة في استخدام مفاهيم وتقنيات تضاف على رصيده المعرفي والمهاري في التعامل مع وضعيات القياس المختلفة، وبالتالي تهدف وحدة القياس النفسي بشكل عام إلى **أن يكون الطالب قادراً على إدراك النماذج المنطقية والرياضية التي تشكل الممارسة المقننة لبناء الاختبارات واستخدامها في الممارسة المهنية،** ويجب أن يؤدي إدراك هذه النماذج وافتراضاتها ومحدداتها إلى ممارسات مطوّرة في بناء الاختبارات والمقاييس، واستخدام معلوماتها في اتخاذ قرارات صائبة حول الأفراد.

المحاضرة الأولى

القياس النفسي: النشأة والمفهوم

الأهداف:

- يتعرّف الطالب على أهم مراحل تطورات القياس النفسي.
- يحدّد الطالب تعريفاً دقيقاً للقياس النفسي.
- يميّز الطالب بين مفهوم القياس ومفهوم التقويم.
- يعرّف الطالب البناء النفسي تعريفاً دقيقاً.

1. نشأة وتطوّر القياس النفسي

ظهر القياس النفسي منذ زمن بعيد يعود إلى ما قبل الميلاد، ولا زال يتطور باستمرار بوتيرة متسارعة في وقتنا الحاضر، ووفقا للأدبيات النفسية والتربوية المتخصصة في هذا المجال فان تطوّر القياس النفسي بمراحل يمكن تلخيصها كما يلي:

1.1. تطوير اختبارات التحصيل:

يعود ظهور القياس إلى الحضارة الصينية خلال الفترة بين (206 ق. م – 220م) عندما شرعت الحكومة الامبراطورية في استخدام الاختبارات لتحديد الأفراد الذين يُسمح لهم بدخول الخدمة المدنية، والحصول على المناصب الرسمية في الحكومة (Chadha, 2009; Finch & French, 2019)، ومن منظور تاريخي أشارت الكتابات اليونانية والرومانية القديمة إلى محاولات تصنيف الأفراد حسب أنواع الشخصية، حيث تضمنت التصنيفات عادةً إشارة إلى زيادة أو نقص في بعض سوائل الجسم (مثل الدم أو البلغم) كعامل يُعتقد أنه يؤثر على الشخصية.

خلال العصور الوسطى أول استخدام للاختبار في التعليم حدث مع ظهور الجامعات في أوروبا في القرن 13، حيث تم استخدام الشهادات الجامعية كوسيلة لإثبات الأهلية للتدريس، وتم تصميم الامتحانات الشفوية الرسمية لمنح المترشحين فرصة لإثبات كفاءتهم، وبعدها انتشر استخدام الامتحانات في التعليم الثانوي، وبعدها حلت الامتحانات الكتابية محل الامتحانات الشفوية في معظم الأوساط التعليمية. وفي أواخر القرن 19 كانت الامتحانات في أوروبا والولايات المتحدة

طريقة راسخة لتحديد من يُمنح شهادات جامعية ومن سيكون قادراً على ممارسة مهنة معينة كالطب أو القانون (Urbina, 2004) .

2.1. تطوير علم النفس التجريبي:

وفي القرن الثامن والتاسع عشر كانت مساهمات العلماء، أمثال "فوندت" Wundt، و"غالتون" Galton، و"كاتل" Cattel، جوهرية في تطوير قياس القدرات المعرفية من خلال الانتقال إلى قياسها باستخدام إجراءات أكثر موضوعية، ويعتبر "فوندت" بإنشائه أول مختبر للبحث في علم النفس التجريبي في ألمانيا بتطوير أجهزة وإجراءات موحدة لقياس القدرات البشرية في مجال الإحساس والإدراك بدراسة قوانين التي تحكم العلاقة بين العلمين الجسدي والنفسي (Urbina, 2004). وفي نفس الفترة تقريباً اهتم "غالتون" بقياس الوظائف النفسية الذي أنشأ مختبراً للقياسات البشرية في لندن الذي جمع لعدة سنوات بيانات عن عدد من الخصائص الفيزيائية والفسيولوجية المتعلقة بالفروق الفردية كالقياسات الجسمية كالطول والوزن ومحيط الرأس، والحسية كزمن الرجع والتمييز الحسي، والقياسات الحركية كالسرعة وقوة اليقظة (رينولدس و لوفنجستون، 2013).

استمرت أعمال "غالتون" في الولايات المتحدة مع "كاتيل" باستخدام الاختبارات الحسية والحركية البسيطة في قياس القدرات العقلية، ونُظر إلى "كاتيل" أنه أول من استخدام مصطلح "الاختبار العقلي"، وبالتالي وأسهما "غالتون" و"كاتيل" في تطوير إجراءات اختبارية، مثل الاستبيانات المقننة، وموازين التقدير، والتي أصبحت شائعة في تقييم الشخصية (رينولدس و لوفنجستون، 2013). كما كانت مساهمات "غالتون" في مجالات الإحصاء والقياسات النفسية كبيرة، حيث اكتشف ظواهر الانحدار والارتباط التي قدمت الأساس لكثير من الأبحاث النفسية اللاحقة وتحليل البيانات، وبدأ في استخدام الاستبيانات وترابط الكلمات في البحث النفسي (Urbina, 2004) .

إحدى المساهمات الإضافية كانت في أواخر القرن 19 عندما ابتكر "إبنغهاوس" Ebbinghaus – المعروف بأبحاثه في مجال الذاكرة – تقنية تُعرف باسم اختبار التكملة، دعت هذه التقنية الأطفال إلى ملء الفراغات في مقاطع النص التي تم حذف الكلمات أو أجزاء الكلمات منها، حيث أثبتت التقنية أنها مقياس فعال للقدرة الفكرية لأن الدرجات المستمدة منها تتوافق مع القدرة العقلية للطلاب على النحو الذي تحدده الرتبة في الفصل (Urbina, 2004) .

وفي أوائل القرن العشرين، كان ظهور أولى الاختبارات النفسية الحديثة نتيجة الاختبارات والأدوات المعملية التي أنشأها علماء النفس التجريبيون الأوائل في ألمانيا، وأدوات القياس والتقنيات

الإحصائية التي طورها "جالتون" وطلابه لجمع وتحليل البيانات حول الفروق الفردية، وتراكم النتائج الهامة في العلوم الناشئة في علم النفس والطب النفسي وعلم الأعصاب، حيث قدمت كل هذه التطورات الأساس لظهور الاختبارات الحديثة.

3.1. تطوير تقييمات القدرات:

اختبارات الذكاء: في عام 1904 اهتم "بينيه" بفحص الفروق الفردية عن طريق مجموعة متنوعة من التدابير البدنية والفسيولوجية، واختبارات العمليات العقلية الأكثر تعقيدًا كالذاكرة والفهم اللفظي، وابتكر مع "سيمون" طريقة لتقييم الأطفال الذين يعانون من التخلف العقلي، والذي اعتبر أول اختبار للذكاء Binet–Simon، وقد تم إدخال مفهوم "العمر العقلي" وتعريف مفهوم معامل الذكاء من قبل الألماني "ستيرن" Stern في عام 1914 (Finch & French, 2019) . أجريت تقييمات لاختبار "بينيه وسيمون" عام 1908 و1911، وحظيت بقبول واسع، وتمت ترجمتها وتقنينها في الولايات المتحدة من قبل "تيرمان" Terman بجامعة ستانفورد، ونتج عنه اختبار "ستانفورد–بينيه" للذكاء الذي أصبح شائعاً أكثر من الاختبار الأصلي، واستمر استخدامه لعدة عقود (رينولدس و لوفنجستون، 2013؛ Urbina. 2004).

وفي الفترة نفسها التي تم فيه تطوير اختبارات الذكاء، نشر "سبيرمان" في عام 1904 ورقتين بحثيتين وضعتا الأساس لتحليل العوامل والنظرية الكلاسيكية للاختبار، كما قدم أيضًا مفهوم الذكاء البشري العام، وفي سنوات 1930 توسع "ثورستون" Thurstone في فكرة الذكاء البشري كبنية وحدوية إلى أنه يتكون من جوانب متعددة، مثل الذكاء اللفظي وغير اللفظي. في المقابل، تطلب هذا التوسع في ما كان يقصد بالذكاء توسعًا مصاحبًا في النماذج الإحصائية اللازمة لفهمه، مما أدى إلى ظهور الرياضيات التي من شأنها أن تكون بمثابة الأساس لتحليل العوامل (Finch & French, 2019).

وفي سنة 1930 طوّر "ويكسلر" اختبار ذكاء اشتمل على مقاييس القدرة اللفظية وغير اللفظية في الاختبار نفسه، بعدما كانت اختبارات الذكاء بما فيها مقياس Wechesler– Bellevue I تقيس الذكاء اللفظي أو غير اللفظي، وليس كلاهما، وأصبحت مقاييس "ويكسلر" أكثر اختبارات الذكاء شيوعاً واستخداماً وقتنا الحاضر (رينولدس ولوفنجستون، 2013؛ Chadha. 2009).

بطاريات القدرات المتعددة: تم تطوير بطاريات القدرات المتعددة Multiple Aptitude Batteries لاستخدامها بشكل متسع في التوجيه المهني في إنجلترا والولايات المتحدة خلال

1920 و1930 من القرن الماضي، وهي عبارة عن مجموعات من الاختبارات ذات صيغة وتصحيح مشترك، تحدد نقاط القوة والضعف للفرد بتقديم درجات في عوامل مختلفة كالتفكير اللفظي والعددي والمكاني والمنطقي والقدرات الميكانيكية، بدلاً من الدرجة العامة التي يوفرها اختبار الذكاء "بينيه" واختبارات "ألفا وبيتا" (Urbina, 2004) . وقد ظهرت بطاريات القدرات المتعددة بعد اتساع استخدام التحليل العاملي بإدراك أن الذكاء ليس مفهوما أحادي البُعد، وأن القدرات البشرية تشتمل على مجموعة واسعة من المكونات أو العوامل المنفصلة والمستقلة نسبياً.

وفي أثناء الحرب العالمية الأولى أعدّ "يركس" Yerkes مع مجموعة عمل سلسلة اختبارات الاستعداد تُعرف باسم اختبارات "ألفا" و"بيتا"، أحدهما لفظي والآخر غير لفظي تستخدم لتقييم وتصنيف المجندين للقبول في الخدمة العسكرية (رينولدس و لوفنجستون، 2013). ساعدت هذه الاختبارات القادة العسكريين في وضع الأفراد في وظائف داخل الجيش، بالإضافة إلى ظهور مدرسة جديدة لعلم النفس العسكري، ووحدة إحصائية لتحليل البيانات من بطاريات التقييم، واستمر هذا العمل اليوم مع تطوير واستخدام نماذج القياس النفسي المتقدمة، والتي تم دمجها في مجال تطوير القوى العاملة واختيار المتقدمين من خلال جميع مستويات الأعمال (Finch & French, 2019).

الاختبارات المقننة: وفي بداية القرن 20 نشر ثورندايك Thorndike أول كتاب في نظرية القياس "مقدمة في نظرية القياس العقلي والاجتماعي" الذي حاول أن يضع اعتبارات منهجية لمشكلات القياس في علم النفس (Crocker & Algina, 2006) . وطوّر "ثورندايك" اختبارات التحصيل المقننة، واختبار الخط اليدوي الذي نُشر عام 1910، وهذا ما فتح آفاقاً جديدة لإنشاء سلسلة من نماذج الكتابة اليدوية تتراوح بين الضعيفة جداً والممتازة التي يمكن مقارنة أداء الأشخاص. وبعدها تبعتها الاختبارات المقننة المصممة لتقييم مهارات الحساب والقراءة والتهجئة التي أصبحت عناصر أساسية في الأوساط التعليمية، وفي منح الترخيص والاعتماد للمهنيين الذين أكملوا تدريبهم، واختيار الموظفين (Urbina, 2004) .

وبدأ استخدام الاختبارات الموضوعية في المدارس الثانوية لغرض اتخاذ قرارات القبول في الكليات والجامعات، أدى هذا التطور إلى إعداد اختبار القدرات الدراسية (SAT) في عام 1926، إلى وصول العديد من الأدوات التي تُستخدم لاختيار المرشحين للخريجين والمدارس المهنية. من أمثلتها المعروفة "امتحان سجل الخريجين" (GRE)، واختبار القبول في كلية الطب (MCAT)،

واختبار القبول في كلية الحقوق (LSAT) التي تؤكد على القدرات اللفظية والكمية والاستدلالية اللازمة للنجاح في معظم المساعي الأكاديمية. وعلى الرغم من اختلاف الغرض منها عن أهداف اختبارات التحصيل المقننة، إلا أن محتواها غالبًا ما يكون متشابهاً (Urbina, 2004) .

4.1. تطوير التقييمات الاكلينيكية:

قوائم الشخصية: قام "وودسوورث" Woodsworth قائمة بيانات الشخصية من خلال جمع تقارير من الأطباء النفسيين بشأن السمات التي ارتبطت بالأمراض النفسية واستخدمها في تطوير بنود لتوفير معلومات تشخيصية تساعد على تحديد المرضى الذين يعانون من اضطرابات نفسية. كما تم تطبيق الاختبار النفسي وتحليل العوامل على قياس الشخصية خلال هذه الفترة من قبل "ايزنك" Eysenck الذي طوّر أداة لقياس السمات كالانبساط والتوافق(Finch & French, 2019)

وتم تطوير استبيان "مينيسوتا للشخصية المتعدد الأبعاد" MMPI في بداية 1940 وهو اختبار موضوعي لتشخص الاضطرابات الطبية النفسية، وقد خضع لعدد كبير من البحث، ولا يزال الاستبيان MMPI–2 من أكثر مقاييس الشخصية استخداماً وشويعاً في وقتنا الحاضر، كما طوّر "رورشاخ" Rorschach اختبار بقع الحبر بداية من 1920 كأسلوب اسقاطي ظل من أساليب تقييم الشخصية الأكثر شيوعاً واستخداماً في القرن 21 (رينولدس و لوفنجستون، 2013).

منذ الأربعينيات، ازدهرت قوائم الشخصية التي تم إدخال العديد من التحسينات في بنائها، بما في ذلك استخدام المنظورات النظرية، مثل نظام "موراي" Murray للاحتياجات (1938) وطرق التناسق الداخلي لاختيار البنود واستخدام التحليل العاملي الذي كان حاسماً للغاية في دراسة القدرات وتمييزها في تطوير قوائم الشخصية، ووقد كان "جيلفورد" Guilford رائداً في استخدام التحليل العاملي لتجميع البنود في مقاييس متجانسة، بينما في الأربعينيات طبق "كاتيل" Cattel التحليل العاملي دوراً أساسيًا في معظم جوانب نظرية الاختبار وبناء الاختبارات.

قياسات الاهتمامات والاتجاهات: مثلما استخدمت اختبارات المهارات والقدرات الخاصة في الصناعة نشأت مقاييس الاهتمامات لغرض التوجيه المهني التي استخدمت فيما بعد في اختيار الموظفين، حيث أنتج "كيلي" Kelley عام 1914 اختباراً بسيطاً للاهتمامات لأول مرة مع بنود تتعلق بتفضيلات مواد القراءة والأنشطة الترفيهية بالإضافة إلى معلومات عن معرفة الكلمات والمعلومات العامة. وطوّر "ريم" Ream عام 1924 مفتاح امبريقي يميّز الاستجابات الناجحة وغير الناجحة لموظفي المبيعات في قائمة "كرنيجي للاهتمامات" المطوّرة من طرف "ياكيوم"

Yoakum وطلابه عام 1921. يمثل هذا الحدث بداية تقنية تُعرف باسم مفتاح المحك التجريبي، والذي سيتم استخدامه بعد التحسينات والاضافات التي تتلاءم مع المهن الأخرى، ونشرت لأول مرة "قائمة الاهتمام القوي" التي تعتبر أحد قوائم الأكثر استخداماً (Urbina, 2004) .

وقدّم "ثيرستون" و"تشيف" Thurstone & Chave بإضافة تقنيات لتدريج مقاييس الاتجاهات إلى الأدبيات المتنامية في تطوير الاختبارات، وطوّر "ثيرستون" Thurstone و"كيلي" Kelley و"هولتزنجر" Holtzinger إجراءات جديدة في التحليل العاملي (Crocker & Algina, 2006).

الاختبارات العصبية النفسية: في القرن الماضي اتجهت الدراسات العلمية والسريرية نحو البحث في علاقات الدماغ والسلوك، وهو موضوع علم النفس العصبي الذي اهتم بها "غولدشتاين" Goldstein الذي لاحظ في الجنود الذين تعرضوا لإصابات دماغية خلال الحرب العالمية الأولى نمط من العجز يتضمن مشاكل في التفكير المجرد، والذاكرة، وتخطيط وتنفيذ المهام البسيطة نسبياً. طوّرت الاختبارات الأدائية عبارة عن اختلافات في الأداء – على عكس الاختبارات اللفظية – التي تم تطويرها لتقييم القدرة الفكرية العامة للأفراد الذين لا يمكن فحصهم باللغة الإنجليزية أو الذين يعانون من إعاقات في السمع أو النطق، حيث تضمنت الاختبارات مواد مثل ألواح التشكيل وألغاز الصور المقطعة والكتل بالإضافة إلى مهام الورق والقلم كالمتاهات والرسومات (Urbina, 2004). وقد استمر مجال التقييم النفسي العصبي في النمو في عدد وأنواع الأدوات المتاحة، وساهم في الفهم السريري والعلمي للعلاقات العديدة والمتنوعة بين وظائف الدماغ والإدراك والعواطف والسلوكيات.

2. تعريف القياس:

من المعروف حسب "ثورندايك" Thorndike أن كل شيء موجود في الطبيعة موجود بمقدار، وعملية تحديد مقداره أو كميته هو موضوع أو مجال اهتمام القياس. وقد قدمت العديد من التعريفات للقياس، ويعدّ تعريف Stevens (1946) من أدق التعريفات، والذي يشير إلى تقديم أو تأشير بالأرقام للأشياء والأحداث وفق قواعد معينة (Crocker & Algina, 2006).

وبشكل عام، وفي مجالات العلم المختلفة يتضمن القياس القواعد التي يتم بواسطتها تقديم رموز للأشياء بهدف تمثيل كميات السمات رقمياً أو تحديد مدى انتماء الأشياء إلى الفئة نفسها أو إلى فئات مختلفة فيما يتعلق بسمة معينة. وتجدر الاشارة إلى أن القياس يكون لخصائص القياس الأشياء وليس للأشياء نفسها، والأشياء أو المواضيع في علم النفس عادة هم الأفراد، فقياس قدرة

الفرد على التفكير يتم بقياس تجلياتها السلوكية التي يُعبر عنها بمؤشرات تثبت مقدار السمة عند الفرد، وليس قياس التفكير في حد ذاته كظاهرة إنسانية بديهية يستدل عليها مباشرة.

تسمح عملية إعطاء الأرقام بتحويل الأبنية (أو المفاهيم النظرية) من نظريات مجردة إلى مؤشرات ملموسة، فالقياس حسب (Bertrand & Blais, 2004) هو العملية التي تسمح لنا بالانتقال من المفاهيم المجردة إلى المؤشرات الملموسة، وبالتالي فهو مجموعة العمليات الامبريقية التي تسمح بالحصول عن طريق أداة معينة أو مجموعة أدوات على بيانات لتصنيف موضوع في فئة أو خاصية معينة.

وفي السياق نفسه يعرف (Hubley & Zumbo, 2013) القياس بأنه الوصف الرقمي للسمات أو الخواص في شكل أرقام، وتستخدم قواعد وإجراءات لتعيين تلك الأرقام، ويعرفان **القياس النفسي** بأنه ميدان للدراسة يركز على النظرية والتقنيات التي تتعلق مبدئيا بقياس الأبنية (المفاهيم النفسية)، وكذا تطوير وتفسير وتقييم الاختبارات والقياسات. وفي القياس الاجراءات والقواعد التي تطبق لتعيين هذه الأعداد، ومن المهم التذكير بأن الأبنية التي يعينها السيكولوجيون، والاختبارات والمقاييس التي يطورونها لقياس هذه المفاهيم، والاجراءات والقواعد التي يستخدمونها ليست قيماً اختيارية بالعكس كل مرحلة من العملية تتضمن قرارات تعكس قيماً شخصية واجتماعية وثقافية.

لا تنحصر عملية القياس على الاختبار فقط، وإنما أكثر من ذلك، فوضعية العملية الاختبارية عملية للحصول على درجات الاختبار يجب أن تؤخذ بعين الاعتبار، فعملية القياس تشمل كل جوانب وضعية العملية الاختبارية، مثل فترة أو وقت تطبيق الاختبار، واستخدام المقدرين (المصححين)، واختيار بنود معينة للاختبار، وكيفية تطبيق الاختبار، والشروط المقننة لتطبيق الاختبار. عملية القياس تتضمن جوانب متعددة في العملية الاختبارية ولا تتوقف ببساطة على الاختبار فقط، فكل جوانب عملية القياس يمكن أن تؤثر على اتساق الدرجات (Meyer, 2010)

يختلف **التقويم** عن القياس رغم أن القياس مرحلة أساسية ومهمة في عملية التقويم، على اعتبار أن التقويم عملية تتم من خلال أربع عمليات أساسية متتابعة؛ القياس، والتقدير، وإصدار الحكم، واتخاذ القرار.

التقويم = القياس + التقدير + إصدار الحكم + اتخاذ القرار

يشير التقويم إلى أنه عملية شاملة لجمع معلومات حول الفرد واستخدامها للقيام باستدلالات حول خصائص الفرد أو التنبؤ بأدائه، ويتضمن التقويم تجميع ومقارنة المعلومات المحصلة من مصادر متنوعة مثل: المقابلات، والسجلات، والملاحظة، ونتائج الاختبار والمعلومات المحصلة من مصادر أخرى متضمنة العائلة، والأصدقاء أو المهنيين (Hubley & Zumbo, 2013) .

وفي هذا السياق يعرف "دوكيتيل" (De Ketele, 1993) التقويم بأنه عملية جمع مجموعة معلومات ملائمة وثابتة وصادقة ما فيه الكفاية، واختبار درجة ملاءمة هذه المعلومات مع مجموعة المحكات المحددة للأهداف أو المعدّلة أثناء العملية بهدف اتخاذ قرار معين.

يختص هذين التعريفين بتقويم أنواع مختلفة من السمات النفسية والتربوية والاجتماعية، لذلك رغم اشتراك هذه السمات من حيث الخصائص إلا أنّها تركز على فئة معينة من السمات. لذلك يقدم بعض المؤلفين تعريفات للتقويم التربوي، حيث يعرّفه (Miller, Linn, & Gronlund, 2009) بأنه مصطلح عام يتضمن مجموعة واسعة من الاجراءات المستخدمة في الحصول على معلومات عن تعلم الطلاب (ملاحظات، تقديرات حول الانجازات أو المشاريع، اختبارات الورقة والقلم) وتكوين أحكام قيمة حول تقدم التعلم.

وفي مجال التقويم الصفي تتفق الكثير من الكتابات بأن عملية التقويم تتلخص في أربع خطوات وأربع فترات؛ 1) القصد الذي يكشف عن الموضوع، وتتضمن تحديد غرض وفترة التقويم، 2) القياس الذي يتكون من جمع وتفسير المعلومات، 3) إصدار الحكم، 3) اتخاذ القرار. وكل خطوة من العملية التقويمية تستند إلى مجموعة من الأسئلة، والخيارات، والقرارات التي يجب على المقوّم اتخاذها بغرض تقويم التعلمات.

3. تعريف البناء النفسي:

قبل تصميم اختبار معين يجب تحديد البناء (المفهوم أو السمة الكامنة) الذي نريد قياسه بوضوح، فإذا أردنا مثلا قياس كفاءة طالب في الرياضيات نحتاج إلى تحديد المقصود بـ "كفاءة الرياضيات" فهل تتضمن كفاءة الحساب، أو الهندسة، أو القياس أو تتضمن الحساب فقط؟ وإذا انصب اهتمامنا على كفاءة الحساب توجد جوانب مختلفة من الحساب التي نحتاج إلى أخذها بعين الاعتبار، فهل نهتم فقط بالمعادلات أو بالمتراجحات أو بالقسمة الاقليدية. فاذا لم يتم بوضوح تحديد البناء فلا نستطيع أن نحدّد بشكل مفصل ما نريد قياسه بالضبط.

من الممكن أن يصمم العديد من المطورين اختبارات تختلف إلى حد ما إذا لم يتم تحديد بشكل جيّد البناء، كما من أنه المحتمل أن تختلف (تتغير) درجات الأفراد في الاختبار تبعا للاختبارات الأخرى المعدة، وسوف تكون أيضاً عملية تفسير درجات الاختبار موضوعا للجدل.

يشير البناء النفسي حسب (Cronbach & Meehl, 1955, p. 283) إلى أنه خاصية معينة مفترضة عند الأفراد، ويفترض أن يتم الكشف عنها أثناء أداء الاختبار، وأثناء جمع أدلة صدق الاختبار فان السمة التي نقوم بالكشف عنها في تفسير الاختبار هو البناء.

كما يشير البناء إلى المفهوم أو السمة أو المتغير الذي يكون هدفاً للقياس، وفي عملية القياس النفسي يعبّر البناء عن سمة موضع الاهتمام غير قابلة للملاحظة مباشرة، ومعظم أهداف القياس في التقويم النفسي، بغض النظر عن مستوى خصوصيتها، فإنها بُنى (تكوينات) من حيث أنها سمات أو أبعاد لدى لأفراد محددة نظرياً.

في العلوم النفسية والتربوية يهتم الباحثون عادة بدراسة الأبنية النظرية التي لا يمكن ملاحظتها مباشرة، وهذه الظواهر المجردة تسمى متغيرات كامنة أو عوامل، ومن أمثلة المتغيرات الكامنة في علم النفس تقدير الذات، والدافعية، وفي التربية القدرة اللفظية، وكفاءة المعلم. لأن المتغيرات الكامنة غير قابلة للملاحظة مباشرة، وبالتالي فهي غير قابلة للقياس مباشرة، وبالتالي فان الباحث يجب أن يحدد إجرائياً المتغير الكامن مجال الاهتمام في شكل سلوكيات يعتقد أنها تمثّله. مثل هذا المتغير غير الملاحظ يرتبط بها بشكل ملاحظ، وبذلك يجعل قياسه ممكنا، ويكون تقييم السلوك حينها يشكل قياس مباشر للمتغير الملاحظ، ولو أن القياس المباشر يكون للمتغير غير الملاحظ.

عكس الخصائص الفيزيائية فان الخصائص النفسية للفرد لا يمكن قياسها مباشرة كما في قياس الطول أو الوزن لأنها عبارة عن أبينة أو مفاهيم، إنها مفاهيم افتراضية لا يمكن اثباتها بشكل مطلق بل يمكن استنتاجها من الملاحظات السلوكية للفرد، فمن الضروري عند قياس البناء عرض قوانين التطابق بين البناء النظري والسلوكات الملاحظة التي تعتبر مؤشرات منطقية لهذا البناء، وتسمى هذه العملية بتأسيس التعريف الاجرائي للبناء (Crocker & Algina, 2006) .

المحاضرة الثانية

نظريات القياس النفسي

الأهداف:

- يتعرف الطالب على مفاهيم ومبادئ نظريات القياس النفسي.
- يميّز الطالب بين تطبيقات أهم نظريات القياس النفسي.

يرى هيوبلي وزيمبو (Hubley & Zumbo, 2013) بأنه لا توجد نظرية أو مقاربة واحدة في القياس النفسي، وإنما توجد على الأقل ست نظريات للقياس مترابطة فيما بينها في شكل فئات يمكن أن تجتمع ضمن ما يُعرف بنظريات الدرجة الملاحظة Observed، ونظريات المتغير الكامن Latent Variable Theory. تتضمن مقاربات الدرجة الملاحظة؛ النظرية الكلاسيكية للاختبارات Classical Test Theory ونظرية إمكانية التعميم Generalizability Theory، وتتضمن مقاربات المتغير الكامن نظرية التحليل العاملي Factor analysis theory، ونظرية الاستجابة للبند Item response theory، ونظرية راش Rash theory، والنماذج المختلطة Mixed Models، ويعتبر العديد من السكومتريين أن نظرية "راش" حالة خاصة من نظرية الاستجابة للمفردة، في حين يميّز آخرون بين النظريتين.

سوف نستعرض هنا ثلاث نظريات؛ نظرية الاستجابة للبند، ونظرية إمكانية التعميم، ونظرية الاستجابة للبند.

1. النظرية الكلاسيكية للاختبارات:

النظرية الكلاسيكية للاختبارات Classical Test Theory أو نظرية الدرجة الحقيقية True Score Theory هي نظرية تقليدية أسسها سبيرمان Spearman في بداية القرن العشرين، حيث قدم أدلة رياضية ومنطقية على أن درجات الاختبار عرضة للخطأ أثناء قياس السمات الانسانية، والارتباط الملاحظ بين درجات الاختبار المعرضة للخطأ يكون أقل من الارتباط بين الدرجات الموضوعية الحقيقية (Crocker & Algina, 2006) .

النظرية الكلاسيكية للاختبارات عبارة عن نموذج بسيط ومفيد يصف كيفية تأثير أخطاء القياس على الدرجات الملاحظة، ويفترض هذا النموذج شروطا حتى يكون واقعياً فاذا كانت افتراضاته منطقية فان الخلاصات المستمدة تكون منطقية، ولكن إذا كانت افتراضاته غير منطقية فان النموذج يؤدي إلى خلاصات خاطئة.

وتعد النظرية الكلاسيكية من النظريات الأولية التي ظهرت في القياس النفسي التي وصفها العديد من المؤلفين بشكل مستفيض خاصة في المراجع الكلاسيكية، وتبقى من أكثر النظريات تأثيرا في علم النفس، واعتبرت كلاسيكية لأن مبادئها وافتراضاتها تقليدية تعتمد على مبادئ النموذج البسيط للعلاقة الخطية بين الدرجة الملاحظة Observed score، والدرجة الحقيقية True score، ودرجة الخطأ Error score، حيث تقوم النظرية الكلاسيكية في جوهرها على أن الدرجة الملاحظة للفرد (X) في الاختبار (أو أي أداة أخرى) يمكن أن ينظر إلى أنّها تتضمن مكونين افتراضيين هما؛ الدرجة الحقيقية (T) والخطأ العشوائي (E) المعبر عنها بالصيغة التالية:

الدرجة الملاحظة = الدرجة الحقيقة + درجة الخطأ

$$\boldsymbol{X = \tau + E}$$

تشير هذه الصيغة إلى أن الدرجة الملاحظة تساوي مجموع الدرجة الحقيقية ودرجة الخطأ، بحيث يُعبّر الرمز **(X)** عن الدرجة الملاحظة التي يحصل عليها الفرد في الاختبار، ويُعبّر الرمز **(τ)** إلى الدرجة الحقيقية للفرد التي تعكس قدراته أو مهاراته أو اتجاهاته الحقيقية أو ما يقيسه الاختبار قياسا تاما من غير خطأ قياس، في حين يعبّر الرمز **(E)** إلى خطأ القياس. ووفقا لهذا النموذج فان الدرجة الحقيقية للفرد غير قابلة للقياس مباشرة، ولكن يمكن تقديرها بتحديد متوسط الدرجة التي يحصل عليها الفرد عن طريق تطبيق عدد افتراضي غير محدود من القياسات المتوازية أو المتكافئة (APA, AERA, & NCME, 2014) .

ويكون للفرد في الاختبار درجة حقيقية واحدة غير أنّه يمكن أن يحصل على درجات ملاحظة مختلفة عندما يُطبق عليه الاختبار في فترات مختلفة، ويعود ذلك إلى القيم المختلفة لأخطاء القياس التي يمكن أن تزيد من درجته الملاحظة أو تخفضها. ويوجد في الواقع العديد من العوامل التي تزيد أو تخفض من الدرجة الملاحظة للفرد في الاختبار، مثل التخمين أو المشتتات في البنود، والتغيّرات في سلوكات المفحوصين، وفي إجراءات تطبيق الاختبار ، وأخطاء في تصحيح الاختبار. مثل هذه الأخطاء العشوائية يمكن تؤثر على الأداء العام للفرد على الاختبار، ومن

الممكن أثناء إعادة تطبيق الاختبار أن لا يتكرر الخطأ العشوائي، ويمكن أن تظهر أخطاء عشوائية أخرى مما يؤدي إلى اختلاف في درجة الفرد باختلاف فترة تطبيق الاختبار أو باختلاف صيغة الاختبار أو باختلاف المصحح القائم بوضع الدرجات، أو اختلاف فترة التصحيح.

اشتقت النظرية الكلاسيكية للاختبارات مجموعة من الافتراضات المهمة التي وردت في العديد من أدبيات القياس (مثلا، de Gruijter & van der Kamp, 2008; Laveault & Grégoire, 2014) وهي أن:

1)- درجة الفرد الملاحظة في اختبار معين تنقسم إلى درجة حقيقية، ودرجة خطأ عشوائية $\boldsymbol{X = \tau + E}$

2)- القيمة المتوقعة للدرجة الملاحظة هي الدرجة الحقيقية $\boldsymbol{\sum(x) = \tau}$

3)- لا يوجد ارتباط بين الدرجة الحقيقية وخطأ القياس في مجتمع الأفراد الذين يطبق عليهم الاختبار $\boldsymbol{\rho\tau E = 0}$

4)- تكون أخطاء القياس في اختبارين مختلفين غير مرتبطة $\boldsymbol{\rho E_1\, E_2 = 0}$

5)- لا يوجد ارتباط بين خطأ القياس في اختبار معين والدرجة الحقيقية في اختبار آخر $\boldsymbol{\rho E_1\, \tau_2 = 0}$

6)- يعتبر الاختباران متكافئان إذا كانت درجاتهما الحقيقية متساوية $(\boldsymbol{\tau = \tau'})$، وأخطائها المعيارية للقياس أيضاً متساوية $(\boldsymbol{\sigma'_E = \sigma_E})$

7)- يعتبر اختباران متكافئان عندما تختلف درجاتهما الحقيقية بثابت إضافي (Tau- $\boldsymbol{\tau}$ Equivalent) $\boldsymbol{\tau_1 = \tau_2 + K}$

قدمت افتراضات النظرية الكلاسيكية طرقاً متعددة لتقدير ثبات درجات الاختبارات (الصيغ المتكافئة، الاختبار-إعادة الاختبار، الاستقرار والتكافؤ، الاتساق الداخلي، الاتفاق بين التقديرات) تبعا لتعدّد مصادر خطأ القياس التي تؤثر على درجات الاختبارات. فمن أجل تقدير خطأ القياس الراجع إلى تغيّر (أو اختلاف) الدرجات عبر الزمن تستخدم طريقة الاستقرار (الاختبار-إعادة الاختبار)، ولتقدير خطأ القياس الراجع إلى اختلاف درجات الاختبارات المتكافئة أو البديلة تستخدم طريقة الصيغ المتكافئة أو البديلة، ولتقدير خطأ القياس الراجع إلى اختلاف تجانس البنود

تستخدم طرق الاتساق الداخلي، ولتقدير خطأ القياس الراجع إلى اختلاف درجات التقديرات تستخدم طرق الاتفاق بين المقدرين أو المصحّحين.

2. نظرية إمكانية التعميم:

تعتبر نظرية إمكانية التعميم Generalizability Theory أو النظرية المتعددة الأبعاد Multi-faceted Theory من نظريات القياس الحديثة التي تعتبر امتداداً للنظرية الكلاسيكية للاختبارات؛ لأنها استُخدمت بشكل عام لتجزئة خطأ القياس إلى أبعاد أو مصادر متعددة الناتجة عن العبارات المختارة أو المصحّحين المعتمدين أو فترات تطبيق الاختبار، حيث يتم ضمنيا تفكيك الخطأ إلى مصادر متعددة من خلال إعادة تحديد الدرجة الحقيقية (الدرجة الشاملة).

طُوّرت نظرية إمكانية التعميم في بداية الستينات من القرن الماضي من طرف كرونباخ وراجاراتنم وجليزر (Cronbach. Rajaratnam & Gleser. 1963) لتحرير النظرية الكلاسيكية من خلال تقديم طرق تسمح بمعالجة مصادر متعددة للخطأ في الوقت نفسه، والتي يمكن أن ترجع إلى الأفراد المختبرين في حدّ ذاتهم، أو صيغ الاختبار المطبقة، أو نوع أسئلة الاختبارات، أو فترات تقديم الاختبارات.

ونظرية إمكانية التعميم هي إطار فكري وإحصائي موسّع يسمح بتقدير ثبات القياسات (السلوكية أو التربوية) في وضعيات تتأثر بمصادر خطأ متعددة للقياس، كما تزوّد الفاحصين بطرق ذات فاعلية لتحسين دقة القياس في المستقبل(Cardinet, Sandra, & Pini, 2010) .

تعتمد نظرية إمكانية التعميم في تجزئة مصادر تباين الخطأ المتعددة على طرق تحليل التباين (ANOVA) بهدف معالجة الوضعيات المتعددة والمعقدة التي تتصف بها القياسات النفسية والتربوية. فالنظرية الكلاسيكية للاختبارات غير قادرة على التحكم في المصادر المتعددة للخطأ التي يمكن أن تؤثر على الثبات، باعتبارها تسمح بضبط مصدر واحد غير مميّز لخطأ القياس، في حين يمكن أن تتحكم نظرية إمكانية التعميم في مختلف أبعاد القياس في الوقت نفسه سواء الراجعة إلى عبارات الاختبار أو صيغه المتكافئة أو فترات تطبيقه.

تتأثر وضعية القياسات السلوكية بمصادر متعددة راجعة إلى عبارات الاختبار التي يجيب عليها الأفراد، والفترات التي يتم فيها تطبيق الاختبار، والمصحّحين الذين يُعتمد عليهم في تقديم الدرجات (في اختبارات الأداء)، هذه الأبعاد Facets منفردة أو متفاعلة فيما بينها (مثلا: تفاعل الأفراد مع

العبارات) تعدّ مصادرا لخطأ القياس يمكن أن تأخذها نظرية إمكانية التعميم بعين الاعتبار في الوقت نفسه.

تميز نظرية إمكانية التعميم بين نوعين من الدراسات: دراسات إمكانية التعميم Generalizability Study، ودراسات القرار Decision Studies، حيث تركز دراسات إمكانية التعميم على تقدير مدى تباين درجات القياس الراجعة إلى مختلف مصادر التباين أو الأبعاد (مثلا، فترات، مقدرين، بنود). وتركز دراسات القرار Decision study على تحديد كيفية تغير معاملات إمكانية التعميم والموثوقية ضمن مختلف ظروف تحسين القياس (مثلا، تغيير عدد مستويات البعد، تغيير تصميم القياس) (Briesch, Swaminathan, Welsh, & Chafouleas, 2014) .

يكمن الفرق بين دراسات إمكانية التعميم ودراسات القرار في أن صانع القرار يستخدم المعلومات المحصلة من دراسة إمكانية التعميم لتقييم مدى فعالية التصميمات البديلة لخفض الخطأ وزيادة الثبات(Shavelson & Webb, 1991) . فالمعلومات التي تمدنا بها دراسات إمكانية التعميم تستخدم في جمع بيانات تفيد في تقدير مكونات تباين القياسات التي نحصل عليها بطريقة ما من الطرق، أما دراسات القرار فتستخدم في جمع بيانات تفيد في صنع العديد من القرارات، والتوصل إلى كثير من التفسيرات والاستنتاجات بناء على النتائج التي توصلت إليها دراسة إمكانية التعميم. التمييز بين دراسات إمكانية التعميم ودراسات القرارات لا يعني أنَ كلا منهما منفصل عن الآخر، ففي الحقيقة أن التصميم التجريبي الأمثل الذي تُبنى عليه دراسة إمكانية التعميم يعتمد على القرار الذي يهم الباحث صنعه (علام، 2000).

3. نظرية الاستجابة للبند:

نظرية الاستجابة للبند The Item Response Theory أو نظرية السمات الكامنة Latent Trait Theory أو نظرية القياس الحديثة Modern Measurement Theory هي بديل للنظرية الكلاسيكية للاختبارات، ظهرت في سنوات 1960 على يد "لورد ونوفيك" Lord & Novick (1968) عندما نشرا كتابًا بعنوان " Statistical Theories of Mental Test Score"، والذي نتجت عنها تطورات سريعة في نظرية الاختبارات. ولقد لقيت هذه النظرية اهتماماً متزايداً في السنوات الأخيرة خاصة في مجال علم النفس والتربية، وغالباً ما يتم تقديمها كبديل حديث أفضل من النظرية الكلاسيكية للاختبارات، رغم أنهما تتشاركان في بعض المبادئ والخصائص الأساسية.

نظرية الاستجابة للبند هي نظرية أو نموذج للقياس العقلي مفاده أن الاستجابات على البنود في اختبار معين تُعزى إلى سمات كامنة، والسمة الكامنة هي قدرة أو خاصية لفرد معين يُستدل عليها اعتماداً على نظريات السلوك وعلى أدلة امبريقية، ولكن لا نستطيع تقييم سمة كامنة معينة بطريقة مباشرة كالذكاء (رينولدس و لوفنجستون، 2013). فهي قياس قائم على النموذج الذي يعتمد على تقديرات مستوى السمات على كل من استجابات الأفراد وخصائص البنود التي تم تطبيقها.

كما تشير نظرية الاستجابة للبند إلى فئة من النماذج الرياضية التي حاولت تفسير العلاقة بين السمات الكامنة (خاصية أو سمة غير قابلة للملاحظة) ومظاهرها (أي النتائج الملاحظة أو الاستجابات أو الأداء)، لغرض إنشاء رابطة بين خصائص البنود والأفراد الذين يستجيبون لها والسمات الأساسية التي يتم قياسها.

تفترض هذه النظرية أن القيمة الاحتمالية لاستجابة فرد معين عن بند اختباري تكون دالة متغيرين رئيسيين؛ المتغير المراد قياسه، وخصائص البند الذي يحاول الفرد الاجابة عنه. ويسمى المتغير المراد قياسه بالسمة الكامنة Latent variable غير قابل للملاحظة المباشرة، ولا يمكن الاستدلال عليه من خلال تغير يسهل قياسه بدقة.

على عكس النظرية الكلاسيكية للاختبار ونظرية إمكانية التعميم تتكون نظرية الاستجابة للبند من فئة من النماذج الرياضية التي لها إجراءات تقدير لمعلمات النموذج (أي، معلمات الفرد والبند) والإجراءات الإحصائية الأخرى للتحقق إلى أي مدى ملاءمة النموذج للبيانات أو استجابات الأفراد على مجموعة من البنود (de Gruijter & van der Kamp, 2008) .

الهدف من نظرية الاستجابة للبند هو تمكين الفاحص من تحديد خصائص معينة للبنود التي تكون مستقلة عمن يجيب عليها، وهذا مماثل للقياس في علوم المادة الذي يمكن فيه قياس سمة كائن معين (مثل الوزن) دون النظر إلى الطبيعة المحددة للكائن (DeVellis, 2016) . فالوزن 50 كلغ تعني الشيء نفسه بغض النظر عما يتم وزنه، وهذا ما تسعى نظرية الاستجابة للبند أن تقوم به مع بنود الاختبار. حيث تنظر هذه النظرية إلى أن استجابة الشخص على بند من بنود الاختبار كدالة لخصائص الفرد وخصائص البند، حيث يُفترض أن استجابة الفرد (أي أداء المفحوص) تعتمد على واحد أو أكثر من العوامل تسمى السمات أو القدرات الكامنة، ويُفترض أن كل بند من مجموعة البنود يقيس السمة أو السمات الأساسية.

كمثال على النموذج البسيط لنظرية الاستجابة للبند هو أن أداء الفرد على بند معين يعتمد فقط على سمة ضمنية واحدة، وأن العلاقة بين أداء الأفراد على بند وأداء البند الضمني يمكن وصفها من خلال وظيفة زيادة متناغمة، وتُسمى الوظيفة الأخيرة عادةً بدالة خاصية البند item characteristic function أو منحنى خاصية البند item characteristic curve التي تحدّد كيف يزيد احتمال الاستجابة الصحيحة على بند معين مع زيادة السمة أو القدرة.

المعيار الذي تقاس عليه درجة الفرد هو البنود التي يتضمنها الاختبار، وفي نظرية الاستجابة للبند فان الاجابة على البنود تعتمد على عاملين وراء الاستجابة التي يظهرها الفرد والمعالم التي يتصف بها البند، والتي تتمثل في درجة صعوبتها وتمييزها. وفي نظرية الاستجابة للبند يتم افتراض وجود سمة أو عدد من السمات لدى الفرد تكون وراء استجاباته عن البنو، وتستخدم هذه السمة (أو السمات) في تفسير الاستجابات، ونظراً لأن هذه السمة (أو السمات) غير ملاحظة مباشرة، لذا يُطلق عليها بالسمة (أو السمات) الكامنة (التقي، 2009).

نظرية الاستجابة للبند تتضمن مؤشرات تحليل البنود الشائعة الصعوبة، والتمييز ، والتخمين، والثبات الشرطي، والخطأ المعياري الشرطي للقياس، فعلى عكس النظرية الكلاسيكية التي تأخذ بعين الاعتبار صعوبة البند والتمييز ذات الصلة بعينة المستجيبين، فان نظرية الاستجابة للبند تقوم على فحص الصعوبة والتمييز عبر نطاق المتغير الكامن (Hubley & Zumbo, 2013) .

انقسمت نظرية الاستجابة للبند إلى مجموعة من النماذج حسب عدد السمات (أو القدرات) التي يقيسها المتغير إلى نماذج أحادية البُعد، ونماذج متعددة الأبعاد، وتختلف هذه النماذج باختلاف نوع البنود (ثنائية الاستجابة أو متعددة الاستجابات)، وخصائص (أو معلمات) البنود (أحادية المعلم، أو ثنائية المعلم، أو ثلاثية المعالم).

وتقوم نظرية الاستجابة للبند على مجموعة من الافتراضات؛ أحادية البُعد Unidimensionality، الاستقلال الموضعي Local Independent، منحنى خاصية البند Item Characteristic Curve، والتحرر من السرعة في الاجابة، حيث أن:

1. **أحادية البُعد** يشير إلى وجود عامل واحد يكمن وراء الأداء في الاختبار يعكس القدرة أو السمة التي يقيسها الاختبار.

2. **الاستقلال الموضعي** يشير إلى عدم اعتماد إجابة الفرد على أي بند من بنود الاختبار على إجابته على أي بند آخر، أي استقلالية البنود عن بعضها البعض.

3. **منحنى خاصية البند** يشير إلى رسم بياني يبيّن احتمالية إجابة الأفراد ذوي القدرات المختلفة إجابة صحيحة على كل بند من بنود الاختبار.

4. **التحرر من السرعة في الاجابة:** يشير إلى أن استجابة الفرد للبنود تتوقف على مقدار ما يمتلكه من القدرة أو السمة المقاسة، وليس لعامل السرعة في الاجابة.

المحاضرة الثالثة

بين القياس المعياري والقياس المحكي المرجع، ومستويات القياس

الأهداف:

- يميّز الطالب بين القياس المحكي المرجع والقياس المعياري المرجع.
- يميّز الطالب بين مستويات قياس المتغيرات النفسية.
- يحدّد الطالب مستوى القياس المناسب للمتغيرات النفسية.

1. بين القياس المعياري المرجع القياس محكي المرجع:

يوجد نوعان من استراتيجيات القياس التي تختلف إلى حد ما فيما بينها، ولكنها تستخدم على نطاق واسع، والمتاحة للقائمين بعملية القياس، وهما: القياس المعياري المرجع والقياس المحكي المرجع، ويتمثل الاختلاف الأساسي بين المقاربتين في طبيعة التفسير المستخدم لفهم أداء الأفراد.

1.1. القياس المعياري المرجع:

من خلال القياس المعياري المرجع Norm-referenced measurement يفسّر أداء الفرد بالنظر إلى انجازات الأفراد الذين تقدّموا من قبل إلى الاختبار نفسه، ويُشار إلى المجموعة السابقة للمتقدمين للاختبار بالمجموعة المعيارية. وبالتالي عندما يحاول المتخصصون إعطاء معنى لدرجة الفرد في الاختبار يتم الرجوع إلى الدرجة السابقة لأداء المجموعة المعيارية، ومن الواضح أن يوصف هذا النوع من التفسيرات بأنها معيارية المرجع (Popham, 2011) .

للتوضيح، عندما يؤكد الفاحص بأن الفرد تحصل على الدرجة المئيني 90 في اختبار القدرات الدراسية، فانه يعني أن أداء الفرد في الاختبار قد تجاوز 90 بالمائة من الأفراد في المجموعة المعيارية للاختبار. وباختصار، التفسيرات المعيارية هي تفسيرات نسبية لأداء الأفراد لأن مثل هذه التفسيرات تركز على كيف يتجمع أداء فرد معين فيما يتعلق بالأداء السابق للأفراد الآخرين.

وبدقة أكثر، فان القياس المعياري المرجع هو مقاربة للتقييم يتم فيها تفسير أداء اختبار الفرد نسبيًا، أي وفقاً لطريقة مقارنة أداء الفرد بأداء الآخرين المتقدمين للاختبار، ويتم تعيين الدرجات من خلال مقارنة أداء الفرد بمجموعة محددة من المعايير التي يتعين تحقيقها، أو الأهداف التي يجب تعلمها، أو المعارف التي سيتم اكتسابها. في حين أن القياس المحكي المرجع هو مقاربة للتقييم يتم

فيها تفسير أداء اختبار الفرد وفقاً لمدى إتقان الفرد لنطاق التقييم المحدد، حيث يتم تفسير الدرجات بمقارنة أداء الفرد في الاختبار بمجال الأداء الذي يعيّنه التقييم للإجابة على سؤال يتعلّق بمقدار الأداء المستهدف الذي حققه هذا الفرد.

2.1. القياس المحكي المرجع:

بالمقابل، يعدّ القياس المحكي المرجع Criterion-referenced measurement تفسيراً مطلقاً لأنه يتوقف على مدى إتقان الفرد للمحك الذي يمثله الاختبار، وبمجرد وصف طبيعة الهدف الذي تم تقييمه بدقة، يمكن تفسير أداء اختبار الفرد وفقاً لدرجة إتقان الهدف (Popham, 2011) . وبدلاً من التفسير المعياري مثلاً الفرد الذي "سجل أفضل من 85 بالمائة من الأفراد في المجموعة المعيارية"، قد يكون التفسير المحكي هو أن الفرد "أتقن 85 بالمائة من محتوى الاختبار. يُلاحظ أن التفسير المحكي لا يعتمد إطلاقاً على أداء الأفراد الآخرين في الاختبار، وينصب التركيز على هدف الاختبار، فمغزى التفسيرات المحكية ترتبط ارتباطاً مباشراً بالوضوح الذي يتم من خلاله تحديد هدف الاختبار.

أوضح (رينولدس و لوفنجستون، 2013) بأنه من الناحية الفنية لا تشير الاختبارات بأنها معيارية المرجع أو محكية المرجع، وإنما تفسير الأداء على الاختبار هو الذي يُشر إليه معياري المرجع أو محكي المرجع. وتجدر الاشارة إلى أنه يمكن بناء اختبارات تقدّم كلاً من تفسيرات مرجعية المعيار ومرجعية المحك. كما أن اختبارات الأداء الأقصى (الاستعدادات والذكاء والتحصيل) تنطبق مع الاختبارات المحكية المرجع بشكل أفضل، أما اختبارات الأداء المميّز (الميول، والاتجاهات، والسلوك) فتنطبق بشكل أفضل مع الاختبارات المحكية المرجع. والجدول رقم يوضح خصائص الدرجات مرجعية المعيار ومرجعية المحك.

جدول (1): خصائص الدرجات في الاختبارات مرجعية المعيار والاختبارات مرجعية المحك

الدرجات في الاختبارات المعيارية المرجع	الدرجات في الاختبارات المحكية المرجع
• تقارن الأداء بمجموعة مرجعية معينة –تفسير نسبي.	• تقارن الأداء بمستوى أداء محدد –تفسير مطلق.
• التفسيرات المفيد تتطلب مجموعة مرجعية موائمة.	• التفسيرات المفيدة تتطلب تعريفاً واضحاً لنطاق المعارف أو المهارات
• يُقيّم الاختبار عادة نطاقاً متسعاً من المعارف أو المهارات.	• يُقيّم الاختبار عادة نطاقاً محدوداً أو ضيقاً من المعارف أو المهارات
• يشتمل الاختبار عادة على عدد محدود من البنود لقياس كل هدف أو مهارة.	• يشتمل الاختبار عادة على عدة بنود لقياس كل هدف أو مهارة
• تُختار عادة البنود بحيث تكون متوسطة الصعوبة، وتُزيد التباين، وتُستبعد عادة البنود الصعبة أو السهلة للغاية.	• تُختار عادة البنود التي تُغطي جيداً نطاق المحتوى، وتُزاوج صعوبة البنود صعوبة نطاق المحتوى.

(رينولدس و لوفنجستون، 2013)

2. مستويات القياس

1.2. المستوى الاسمي:

المستوى الاسمي هو أدنى مستويات القياس الذي يمكن فيه وضع رموز أو أرقام للأسماء والأشياء للتمييز بينها وتصنيفها في فئات معينة، ويتناسب هذا المستوى مع المتغيرات النوعية أو الكيفية، والأرقام التي تستخدم فيها هي بمثابة رموز لا يمكن إجراء العمليات الحسابية المعروفة عليها أو المقارنة بينها. ومن أمثلة المتغيرات من المستوى الاسمي الجنس، الحالة العائلية، التخصص.

2.2. المستوى الترتيبي:

المستوى الترتيبي هو أدق من المستوى الاسمي لأنه بالإضافة إلى وضع رموز أو أرقام للأسماء والأشياء أو الأفراد يمكن ترتيب ومقارنة الأفراد في السمة المقاسة تصاعديا أو تنازليا، ولا يفترض أن تكون المسافة بين الرتب متساوية، كما يكون المستوى الترتيبي في العادة من مستوى مجال في

الأصل يتم تحويله وفق معايير معينة إلى مستوى ترتيبي، ومن أمثلة المتغيرات: تقديرات الطلاب، ترتيب الأطفال في العائلة، مستويات القلق.

3.2. مستوى المجال:

مستوى المجال (المسافات المتساوية، الوحدات المتساوية) هو أدق من المستويين الاسمي والترتيبي، ويتناسب هذا المستوى مع المتغيرات الكمية حيث يمكن إجراء العمليات الحسابية عليها ومقارنتها، ولكن يكون الصفر غير حقيقي أي لا تنعدم السمة عند الصفر المطلق، والمسافة بين التي تفصل درجتين متتاليتين هو نفسه بين درجتين أخريين متتاليتين، لذا سمي بمستوى المسافات المتساوية، ومعظم المتغيرات النفسية والتربوية تندرج ضمن مستوى المجال، مثل الفعالية الذاتية، التوافق النفسي، التحصيل الدراسي لأنّه حاليا طوّر العديد من العلماء المختصين مقاييس لتقييم المتغيرات النفسية والتربوية.

4.2. المستوى النسبي:

المستوى النسبي من أدق مستويات القياس يتناسب مع المتغيرات الكمية المطلقة، ويمتلك نفس خصائص مستوى المجال إضافة إلى الصفر حقيقي (مطلق)، بمعنى أن السمة تنعدم عند الدرجة الصفر. ويمكن إجراء العمليات الحسابية عليه، وتنتج بيانات المستوى النسبي عادة من قياس المتغيرات في العلوم الطبيعية والتقنية، مثل: الوزن، الطول.

وتجدر الاشارة إلى أن المستويين الاسمي والترتيبي يتناسبان مع المتغيرات النوعية أو الكيفية أما المستويين المجال والنسبي فيتناسبان مع المتغيرات الكمية، بالإضافة إلى إمكانية التحويل من المستوى الأدق إلى الأقل دقة بمعنى أنه يمكن تحويل المتغير من المستوى النسبي أو من مستوى المجال إلى المستوى الترتيبي أو الاسمي، والمستوى الترتيبي إلى المستوى الاسمي بينما لا يمكن أن تعكس العملية. فعلى سبيل المثال يمكن تحويل درجات الطلاب (مستوى مجال) إلى مستوى ترتيبي أي إلى تقديرات يمكن ترتيبها تصاعديا أو تنازليا (ممتاز، جيد، حسن، متوسط، ضعيف) أو تحويلها إلى مستوى اسمي يصنف خلالها الطلاب إلى (ناجح، راسب).

يمكن تلخيص مستويات القياس من حيث وظيفتها، والعمليات الحسابية التي يمكن أن يستخدمها، وأمثلة عن كل مستوى وفقا للجدول رقم (2).

جدول (2): ملخص مستويات القياس من حيث وظيفتها، والعمليات الحسابية، وأمثلة عنها

المستوى	الوظيفة	العمليات الحسابية	أمثلة
الاسمي	تستخدم الأعداد في تصنيف الأشياء والأفراد	حساب عدد الحالات في فئة معينة لا يمكن إجراء العمليات الحسابية الأربعة	الجنس الحالة العائلية
الرتبي	تستخدم الأعداد في ترتيب الأشياء أو الأفراد تصاعديا أو تنازليا.	استخدام **>** ، **<** ، ≥ ، ≤ استخدام العمليات الحسابية لمقارنة الرتب	تقدير جيد أفضل من تقدير حسن.
المجال	تستخدم الأعداد في مقارنة قياسات أو درجات الأفراد	الصفر غير مطلق (غير حقيقي) مقارنة مدى الفروق بين قياسين	درجة الفرد A في الذكاء تفوق من درجة الفرد B بمقدار 20.
النسبي	تستخدم الأعداد في تحديد علاقات دقيقة بين الأشخاص أو الأشياء	الصفر مطلق (حقيقي) إجراء العمليات الحسابية المختلفة	الفرد الذي وزنه 85kg ضعف الفرد الذي وزنه 42.5 kg

نشاط تدريبي:

في إحدى الدراسات في مجال علم النفس اهتم أحد الباحثين بدراسة أثر استخدام الحوافز المادية (الأجر، الرحلات، الدعم المعنوي) المقدمة للموظفين في إحدى مؤسسات إنتاج النسيج على تحسين الأداء الوظيفي لديهم. أجرى الباحث قياساً قبليا باستخدام شبكة ملاحظة لتقييم أداء الموظفين، وبعد مدة سنة من تقديم الحوافز المادية لهم، أعيد تقييم أدائهم للتعرف على مدى التحسّن في أدائهم أثناء العمل.

– اعتمد الباحث في إجراء هذه الدراسة على قياسات للمتغيرات، حدد هذه المتغيرات بدقة، وطبيعتها (نوعها)، ومستوى قياسها.

الحل:

المتغير	النوع	مستوى القياس
الحوافز المادية	كيفي	اسمي
الأداء الوظيفي	كمي	مجال

المحاضرة الرابعة

الاختبار: مفهومه وتصنيفاته وخطوات بنائه

الأهداف:

- يحدّد الطالب تعريفاً دقيقا للاختبار.
- يميّز الطالب بين مصطلحي الاختبار والمقياس.
- يصنّف الطالب مختلف أنواع الاختبار وفق معايير محددة.
- يوظف الطالب خطوات بناء الاختبار في إعداد اختبار أو مقياس معين.

1. تعريف الاختبار

تعتبر الاختبارات النفسية أو التربوية من أكثر الأدوات استخداما لقياس سمات وقدرات الأفراد، فيعتمد المتخصصون في قياس خصائص الأفراد لتحقيق العديد من الأغراض منها؛ التشخيص، والانتقاء، والتصنيف، وتقويم الأداء، ومنح التراخيص، وتقويم البرامج. وفي هذا العرض نستخدم كلمة "اختبار" للدلالة على مختلف أدوات قياس السلوك الأخرى، مثل: قوائم التقدير، شبكات الملاحظة، الاستبيانات، المقاييس... وغيرها.

يعرف (Crocker & Algina, 2006) الاختبار بأنه طريقة معيارية للحصول على عينة من السلوك في مجال معين، ويستخدم مصطلح الاختبار ليس للدلالة على المعنى المألوف فحسب، وإنما يستخدم ليشير إلى الطرائق المتبعة للحصول على عينة أقصى أداء للفرد كما في اختبارات الاستعداد والتحصيل أو الأداء المميز للفرد كما في الاستبيانات والمقابلات للدلالة على ميوله ومشاعره واتجاهاته أو الحصول عينة أداء مميز القوائم المعيارية وقوائم السلوك الملاحظة.

يعرّف (Hubley & Zumbo, 2013) الاختبار بأنه طريقة مقننة لمعاينة السلوك، ويمكن أن يشير إلى مجموعة من البنود أو العبارات التي يجيب عليها شخص معين في استبيان أو مقابلة أو في قياس زمن الرجع.

يعرف (Miller, Linn, & Gronlund, 2009) الاختبار على أنّه نوع خاص يشتمل نموذجيا على مجموعة من الأسئلة التي تطبق خلال فترة محددة من الوقت في ظروف متشابهة منطقيا بالنسبة لكل الأفراد.

بناء على التعريفات المقدمة يشير الاختبار إلى مجموعة من المثيرات التي تعبّر عن عينة من سلوك الفرد، والتي يستجيب لها في ظروف معينة خلال فترة زمنية محددة، ويشتمل الاختبار على مجموعة من العبارات فلا يكفي تقديم مثير واحد، وإنما يتطلب عينة ممثلة من المثيرات لتغطية السمة المراد قياسها.

في بعض الأحيان، يتم استخدام مصطلح الاختبار والمقياس بشكل متبادل، ومع ذلك يمكن أن تكون بعض الاختلافات فيما بنهما تتعلق بنوع السمات النفسية واتساع الأداة. حيث يعرف الباحثان (Niko & Brookhart, 2014) الاختبار بأنه أداة أو إجراء منظم لملاحظة أو وصف خاصية أو أكثر للفرد باستخدام مقياس عددي أو مخطط تصنيفي. وتعرف (Urbina, 2004) الاختبار النفسي بأنه إجراء منهجي للحصول على عينات من السلوك ذات الصلة بالأداء المعرفي أو العاطفي، ولتصحيح وتقييم تلك العينات وفقاً للمعايير.

يتم استخدام اختبار بشكل مناسب عندما يجب أن تكون الأداة تعتمد على الكفاءة أو الأداء، أي أنه يعكس القدرة على تقديم إجابة صحيحة لبند أو سؤال معين، بمعنى أن الاختبار يستخدم في العادة عندما يستخدم لقياس أو تقييم القدرات المعرفية (معارف، مهارات، كفاءات).

يعرف المقياس Scale حسب قاموس الجمعية الأمريكية لعلم النفس (American Psychological Association, 2023) بأنه نظام لترتيب الاستجابات في سلسلة متدرجة، وذلك لقياس سمة أو قدرة أو موقف أو ما شابه ذلك. وعلى سبيل المثال، قد يكون لمقياس الاتفاق المستخدم في مقياس الاتجاهات سبعة استجابات تتراوح من غير موافق بشدة (1) إلى موافق بشدة (7)، مع عدم الموافقة أو الاختلاف (4) كدرجة وسطية. ومن الشائع فيما يتعلق بالاتجاهات أو المفاهيم التي لا توجد فيها إجابة صحيحة لها، ولكنها تتطلب تأييداً (أو تقرير) بديل من بين البدائل المقدمة، وغالباً ما يستخدم لقياس الشخصية والميول المهنية التي لا تتطلب إجابة صحيحة، ولكن يتم الإشارة إلى الخيار الأكثر ملاءمة لوصف الفرد فيما يتعلق ببيان محدد.

فالفرق الجوهري بين الاختبار والمقياس في أن الأول يقيس الأداء الأقصى أي المعارف والمهارات والقدرات والكفاءات والاستعدادات التي يمتلكها الفرد في مجال أو عدة مجالات للتعرف على مدى امتلاكه للحد الأدنى (بلوغ محك معين يكون خارجي)، في حين يقيس الثاني الأداء المميّز أي السمات الشخصية كالاتجاهات والقيم والجوانب الوجدانية كالقلق، الاكتئاب، الدافعية... للبحث عن الاختلافات الموجودة بين الأفراد للتمييز فيما بينهم (معيار داخلي للجماعة).

كما أن اجابة الفرد على الاختبار تكون باختيار أو بإنتاج عمل معين أو اختيار إجابة صحيحة من بين مجموعة من الاجابات تعكس معارفه ومهاراته في مجال محدد، في حين أن الاجابة على المقياس لا تكون فيها إجابة صحيحة أو خاطئة، وإنما يتطلب تقرير أو اختيار ما يناسب الفرد من بين مجموعة من البدائل (مثلا: دائما، أحيانا، نادرا، أبدا).

2. تصنيف الاختبارات:

نظرا لتعقد وتشابك الظواهر التربوية والنفسية وتعدد التطبيقات الميدانية لقياساتها (انتقاء الأفراد، توجيه الأفراد، إرشاد الأفراد، تقييم الأفراد والبرامج والمؤسسات...)، تعددت وتنوعت أدوات القياس والتقييم لتتناسب مع طبيعة الظاهرة وما تتضمنه من متغيرات، لذا تصنف الاختبارات والمقاييس النفسية والتربوية إلى:

- **اختبارات جماعية مقابل اختبارات فردية:** تطبق الاختبارات الجماعية على مجموعة كبيرة من الأفراد في وقت واحد، وتطبق الاختبارات الفردية على فرد واحد في وقت واحد.

- **اختبارات السرعة مقابل اختبارات القوة:** اختبارات السرعة (الموقوتة) تحدد زمن الاجابة عليه تحديدا دقيقا، واختبارات القوة تتطلب حل مشكلات صعبة بغض النظر عن زمن الاجابة.

- **اختبارات القدرات مقابل اختبارات الوجدانية:** اختبارات القدرات (اختبارات الاستعدادات، والذكاء، والتحصيل، والابداع) يبذل فيها الفرد مجهودا للاستدلال على قدراته ومهاراته، والاختبارات الوجدانية (اختبارات الشخصية، والاهتمامات، والقيم، والاتجاهات) تقيس سلوكات الفرد في المواقف الانفعالية أو المزاجية.

- **اختبارات الورقة والقلم واختبارات الأداء الموضعي:** اختبارات الورقة والقلم تتطلب من الفرد استخدام الورقة والقلم أثناء الاستجابة لمواقف معينة، في حين تتطلب اختبارات الأداء الموضعي انجاز عمل في وضعية طبيعية حقيقية.

- **اختبارات لفظية واختبارات غير لفظية:** يجيب الفرد في الاختبارات اللفظية شفويا للمواقف المستهدفة، بينما يجيب الفرد في الاختبارات غير اللفظية للمواقف التي يوضع فيها كتابيا أو أدائيا.

- **اختبارات موضوعية مقابل اختبارات مقالية:** يتطلب من الفرد في الاختبارات الموضوعية اختيار اجابة من إجابتين أو من إجابات متعددة، وتتطلب الاختبارات المقالية من الفرد تحليل الموقف أو المثير، وبناء إجابة من نتاجه الخاص.

- **اختبارات معيارية المرجع مقابل اختبارات محكية المرجع:** تستخدم الاختبارات معيارية المرجع لغرض ترتيب الفرد تحديد مكانته أو مقارنته بأقرانه في المجموعة نفسها، في حين تستخدم الاختبارات محكية المرجع لمقارنة أداء الفرد بالهدف المطلوب أو بمحك خارجي بغض النظر عن زملائه في المجموعة.

- **اختبارات التصحيح الذاتي مقابل اختبارات تصحيح الخبير:** يعتمد في اختبارات التصحيح الذاتي تقييم الفرد منتوجه وإعطائه تقديرات عن أدائه بنفسه، ويعتمد في تقييم أداء الفرد في اختبارات تصحيح الخبير على شخص يتصرف كمقدّر كالمعلم مثلا.

- **اختبارات التصحيح اليدوية واختبارات التصحيح الآلي:** تُصحح اختبارات التصحيح اليدوية بالطريقة التقليدية التي تعتمد على القلم أو يدويا، وتُصحّح اختبارات التصحيح الآلي باستخدام وسائل متطوَرة كالكمبيوتر (برمجيات).

- **اختبارات متحررة ثقافيا مقابل اختبارات متخصصة ثقافيا:** تتوجه الاختبارات المتحررة ثقافيا نحو الجمهور مهما كانت ثقافته أو غير متحيزة ثقافيا نحو فئات معينة من الأفراد، وتتوجه الاختبارات المتخصصة ثقافيا نحو فئات محددة من الأفراد ينتمون إلى ثقافة معينة.

وفي هذا السياق قد قدّمت (Chadha, 2009) تصنيفا مختصراً للاختبارات والمقاييس النفسية والتربوية وفق مجموعة من المعايير التي تقوم على طريقة التطبيق، ومعدّل الأداء، والبعد السلوكي، والوسيلة المستخدمة، وطبيعة البنود، وطريقة التفسير، والتصحيح، وقابلية التطبيق، وذلك وفقا للشكل رقم (1):

شكل (1): معايير تصنيف الاختبارات والمقاييس النفسية والتربوية

3. خطوات بناء الاختبار:

بناء الاختبارات ليست بالعملية السهلة كما يعتقد البعض، فهي صعبة للغاية تتطلب مجهودا كبيراً وتحتاج إلى متخصصين في المجال، لذلك يتطلب إتباع مجموعة من الخطوات المنظمة لبناء الاختبار والمطبقة على مدى واسع من أنواع الاختبارات والمقاييس، وهذه الخطوات حددها بعض المؤلفين (Crocker & Algina, 2006; Laveault & Grégoire, 2014) في عشرة خطوات أساسية، نوجزها فيما يلي:

1. **تحديد الأغراض الأولية التي تستخدم فيها درجات الاختبار:** بمعنى تحديد الوظائف التي يمكن أن يؤديها، والغرض من استخدامه، فاختبار الرياضيات يمكن توظيفه لتشخيص صعوبات التعلم، أو لاختيار الأفراد لبرنامج معين، أو لتقييم الحد الأدنى من الكفاءة. لأن تحديد أغراض الاختبار تساعد في تحديد طبيعة الاختبار الذي يجب بناؤه والطريقة التي نستخدمها في ذلك.

2. **تحديد السلوكات التي تمثل البناء النفسي أو نطاقه:** يتم من خلال تحليل مفاهيمي لأنماط السلوك التي يعتقد أنها تمثل البناء المستهدف، ومن ثمّ يحاول إعداد بنود تغطي هذه السلوكات، وللحصول على صورة واضحة منقحة وواضحة للبناء المراد قياسه على يجب على مطوّر الاختبار توظيف واحدة أو أكثر من الأنشطة التالية: تحليل المحتوى، مراجعة الأبحاث، الأحداث العرضية الحرجة، الملاحظات المباشرة، أحكام الخبراء، الأهداف التدريسية.

3. **إعداد مجموعة مواصفات الاختبار:** بعد تحديد الأهداف وخصائص البنود وفئات السلوك الأخرى فان مطوّر الاختبار يحتاج إلى إعداد خطة لاتخاذ قرار على أن كل العناصر متضمنة في الاختبار. وبالتحديد يجب أن يكون هناك توزناً للبنود الممثلة لعناصر الاختبار من وجهة نظر مطوّر الاختبار. ويجب أن يهتم بخاصيتين للبنود وباستقلالية كل منها عن الأخرى في المحتوى والعمليات العقلية التي يستخدمها المفحوص في حل المهمة المطلوبة، ويتألف جدول المواصفات من بُعدين على الأقل، يتضمن الأول مجالات المحتوى الرئيسية والآخر العمليات العقلية.

4. **إعداد ملف أولي للبنود:** يجب أثناء صياغة بنود الاختبار الأخذ بعين الاعتبار مجموعة من النقاط: صيغة البنود، صعوبة البنود، عدد البنود. بمعنى يجب أن يتخذ قرارين؛ ماذا يقيس؟ وكيف يقيس؟ ويتضمن إعداد مجموعة من البنود التي تقيس البناء مراعيا اختيار صيغة مناسبة للبنود، والتأكد بأن الصيغة المقترحة ملائمة لفئة المفحوصين، وتدريب معدّي البنود، وكتابة البنود، ومتابعة تقدَم معدّي البنود ونوعية البنود.

المحاضرة الخامسة

تحليل البنود معيارية المرجع

الأهداف:

- يتعرف الطالب على دلالة تحليل بنود الاختبار.
- يميّز الطالب بين أساليب تحليل البنود معيارية المرجع والبنود محكية المرجع.
- يقدّر الطالب بدقة معامل صعوبة وسهولة البند.
- يصحّح الطالب معامل الصعوبة من أثر التخمين.
- يقدّر الطالب معامل التمييز بأساليب المقارنة الطرفية والأساليب الارتباطية.
- يفسّر الطالب دلالة معاملات الصعوبة والسهولة والتمييز وفق معايير دقيقة.

1. تعريف تحليل البنود:

يعد تطوير الاختبار عملية واسعة النطاق وتستغرق وقتًا طويلاً، حيث يتضمن تحديد الغرض من الاختبار، وتحديد البناء (التكوين الفرضي) ذات الاهتمام، وصياغة بنود الاختبار، وإجراء اختبار تجريبي، وتحليل بيانات الاستجابة على البند. وهذه الخطوات دورية حيث يؤدي تحليل بيانات الاستجابة على البند إلى تحديد البنود الجيدة وتقديم مراجعة البنود السيئة. تخضع البنود المنقحة بعد ذلك لجولات إضافية من الاختبار والتحليل التجريبي.

يقدم العديد من المؤلفين معلومات مفصلة حول كل خطوة من خطوات بناء الاختبار، بما فيها خطوة تحليل البنود (Crocker & Algina, 2006; Laveault & Grégoire, 2014). تحليل بيانات الاستجابة على البنود محور اهتمامنا، لأنها خطوة مهمة في تطوير قياسات ذات جودة.

يعتمد بناء الاختبار على مجموعة من البنود (العبارات أو الفقرات) Items التي تعتبر وحدات أو مثيرات يتم إعدادها من طرف الفاحص لغرض الحصول على بيانات كمية أو كيفية عن السمات العقلية أو النفسية للفرد، وقد تكون هذه المثيرات شفهية أو كتابية أو أدائية تثير استجابته. وباعتبار أن الاختبار يشتمل على مجموعة من البنود فان خصائص الاختبار تعتمد أيضا على خصائص بنوده، وبهدف بناء اختبار جيد يجب إجراء فحص دقيق باستخدام مجموعة من الأساليب الإحصائية، وتسمى هذه الإجراءات بتحليل البنود.

يتضمن الاختبار مجموعة من البنود الذي تعتمد خصائصه على خصائص بنوده، ولغرض بناء اختبار ذات فاعلية يجب فحص البنود التي يتضمنها بطريقة دقيقة باستخدام الأساليب الكيفية والأساليب الكمية، وتسمى الأساليب الكمية المستخدمة في فحص بنود الاختبار بتحليل البنود، حيث تتعلق هذه الخصائص بمؤشر صعوبة البند، ومؤشر التمييز.

تحليل البنود Item analysis هو إجراء لقياس الخصائص المختلفة لبنود الاختبار، يساعدنا في تحديد البنود التي تكون سهلة للغاية أو شديدة الصعوبة للمفحوص، ويساعدنا في فهم الطريقة التي يميز بها البند بين المفحوصين ذوي الدرجات المنخفضة والمفحوصين ذوي الدرجات العالية، فهي إنها طريقة بسيطة نسبيًا لها تقليد طويل في القياس(Meyer, 2014) . وفي جوهره لا يعد تحليل البنود أكثر من مجرد حساب للمتوسطات والارتباطات، لكن سياق العملية الاختبارية يؤدي إلى تفسيرات محددة لهذه الإحصاءات.

تحليل البنود هو إجراء يتبع لفحص بعض الخصائص القياسية للبنود، ويتم التحقق من خصائص بنود الاختبار بواسطة مؤشر صعوبة البند ومؤشر تمييزه. تكمن أهمية تحليل البنود في المراجعة الفنية للبنود وتحسينها ليساهم كل بند ايجابيا فيما يقيسه الاختبار، ومساعدة القائمين ببناء الاختبارات على التعرف على جوانب الضعف في بعض البنود غير صالحة والإبقاء على البنود التي تفي بالغرض.

2. مؤشر صعوبة البند:

صعوبة البند هو متوسط درجة البند، فالبنود ذات الاختيار من متعدد، والصواب/ الخطأ، والبنود الأخرى التي تم تصحيحها على أنها صحيحة (درجة واحدة) أو خاطئة (0 درجة)، فإن صعوبة البند هي نسبة المفحوصين الذين أجابوا على البند بشكل صحيح، حيث يتراوح من 0 إلى 1، وعلى الرغم من تسميته "صعوبة البند"، فانه تشير القيمة الكبيرة لهذه الإحصائية إلى بند سهل، وتشير القيمة الصغيرة إلى بند صعب. فعلى سبيل المثال، تشير صعوبة بند بمقدار 0.80 إلى أن 80% من المفحوصين أجابوا على البند بشكل صحيح، من ناحية أخرى تظهر صعوبة البند 0.1 أن 10٪ فقط من المفحوصين أجابوا على البند بشكل صحيح. فالبند الذي تقدّر صعوبته 0.1 يكون أكثر صعوبة من بند تقدّر صعوبته 0.8.

وبالتالي يمكن الحصول على صعوبة بنود الاختبار بإيجاد نسبة عدد الأفراد الذين أجابوا عن كل بند إجابة صحيحة، وكلما زادت هذه النسبة دل على سهولة البند، وكلما قلت دل على صعوبة البند. ويرمز لمؤشر صعوبة البند بالرمز $\boldsymbol{P}$ حيث:

$$\boldsymbol{P = \frac{Pi}{n}}$$

$\boldsymbol{Pi}$: عدد الأفراد الذين أجابوا إجابة صحيحة على البند

$\boldsymbol{n}$: عدد الأفراد

إن صعوبة البند متعدد الدرجات Polytomous item الذي يتم تصحيحه وفق أكثر من فئتين ترتيبيتين، هو ببساطة متوسط البند أو متوسط درجة البند، وهي تتراوح بين الحد الأدنى لدرجة البند والحد الأقصى لدرجة البند الممكنة. ويعتمد تفسير صعوبة البند متعدد الدرجات على الحد الأدنى والأقصى لدرجات البنود الممكنة. لذلك فإن البند متعدد التدريج المكون من 4 درجات، والذي تم تصحيحه 0 و1 و2 و3 درجات له متوسط يتراوح بين 0 و3 درجات. وكلما اقترب المتوسط من 0 زادت صعوبة البند (على سبيل المثال، كان من الصعب تحقيق أعلى فئة)، وكلما اقترب المتوسط من 3 كان البند أسهل (على سبيل المثال، كان من السهل الوصول إلى أعلى فئة). ويتم تقدير صعوبة البند بحساب المتوسط الحسابي:

$$\boldsymbol{P = \frac{\sum x_i}{n}}$$

من الممكن تحويل صعوبة البند بالنسبة للبند متعدد الدرجات إلى درجة صحيحة نسبة عن طريق قسمة متوسط البند على أقصى درجة ممكنة للبند. فعلى سبيل المثال، نفترض أن بنداً تم تصحيحه 0 و1 و2 و3 درجات، فان متوسط البند 2.35 يتوافق مع الدرجة الصحيحة للنسبة 2.35/ 3 = 0.78.

وكما هو الحال مع البنود الثنائية التصحيح، عندما يتم تحويل صعوبة البند متعدد الدرجات إلى مقياس 0–1، تشير القيم القريبة من 1 إلى بند سهل، والقيم القريبة من 0 بند صعب. ومع ذلك، لن نثق بأن 78% من المفحوصين أجابوا على هذا البند بشكل صحيح، يسمح لنا هذا التحويل فقط بالنظر إلى أن المفحوصين حصلوا في المتوسط على جزء كبير من الدرجات المتاحة.

وبالتالي يُعطى مؤشر الصعوبة في حالة البند ثنائي الإجابة (1 أو 0) بنسبة الأفراد الذين أجابوا إجابة صحيحة على البند. وفي حالة بند ذات متعددة الدرجات (مثلا 3، 2، 1، 0) فان مؤشر الصعوبة يُشار إليه بمتوسط الدرجات المحصلة في البند.

مؤشر صعوبة البند هو أيضا مؤشر للسهولة لأنه كلما ارتفعت قيمة معامل الصعوبة كلما اعتبر البند سهلا، بمعنى أن مؤشر السهولة يشير إلى نسبة عدد الأفراد الذين أجابوا إجابة خاطئة على البند، ويمكن أن يُعطى معامل السهولة انطلاقا من مؤشر الصعوبة وفقا للصيغة التالية:

$$q = 1 - P$$

يتراوح معامل الصعوبة بين (0) عندما يكون البند في غاية الصعوبة و(1) عندما يكون البند في غاية السهولة، لذلك فان معامل السهولة الذي يتراوح بين (0.40) و(0.60) يشير إلى أنه متوسط الصعوبة أو السهولة.

يعتبر مصطلح "صعوبة البند" منطقيًا في القياس التربوي حيث يكون للأسئلة إجابة صحيحة أو يمكن تصنيفها على درجات من الصواب. ففي القياس النفسي حيث يكون الهدف هو قياس اتجاه أو سمة شخصية، يكون المصطلح أقل ملاءمة. غالبًا ما تشتمل القياسات النفسية على مقاييس "ليكرت" Likert التي تطلب من الأفراد الاجابة على النحو التالي "أوافق بشدة" أو "أوافق" أو "لا أوافق" أو "لا أوافق بشدة". هنا لا توجد إجابة صحيحة لمثل هذا السؤال، أي إجابة مقبولة، ولا ينطبق مصطلح "صعوبة البند" على هذه الوضعية. بالنسبة لبنود "ليكرت" وأنواع البنود المماثلة يمكنك التفكير في متوسط البند على أنه مؤشر المصادقة على البند Index of item endorsability – إلى أي مدى يتم المصادقة على خيار الاستجابة الأعلى، وهذا المصطلح أكثر اتساقًا مع فكرة أن شخصًا ما يؤيد اتجاهاً معينًا، حيث تكون جميع الاتجاهات مقبولة ولا يكون أي منها "صحيحًا".

نشاط تدريبي:

فرضاً أنه تم تطبيق ثلاث (3) بنود اختبار لقياس القدرة على حل المشكلات على (10) أفراد، حيث أن البند1 يُصحّح وفق سلم خماسي (/**5** أي 5، 4، 3، 2، 1، 0)، والبند2 يُصحّح وفق سلم ثلاثي (/2 أي 2، 1، 0)، والبند3 يُصحّح وفق سلم ثنائي (/1 أي 1، 0). حيث توزعت درجاتهم على البنود الثلاثة كما هي موضحة في الجدول.

– أحسب المتوسط الحسابي، ومعامل صعوبة ومعامل سهولة البنود.

الأفراد	البند1 (/5)	البند2 (/2)	البند3 (/1)	المجموع
1	3	2	1	6
2	5	2	0	7
3	5	2	0	7
4	5	2	1	8
5	4	2	1	7
6	3	1	1	5
7	2	1	1	4
8	2	1	1	4
9	0	0	0	0
10	2	1	0	3
المتوسط الحسابي $\bar{x}$	**3.10**	**1.40**	**0.60**	**5.10**
مؤشر الصعوبة p	**0.62**	**0.70**	**0.60**	$\bar{p} = 0.64$
مؤشر السهولة q	**0.38**	**0.30**	**0.40**	$\bar{q} = 0.36$

$$P_1 = \frac{3.10}{5} = 0.62 \qquad P_2 = \frac{1.40}{2} = 0.70 \qquad P_3 = \frac{6}{10} = 0.60$$

$$q_1 = 1 - 0.62 = 0.38 \qquad q_2 = 1 - 0.70 = 0.30$$

$$q_3 = 1 - 0.60 = 0.40 \quad \text{أو} \quad q_3 = \frac{4}{10} = 0.40$$

$$\bar{p} = \frac{\sum P}{j} = \frac{0.62 + 0.70 + 0.60}{3} = 0.64$$

يشير P إلى درجة صعوبة كل بند، ويشير j إلى عدد البنود. ويُفضل قيمة $\bar{p}$ في الغالب على المتوسط الحسابي للاختبار لأنها تتأثر بنظام تصحيح البنود، لذا من الأفضل نشير إليه بمتوسط الصعوبة (0.64) الذي يكون قابلاً للمقارنة بالإشارة إلى المتوسط الحسابي المحصل عليه (5.10) من (8.00).

3. تصحيح صعوبة البند من أثر التخمين:

إذا تم حساب مؤشر صعوبة البند ذات إجابات اختيارية، فانه يجب الأخذ بعين الاعتبار احتمالية النجاح في البند دون المعرفة الفعلية بالإجابة الصحيحة. فعلى سبيل المثال البند ذات الاجابة القصيرة الذي يكون معامل صعوبته 0.75 يمكن أن نعتبره سهل نسبياً، ولكن لا يكون كذلك في

البند "الصواب– الخطأ" الذي مؤشر صعوبته 0.75، والذي تكون فيه احتمالية النجاح بشكل عشوائي 0.50، وبالتالي بالبند "صحيح–خطأ" يجب أن يعتبر أكثر صعوبة نسبياً من البند ذات الاجابة القصيرة.

يمكن أن يتأثّر تفسير مؤشر صعوبة البند بعاملين أساسيين، وهما: عدد الاجابات (أو البدائل) المقدمة، واحتمالية الاجابة على البند إجابة صحيحة بالتخمين. فالبنود التي تتطلب اختيار إجابة من إجابتين (صح– خطأ) أو اختيار إجابة من إجابات متعددة (اختيار من متعدد) تتأثر الدرجة التي يحصل عليها المفحوص بأثر التخمين Guessing effect.

يشير **أثر التخمين** إلى احتمال إجابة الفرد على البند إجابة صحيحة بالصدفة، بمعنى الاجابة على البند إجابة صحيحة دون معرفة الإجابة الصحيحة فعلياً، حينها يجب تصحيح معامل صعوبة البند من أثر التخمين بواسطة الصيغة التالية:

$$P' = P - \left[\frac{1 - P}{M - 1}\right]$$

P' : معامل الصعوبة المصحح من أثر التخمين.

P : معامل الصعوبة قبل التصحيح.

M : عدد بدائل (أو اختيارات) الإجابة.

تصحيح مؤشر الصعوبة من أثر التخمين ليس ضروري للمقارنة بين مؤشرات صعوبة بنود الاختبار الذي يشتمل على أسئلة مماثلة من حيث عدد البدائل، على سبيل المثال عندما تكون كل الأسئلة ذات أربع (4) بدائل للإجابة، لأنه نعرف بأن احتمالية النجاح بالتخمين هي $\frac{1}{M}$ أي 0.25 بالنسبة لكل الأسئلة. في حين إذا كانت صيغ أسئلة الاختبار متنوعة ($M = 2.3.4.5$) فمن الضروري إجراء التصحيح حتى يمكن مقارنة صعوبة البنود انطلاقاً من قاعدة مشتركة.

نشاط تدريبي:

تم تطبيق على مجموعة مكونة من (10) أفراد ثلاث (3) بنود مختلفة من حيث عدد بدائل الاجابة، البند1 يتضمن بديلين "صحيح –خطأ"، والبند2 يتضمن 3 بدائل، والبند 3 يتضمن 5 بدائل، حيث توزعت درجات هؤلاء الأفراد على البنود الثلاث كما هي موضحة في الجدول التالي:

الأفراد	البند1 (الصواب– الخطأ)	البند2 (3 بدائل)	البند3 (5 بدائل)
1	1	1	1
2	1	1	1
3	0	0	1
4	0	0	1
5	1	0	0
6	1	1	1
7	0	0	0
8	1	1	0
9	1	1	0
10	1	1	0
مؤشر الصعوبة	**0.70**	**0.60**	**0.50**
مؤشر الصعوبة المصحح من أثر التخمين P'	**0.40**	**0.40**	**0.38**
الفرق $P - P'$	**0.30**	**0.20**	**0.13**

$$P'_1 = 0.70 - \left[\frac{1-0.70}{2-1}\right] = 0.40$$

$$P'_2 = 0.60 - \left[\frac{1-0.60}{3-1}\right] = 0.40$$

$$P'_3 = 0.50 - \left[\frac{1-0.50}{5-1}\right] = 0.38$$

كلما زادت عدد بدائل أو اختيارات الإجابة على البند كلما زادت احتمالية أثر التخمين، حيث يكون في البند ذات الاجابة الثنائية (صحيح –خطأ) احتمال التخمين الصحيح للإجابة على البند هي 0.50، في حين تكون احتمالية الاجابة الصحيحة على البند ذات خمس (5) بدائل يكون 0.20.

4. تباين البند:

بينما يخبرنا معامل الصعوبة إلى أي مدى يكون البند ناجحاً، فان تباين البند يخبرنا إلى أي مدى تكون نتائج هذا البند مشتتة أم لا. في حالة البنود التي يتم تصحيحها على مقياس مستمر أو متعددة الدرجات يتم حساب التباين باستخدام الصيغة المعتادة، وهي كما يلي:

$$S^2 = \frac{n\sum x^2 - (\sum x)^2}{n\,(n-1)}$$

نشاط تدريبي:

تم تطبيق بندين تم تصحيحهما وفق مقياس متعدد الدرجات من 0 إلى 5 درجات على مجموعة من الأفراد، فتوزعت درجاتهم كما يلي:

الأفراد	**1**	**2**	3	4	5	6	7	8	9	10
البند1	0	3	1	0	2	1	2	1	0	2
البند2	1	2	2	4	5	0	3	3	2	0

– أحسب تباين درجات البندين 1 و2.

الحل:

$$S_1^2 = \frac{10 \times 24 - (12)^2}{10\,(10-1)} = 1.07$$

$$S_2^2 = \frac{10 \times 72 - (22)^2}{10\,(10-1)} = 2.62$$

أما في حالة البنود المصححة بشكل ثنائي (0 أو 1) فان صيغة مبسطة تسمح بحساب التباين بسهولة، ويتم الحصول عليه من خلال ناتج ضرب نسبة الأفراد الذين نجحوا في البند p ونسبة الأفراد الذين فشلوا في البند q (الذي يساوي $1 - p$):

$$S^2 = p \times q$$

نشاط تدريبي:

طُبّقت ثلاث (3) بنود ذات اختيار من متعدد تم تصحيحها (1 أو 0) على عينة متكونة من (10) أفراد، توزعت درجاتهم كما يلي:

q	p	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**الأفراد**
0.6	0.4	0	1	0	1	1	0	0	0	1	0	**البند1**
0.3	0.7	1	1	0	1	1	0	1	0	1	1	**البند2**
0.8	0.2	0	1	0	0	1	0	0	0	0	0	**البند3**

– أحسب تباين درجات البنود.

الحل:

$$S_1^2 = 0.4 \times 0.6 = 0.24$$
$$S_2^2 = 0.7 \times 0.3 = 0.21$$
$$S_3^2 = 0.2 \times 0.8 = 0.16$$

القيمة القصوى لتباين البند الذي تم تصحيحه بطريقة ثنائية (1، 0) هو 0.25، وهذه القيمة ممكنة عندما يكون معامل صعوبة البند 0.5. وبالتالي يمكن أن يصل تشتت البند الثنائي التصحيح إلى الحد الأقصى عندما يكون نصف الأفراد قد نجحوا أو فشلوا في البند.

5. مؤشر تمييز البند:

تمييز البند Item discrimination هو المدى الذي يميز فيه البند بين المفحوصين الذين حصلوا على درجات مختلفة في الاختبار، فإذا كان التمييز مرتفعًا يمكن للبند أن يميّز بسهولة بين المفحوصين الذين حصلوا على درجات اختبار متشابهة، ولكن ليست متطابقة (Meyer, 2014) ومن ناحية أخرى إذا كان التمييز منخفضًا يمكن للبند أن يميّز فقط بين المفحوصين الذين حصلوا على درجات اختبار مختلفة تمامًا.

تعتبر خاصية التمييز أيضاً من الخصائص الهامة التي يجب أن تتوفر في بنود الاختبارات، وتعني مدى قياس بنود الاختبار للفروق الفردية، فقدرة البند على التمييز هو قدرته على التمييز بين الأفراد الأكثر مهارة والأفراد الأقل مهارة، بمعنى التمييز بين الأفراد الذين تحصلوا على مستوى معين من المهارة والأفراد الذين لم يتحصلوا.

يمكن أن نصنّف مؤشرات التمييز إلى نوعين؛ مؤشرات التمييز معيارية المرجع التي تهتم بمقارنة أداء الأفراد في المجموعة التي ينتمون إليها، ومؤشرات التمييز المحكية المرجع التي تهتم بمقارنة أداء الأفراد بمحك أداء خارجي.

يتطلب تقدير مؤشر التمييز أن يتوفر محك نستند إليه في تحديد الأفراد الضعفاء والأفراد الأقوياء في قدرة معينة، وتتوفر العديد من الطرق التي تستخدم في هذا الغرض، وتعتبر الدرجة الكلية التي يحصل عليها الأفراد في الاختبار من محكاً مناسباً، وتستخدم في تقدير معامل تمييز البند عدة طرق من بينها:

1.5. مؤشر تمييز المقارنة الطرفية:

تعتمد هذه الطريقة على تقسيم درجات الاختبار إلى قسمين متمايزين، ويمثل أحد القسمين المجموعة التي حصلت على أعلى الدرجات والقسم الآخر يمثل المجموعة التي حصلت على أقل الدرجات في الاختبار نفسه، وقد حدد Kelley عند تحليل بنود الاختبار نسبة (27%) من الأفراد في كل من المجموعتين الطرفيتين، ولكن تتطلب العملية تطبيق بنود الاختبار على عينة كبيرة للحصول على درجات متسقة مع عينة إلى أخرى (Laveault & Grégoire, 2014) .

لتقدير مؤشر التمييز فان الطريقة سهلة، بالنسبة للبنود ثنائية التصحيح يتم حسابها على أنها الفرق بين صعوبة البند في الفئة الأعلى 27% وصعوبة البند في الفئة الأدنى 27% من المفحوصين، وتشير القيم الكبيرة لمؤشر التمييز إلى أن البنود أسهل كثيرًا بالنسبة للمفحوصين الحاصلين على أعلى الدرجات مقارنة بالمفحوصين ذوي الدرجات المنخفضة.

يتضمن مؤشر التمييز **D** تفسير بديهي، ومن السهل حسابه يدويًا، وربما يكون القيد الرئيسي لمؤشر ***D*** هو أنه لا يستخدم جميع البيانات المتاحة، لأنه يتم استبعاد 46% من المفحوصين في الوسط من الحساب. ويشير معامل التمييز ببساطة إلى الفرق بين صعوبة البند في المجموعة العليا P_+ وصعوبة البند في المجموعة الدنيا **P_**.

يحسب مؤشر التمييز بالصيغة التالية:

$$D = P_+ - P_-$$

تتراوح معاملات التمييز بين (−1) و (+1)، وتشير أي قيمة إلى دلالة معينة، حيث اقترح الباحثان (Ebel & Frisbie, 1991) قيماً مرجعية لتفسير معاملات التمييز:

0.40 أو أكثر: البند مميز جدا

0.30 – 0.39 : بند مميز

0.20–0.29 : بند أقل تمييزا

0.10–0.19 : بند محدود يجب تحسينه

أقل من 0.10 : بند لا فائدة منه في الاختبار.

نشاط تدريبي:

نفترض أننا طبقنا اختبارا مكوناً من مجموعة من البنود على عينة تكونت من (33) فرداً، حيث تم قسمة عينة الأفراد إلى فئتين متباعدتين (27%) وفقاً لأدائهم الكلي على الاختبار (فئة عليا وفئة دنيا)، وبعد فحص درجاتهم على 5 بنود جاءت كما يلي:

البنود	مجموعة عليا (9 أفراد)	مجموعة دنيا (9 أفراد)	مؤشر الصعوبة P	مؤشر التمييز D
1	1 1 1 0 1 1 0 1 1	1 0 0 0 1 0 0 0 1	$\frac{10}{18} = 0.56$	$\frac{4}{9} = 0.44$
2	1 1 1 1 0 1 1 1 1	1 0 1 1 0 1 1 1 1	$\frac{15}{18} = 0.56$	$\frac{1}{9} = 0.11$
3	0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0	$\frac{1}{18} = 0.06$	$\frac{-1}{9} = -0.11$
4	1 0 1 0 0 0 1 0 0	0 0 0 0 1 0 0 0 0	$\frac{4}{18} = 0.22$	$\frac{2}{9} = 0.22$
5	1 1 1 0 1 1 0 1 1	0 1 0 0 0 1 0 0 0	$\frac{9}{18} = 0.50$	$\frac{5}{9} = 0.56$

يمكن أن يتم تقدير مؤشرات تمييز البنود بطريقة مباشرة بطرح عدد الأفراد الذين أجابوا إجابة صحيحة على البند في المجموعة الدنيا من الذين أجابوا إجابة صحيحة في المجموعة العليا، وقسمة ذلك النتائج على عدد الأفراد الكلي للمجموعتين. فعلى سبيل المثال في البند1 تم طرح 7 من 3 (أي $\mathbf{4 = 3 - 7}$) وقسمة الناتج على 9 (أي $\frac{}{9} = \mathbf{0.44}$).

كما يمكن استخدام الطريقة الثانية التي تم التعبير عنها بصيغة مؤشر التمييز $(\boldsymbol{D = P_+ - P_-})$ من خلال تقدير مؤشر الصعوبة في المجموعة العليا ومؤشر الصعوبة في المجموعة الدنيا، ثم طرح مؤشر صعوبة البند في المجموعة الدنيا من مؤشر الصعوبة في المجموعة العليا، كما يلي:

$\boldsymbol{P_+ = \frac{7}{9} = 0.78}$	$\boldsymbol{P_- = \frac{3}{9} = 0.33}$	$\boldsymbol{D = 0.78 - 0.33 = 0.45}$
$\boldsymbol{P_+ = \frac{8}{9} = 0.89}$	$\boldsymbol{P_- = \frac{7}{9} = 0.78}$	$\boldsymbol{D = 0.89 - 0.78 = 0.11}$
$\boldsymbol{P_+ = 0}$	$\boldsymbol{P_- = \frac{1}{9} = 0.11}$	$\boldsymbol{D = 0 - 0.11 = -0.11}$

$P_+ = \frac{3}{9} = 0.33$	$P_- = \frac{1}{9} = 0.11$	$D = 0.33 - 0.11 = 0.22$
$P_+ = \frac{7}{9} = 0.78$	$P_- = \frac{2}{9} = 0.22$	$D = 0.78 - 0.22 = 0.56$

توضح النتائج بأن البند رقم1 ($D = 0.45$) والبند رقم5 ($D = 0.56$) مميّزين جداً مميزين جداً لأن مؤشر تمييزهما أكبر من 0.40. وأن البند رقم 4 ($D = 0.22$) أقل تمييزاً لأن مؤشره يتراوح بين (0.20-0.29).

في حين أن البند رقم 2 ($D = 0.11$) محدود يتطلب تحسينه لأن مؤشر تمييزه يتراوح بين (0.10- 0.19)، أما البند رقم3 ($D = -0.11$) الذي جاء أقل من (0.10) فلا فائدة منه في الاختبار يتطلب حذفه.

2.5. مؤشر التمييز الارتباطي:

التمييز بين البنود هو ارتباط بين درجة البند والدرجات الكلية للاختبار، ولهذا السبب غالباً ما يطلق عليه الارتباط الكلي للبند، ومعامل الارتباط "بيرسون" هو النوع الأساسي من الارتباط المتضمن في تحليل البند، ويمكن تطبيقه على البنود الثنائية والمتعددة الدرجات، ويكون له اسم مختلف عند حسابه بين بند ثنائي والدرجات الكلية للاختبار. في هذه الحالة يُشار إليه بشكل أكثر تحديدًا على أنه "معامل الارتباط الثنائي التسلسل الخاص"، حتى أن هناك معادلة مبسطة لمعامل الارتباط الثنائي المتسلسل الخاص، لكنها ليست إلاّ معامل ارتباط "بيرسون".

يمكن تقدير تمييز بنود الاختبار بإيجاد معامل الارتباط بين درجات كل بند والدرجة الكلية في الاختبار، تفترض هذه الطريقة توزيع الدرجات على متغير متصل، حيث يكون لدينا متغيران أحدهما ثنائي الدرجة (درجة البند تكون 1 أو 0)، والآخر يكون متغير متصل يشتمل على الدرجة الكلية للاختبار. يمكن استخدام معامل الارتباط ثنائي المتسلسل الخاص (الحقيقي) أو معامل الارتباط ثنائي المتسلسل.

من بين قيود التي تحدّ من استخدام معامل الارتباط "بيرسون" هو حساسيته لتوزيع درجات قدرة المفحوص، فإذا اشتمل الاختبار على فاحصين ذوي قدرة عالية، فسيكون له قيمة تمييز البند مختلفة كثيراً عن نفس الاختبار الذي يتم إجراؤه على مجموعة من المفحوصين ذوي قدرة

منخفضة. وللتغلب على هذا القيد، يمكننا أن نفترض أن المتغير الكامن يتوزع بشكل طبيعي يكمن وراء درجة البند (Meyer, 2014) .

يُشار إلى العلاقة بين درجة البند الكامن ومجموع درجات الاختبار على أنها معامل الارتباط ثنائي المتسلسل Biserial correlation عندما يتم تطبيقها على البنود الثنائية، ويُشار إليه كمعامل الارتباط المتسلسل المتعدد Polyserial correlation عند استخدامه مع البنود متعددة الدرجات.

جميع معاملات الارتباط بين البند والدرجة الكلية لها التفسير نفسه، وتعني معاملات الارتباط الإيجابية بين البند والدرجة الكلية أن المفحوصين الحاصلين على درجات عالية في الاختبار يميلون إلى الحصول على درجة صحيحة على البند، ويميل المفحوصون ذوو الدرجات المنخفضة في الاختبار إلى الاجابة بطريقة خاطئة. وتشير القيم الإيجابية العالية للارتباط بين البند والدرجة الكلية إلى مقدار كبير من التمييز، والقيم القريبة من 0 تعكس تمييزًا ضئيلًا أو معدومًا.

يمكن أن يأخذ معامل الارتباط بين البند والدرجة الكلية قيمًا سالبة، لكن مثل هذه النتيجة قد تشير إلى وجود مشكلة، وهذا يعني أن المفحوصين ذوي الدرجات المنخفضة يميلون إلى الاجابة الصحيحة على البند. فإذا لوحظ وجود تمييز سلبي للبند، فيجب التحقق من البند نفسه وتصحيحه، ربما قدمت رمز إجابة خاطئة، أو قد يكون البنود مكتوبًا بصيغة معكوسة، إذا كان مفتاح الإجابة صحيحًا، تشير قيمة تمييز البند السالبة إلى وجود مشكلة خطيرة في البند.

1.2.5. معامل الارتباط الثنائي الخاص:

يستخدم في إيجاد درجة الارتباط بين درجات البند والدرجات الكلية في الاختبار بواسطة تقسيم توزيع الدرجات الكلية في الاختبار إلى مجموعتين، تمثل إحداهما المجموعة العليا وتمثل الأخرى المجموعة الدنيا، وتشير القيمة الناتجة عن معامل تمييز البند إلى مدى اتساق درجات كل بند مع درجات الاختبار ككل.

يتم تقدير معامل الارتباط الثنائي الخاص بواسطة الصيغة التالية:

$$r_{pbis} = \frac{\overline{X}_+ - \overline{X}}{S_x}\sqrt{\frac{P}{q}}$$

$\overline{x}_+$: متوسط توزيع الدرجات الكلية للمجموعة التي أجابت إجابة صحيحة عن البند.

$\overline{x}$: متوسط توزيع الدرجات الكلية في الاختبار.

S_x : الانحراف المعياري للدرجات الكلية في الاختبار.

P : نسبة عدد الأفراد الذين أجابت إجابة صحيحة على البند (مستوى صعوبة البند).

q : نسبة عدد الأفراد الذين أجابوا إجابة خاطئة على البند (مستوى سهولة البند).

نشاط تدريبي:

قام أحد الفاحصين بتطبيق اختبار لقياس القدرة على حل المشكلات على عينة من الأطفال، حيث توزعت درجاتهم على الاختبار وعلى أحد بنوده كما يلي:

الأفراد	1	2	3	4	5	6	7	8	9	10
درجات البند	1	1	0	1	0	1	1	0	1	0
درجات الاختبار	7	8	6	5	7	8	9	6	9	5

– ما مدى قدرة البند على التمييز.

الحل:

$$\bar{X} = \frac{70}{10} = 7 \qquad \bar{X}_+ = \frac{46}{6} = 7.67$$

$$p = \frac{6}{10} = 0.6 \qquad q = \frac{4}{10} = 0.4$$

$$S_x = \sqrt{\frac{n\Sigma x^2 - (\Sigma x)^2}{n\,(n-1)}} = \sqrt{\frac{10 \times 510 - (70)^2}{10\,(10-1)}} = 1.49$$

$$r_{pbis} = \frac{\bar{X}_+ - \bar{X}}{S_x}\sqrt{\frac{P}{q}} = \frac{7.67 - 7}{1.49}\sqrt{\frac{0.6}{0.4}} = 0.55$$

بلغ معامل الارتباط بين البند والدرجة الكلية للاختبار (0.55)، وبالتالي فان البند مميّز جداً أي متسق مع الدرجة الكلية للاختبار.

2.2.5. معامل الارتباط الثنائي:

يستخدم معامل الارتباط الثنائي أيضاً في إيجاد درجة الارتباط بين درجات البند ودرجات الاختبار ككل، ولكن يكمن الفرق بينه وبين معامل ثنائي الحقيقي أنه يستخدم في حالة تقسيم المتغير تقسيما اصطناعيا (غير حقيقي. فإذا اعتبرنا المتغير (ناجح– راسب) في الإجابة على بنود الاختبار متغيرا متصلا يمثل قدرة أو سمة معينة تنطوي عليها درجات البنود فان نقطة التقسيم

تعتمد على صعوبة البنود، حينها يفضل استخدام معامل الارتباط ثنائي المتسلسل. يتم تقدير معامل الارتباط ثنائي المتسلسل وفقا للصيغة:

$$r_{bis} = \frac{(\bar{X}_+ - \bar{X})}{S_x} \frac{P}{Y}$$

تتضمن صيغة معامل الارتباط ثنائي نفس رموز صيغة معامل الارتباط ثنائي الخاص، باستثناء قيمة **Y** التي ترمز إلى إحداثيي المنحنى الاعتدالي المعياري عند الدرجة الزائية المرتبطة بقيمة P عوضاً عن قيمة معامل السهولة q.

على سبيل المثال إذا تحصلنا على معامل صعوبة البند ($p = 0.60$)، وعلى اعتبار أن معامل الصعوبة أكبر من 0.50 بالعودة إلى جدول المنحنى الاعتدالي المعياري فإننا سوف نختار عمود الاحتمالات (المساحات إلى اليسار) للدرجات الزائية الموجبة، وبالتالي قيمة المساحة الأقرب إلى (0.60) هي (0.55) والإحداثية المرتبطة بها تساوي (0.3867).

نشاط تدريبي:

إذا أخذنا بيانات النشاط التدريبي السابق نفسها، والموضحة في الجدول التالي:

الأفراد	1	2	3	4	5	6	7	8	9	10
درجات البند	1	1	0	1	0	1	1	0	1	0
درجات الاختبار	7	8	6	5	7	8	9	6	9	5

الحل:

$\bar{X} = 7$ $\quad\quad$ $\bar{X}_+ = 7.67$

$p = 0.6$ $\quad\quad$ $q = 0.4$

$S_x = 1.49$ $\quad\quad$ $Y = 0.3867$

$$r_{bis} = \frac{(\bar{X}_+ - \bar{X})}{S_x} \frac{P}{Y} = \frac{7.67 - 7}{1.49} \times \frac{0.6}{0.387} = 0.70$$

معامل الارتباط الثنائي يساوي تقريباً (0.70)، وهذا يعني بأن البند يسمح بالتمييز بشكل جيد بين الأفراد الأقوياء والأفراد الضعفاء، وهو يتعلق ببند يجب الاحتفاظ به إذا كان الغرض منه التمييز بين الأفراد.

أثبت Lord et Novick (1968) بأن معامل الارتباط الثنائي المحصل عليه يكون أكبر بـ 20% من معامل الارتباط الثنائي الخاص، حيث يمكن تحويل معامل الارتباط الثنائي الخاص إلى معامل الارتباط الثنائي كما يلي: (Cited in Laveault & Grégoire. 2014)

$$r_{bis} = \frac{\sqrt{Pq}}{Y} r_{Pbis}$$

وفي حالة المثالين السابقين يمكن أن نتأكد بأن العلاقة المعبر عنها في هذه المعادلة تسمح بإيجاد معامل الارتباط الثنائي بالاعتماد على معامل الارتباط الثنائي الخاص، وفي الواقع يمكن أن نستبدل القيم في المعادلة، ونجد أن:

$$r_{bis} = \frac{\sqrt{Pq}}{Y} r_{Pbis} = \frac{\sqrt{0.6 \times 0.4}}{0.387} \times 0.55 = 0.70$$

3.2.5. معامل الارتباط "فاي":

عندما نريد حساب معامل الارتباط بين البند وبند آخر يعدّ محكاً صادقاُ وثابتاُ (موثوق منه) يمكن استخدام معامل الارتباط "فاي" Phi، كما يمكن استخدام هذا المعامل في تحديد مدى استقرار درجات الاستجابات على البند نفسه ثنائية التصحيح (1، 0) للأفراد أنفسهم في موقف آخر أو ارتباط بنود ثنائية التصحيح ودرجات محك ثنائي مثل ناجح وراسب في برنامج معين.

ويكون استخدام معامل الارتباط "فاي" مناسباً أكثر عندما تكون المتغيرات ثنائية حقيقية، ويُشار إليه بالصيغة التالية:

$$\emptyset = \frac{P_{jk} - P_j P_k}{\sqrt{P_j q_j - P_k q_k}}$$

P_{jk} : النسبة المشتركة للأفراد الذين أجابوا على البندين j و K إجابة صحيحة.

P_j : إلى نسبة الأفراد الذين أجابوا على البند j إجابة صحيحة.

P_k : نسبة الأفراد الذين أجابوا على البند K إجابة صحيحة.

q_j : نسبة الأفراد الذين أجابوا على البند j إجابة خاطئة.

q_k : نسبة الأفراد الذين أجابوا على البند K إجابة خاطئة.

نشاط تدريبي:

تم تطبيق بندين تم تصحيحهما بطريقة ثنائية (1، 0) على عينة مكونة من (10) أفراد، حيث أن البند رقم 1 نريد التحقق من اتساقه مع البند رقم 2 الذي يعدّ محكاً ملائماً (صادقاً وثابتاً)، ويقيس السمة نفسها.

البند 2 \ البند 1	0		1		
0	3	b	1	a	0
1	2	d	4	c	1

الحل:

$$P_{jk} = \frac{4}{10} = 0.4$$

$$P_j = \frac{5}{10} = 0.5$$

$$P_k = \frac{6}{10} = 0.6$$

$$\emptyset = \frac{0.4 - 0.5 \times 0.6}{\sqrt{0.5 \times 0.5 \times 0.6 \times 0.4}} = 0.41$$

معامل الارتباط بين البند رقم 1 والبند رقم 2 المحكي منخفض يقدّر بـ (0.41)، وهذا يعني أن العلاقة بين البند1 والبند المحكي ضعيفة مما يشير إلى عدم اتساق البند مع المحك.

المحاضرة السادسة

تحليل البنود محكية المرجع

الأهداف:

- يتعرّف الطالب على استخدامات أساليب تحليل البنود محكية المرجع.
- يقدّر الطالب معاملات التمييز محكية المرجع.
- يفسّر الطالب معاملات التمييز محكية المرجع وفق معايير محددة.

1. مؤشرات التمييز محكية المرجع:

مؤشرات التمييز التي تناولناها (طريقة المقارنة الطرفية، والطرق الارتباطية؛ معامل الارتباط الثنائي، ومعامل الارتباط الثنائي الخاص، ومعامل فاي) تعتمد في مبدئها على التمييز بين الأفراد الضعفاء والأقوياء في السمة أو القدرة المراد قياسها بمقارنة الأفراد بالمجموعة التي ينتمون إليها، بمعنى أن المقارنة تكون معيارية المرجع أو الجماعة بغض النظر عن هدف الاختبار أو بغض النظر عن المحك الخارجي للاختبار.

ولكن في إطار بيداغوجيا الاتقان أو في إطار التقويم التكويني اللذان يندرجان ضمن القياس المحكي المرجع الذي يهتم بتمييز أداء الفرد بالهدف المطلوب أو بمحك اتقان خارجي دون الاهتمام بالتمييز بين أفراد المجموعة فيما بينهم، فانه لا ننتظر من الاختبار أو أي أداة قياس أخرى التمييز فقط بين الأفراد، بل بالعكس نريد معرفة إذا كان البند يسمح بالتمييز بين الأفراد الذين يتقنون والأفراد الذين لا يتقنون الهدف عند عتبة نجاح محددة مسبقا.

وتوجد العديد من المؤشرات أو المعاملات لتقدير قدرة البنود على التمييز، من بينها؛ مؤشر الحساسية للتعليم (التدريب) $\boldsymbol{S}$ ، ومؤشر التمييز في عتب الاتقان $\boldsymbol{B}$ ، ومؤشر التوافق "فاي" $\boldsymbol{Phi}$، مؤشر التوافق المرجعي $\boldsymbol{HS}$.

1.1. مؤشر الحساسية للتعليم:

يعدّ مؤشر الحساسية للتعليم (أو التدريب) للبند أساساً مؤشراً لجودة البند بين المفحوصين الذين دُرّسوا والمفحوصين الذين لم يُدرّسوا، حيث اقترح "كوكس وفارجاس" (Cox & Vergas, 1966)

طريقة تعتمد على اختبار المجموعة نفسها قبل وبعد التعليم بما يُعرف بمؤشر الحساسية للتعليم أو التدريب (Crocker & Algina, 2006) .

البنود أكثر فائدة في القياس المحكي هي البنود الأكثر حساسية للتعليم أو التدريب، فاذا كان التعليم مفيداً فان مستوى صعوبة هذه البنود يجب أن تتغير بشكل كبير (Laveault & Grégoire, 2014). يستخدم مؤشر الحساسية للتعليم في تحديد البنود أكثر تأثّرا بالتعليم، ويتم حسابه عن طريق الفرق بين مستوى صعوبة البند بعد التعليم (P_{post}) ومستوى صعوبة البند قبل التعليم ($\mathbf{P_{pre}}$):

$$S = P_{post} - P_{pre}$$

حيث ترمز P_{post} إلى نسبة الأفراد الذين أجابوا إجابة صحيحة عن البند في الاختبار البعدي، وترمز P_{pre} إلى نسبة الأفراد الذين أجابوا إجابة صحيحة عن البند في الاختبار القبلي، حيث تتراوح قيم S بين (-1) و$(+1)$ مع تفضيل القيم الموجبة العالية.

كلما ارتفع الفرق بين صعوبة البند بعد التعليم وصعوبة البند قبل التعليم كلما كان البند أكثر تمييزا، وكلما قل الفرق بينهما كان البند أقل فائدة لأنه نجاح البند أكثر قبل التعليم من نجاحه بعد التعليم. وإذا كانت قيمة معامل الحساسية S منعدمة أو سالبة فإنها تفسّر بطريقتين سواء:

- البند غير متوافق لأنه لا يدخل ضمن نطاق التعليم.
- التعليم ليس له أي أثر على نجاح الأفراد.

نشاط تدريبي:

طُبّق اختبار للقدرة على التفكير المنطقي يشتمل على 05 مهمات ذات (05) بدائل على عينة مكونة من (20) فرداً قبل التعليم وبعد التعليم، فتوزعت درجاتهم على المهمة 1 كما يلي:

الأفراد	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
قبل	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	0	**0**	0	**0**	**1**	**0**
بعد	**1**	**1**	1	**1**	0	**1**	**1**	**0**	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**

- ما مدى حساسية البند للتعليم.

الحل:

$$P_{pre} = \frac{7}{20} = 0.35$$

$$P_{post} = \frac{16}{20} = 0.80$$

$$S = P_{post} - P_{pre} = 0.80 - 0.35 = 0.45$$

بلغ معامل الحساسية للتعليم (0.45) وهو مرتفع، حيث أنه نجح (35%) من الأفراد في الاجابة على البند إجابة صحيحة قبل التعليم، في حين ارتفعت نسبتهم إلى (80%) بعد التعليم بفارق (45%). مما يدل على حساسية البند للتعليم، فالبند له قدرة تمييزية.

2.1. مؤشر عتبة الاتقان:

قدم Brennan (1972) مؤشراً لحساب تمييز البند في عتبة الاتقان، وهذا المؤشر **B** مكافئ للمؤشر D باستثناء أن المجموعات العليا والدنيا عُوّضت بالمجموعة التي بلغت عتبة الاتقان والمجموعة التي لم تبلغ عتبة الاتقان في درجة للاختبار (Laveault & Grégoire, 2014) . يمكن حساب مؤشر التمييز في عتية الاتقان B بالطريقة التالية:

$$B = P_M - P_{NM}$$

حيث تشير P_M إلى مؤشر صعوبة البند للأفراد الذين بلغوا عتبة الاتقان في الاختبار ككل. وتشير P_{NM} إلى مؤشر الصعوبة للأفراد الذين لم يبلغوا عتبة الاتقان في الاختبار ككل.

يستخدم معامل Brennan لحساب الفرق بين نسبة نجاح البند للأفراد الذين بلغوا عتبة الاتقان في الاختبار ككل، ونسبة النجاح للأفراد الذين لم يبلغوا عتبة الاتقان. تتراوح قيمة B بين (−1) و(+1)، حيث يشير المؤشر السالب على أن البند لا يميز بين مجموعة المتقنين وغير المتقنين في الاختبار، بينما يشير المؤشر الموجب إلى أن البند مميز أي أن نسبة الأفراد في مجموعة الاتقان نجحوا في البند أفضل من المجموعة غير المتقنة (Crocker & Algina, 2006) .

نشاط تدريبي:

تم تطبيق اختبار على عينة تكونت من (50) فرداً، فتم من خلال نتائج الأفراد تحديد مجموعة متقنة للاختبار وعينة غير متقنة في الاختبار وفقاً لدرجة قطع محدّدة بـ (80%)، ولغرض التحقق من قدرة أحد بنود الاختبار على التمييز في عتبة الاتقان تم تنظيم البينات في الجدول التالي:

	الاختبـــار	
	متقن	غير متقن
البنـــد ناجح	$b = 15$	$a = 7$
غير ناجح	$d = 4$	$c = 24$

الحل:

$$P_M = \frac{b}{b+d} = \frac{15}{15+9} = 0.79$$

$$P_{NM} = \frac{a}{a+c} = \frac{7}{7+24} = 0.23$$

$$B = P_M - P_{NM} = 0.79 - 0.23 = 0.56$$

توضح النتائج بأن البند يميّز جداً عند عتبة الاتقان ($B = 0.56$) لأنه يوجد 56% من الأفراد أكثر في "عينة الاتقان" نجحوا في البند بالمقارنة مع المجموعة "غير المتقنة"، فمن المؤكد بأنه بند مناسب للتمييز عند عتبة الإتقان.

3.1. مؤشر الارتباط "فاي":

يستخدم معامل فاي Phi لمعرفة درجة التوافق في تصنيف الأفراد بين البند والاختبار، حيث يتم تطبيق الاختبار على مجموعة من الأفراد ثم يتم تصنيفهم وفقاً لدرجة قطع أو درجة فاصلة معينة تمثل مستوى الاتقان، وتتحدّد قدرة البند على التمييز بقدرتها على التمييز بين الأفراد عند درجة قطع (درجة فاصلة) على الدرجة الكلية للاختبار. حيث يتم تنظيم البيانات في جدول رباعي تحدّد فيه أربع خلايا (**a . b . c. d**).

	الاختبـار	
	متقن	غير متقن
البنـد **1**	b	a
0	d	c

يتم حساب معامل التوافق "فاي" وفقاً لخلايا الجدول الرباعي كما يلي:

$$\emptyset = \frac{\boldsymbol{bc - ad}}{\sqrt{(\boldsymbol{a+b})(\boldsymbol{c+d})(\boldsymbol{a+c})(\boldsymbol{b+d})}}$$

$\boldsymbol{a}$ إلى عدد الأفراد غير المتقنين الذين أجابوا على البند إجابة صحيحة.

$\boldsymbol{b}$ عدد الأفراد المتقنين الذين أجابوا على البند إجابة صحيحة.

$\boldsymbol{c}$ عدد الأفراد غير المتقنين الذين أجابوا إجابة خاطئة على البند.

$\boldsymbol{d}$ عدد الأفراد المتقنين الذين أجابوا إجابة خاطئة على البند.

نشاط تدريبي:

تم تطبيق اختبار لتقييم مهارة الكتابة على عينة تكونت من (26) فرداً، حيث تم تقسيم العينة إلى مجموعة بلغت درجة القطع ومجموعة لم تبلغ درجة القطع، وفي أحد بنوده تم تحديد الأفراد الذين أجابوا إجابة صحيحة والأفراد الذين أجابوا إجابة خاطئة في كلاً من المجموعتين، كما يلي:

	الاختبار	
	غير متقن	متقن
البند 1	$a = 4$	$b = 8$
البند 0	$c = 12$	$d = 2$

– ما مدى قدرة البند على التمييز.

الحل:

$$\emptyset = \frac{\boldsymbol{8 \times 12 - 4 \times 2}}{\sqrt{(\boldsymbol{4+8})(\boldsymbol{12+2})(\boldsymbol{4+12})(\boldsymbol{8+2})}} = \boldsymbol{0.54}$$

بلغ معامل فاي (0.54) وهو مؤشر يدل على قدرة البند على التمييز في درجة القطع للاختبار، حيث يعتبر البند جيداً وفق معامل "فاي" إذا بلغت قيمته (0.30) أو أكثر.

4.1. مؤشر التوافق المرجعي:

يهدف إلى معرفة احتمالية التوافق بين الاجابة على بند معين (صحيحة وخاطئة) ومستوى التمكن أو الاتقان (متقنين وغير متقنين) في الاختبار، وتعتمد هذه الطريقة على تحديد عدد الأفراد

المتقنين الذين أجابوا على البند إجابة صحيحة وعدد الأفراد غير المتقنين الذين أجابوا على البند إجابة خاطئة.

وبالتالي يعتمد على تطبيق الاختبار مرة واحدة على مجموعة واحدة من الأفراد، ويتم تصنيف أفراد هذه المجموعة إلى متقنين وغير متقنين بناء على مدى تحقيقهم لمستوى الاتقان المطلوب، وقد اقترح Harris & Subkoviak الصيغة التالية لحساب معامل التوافق المرجعي:

$$HS = \frac{A + D}{N}$$

A : عدد الأفراد المتقنين الذين أجابوا على البند إجابة صحيحة.

B: عدد الأفراد غير المتقنين الذين أجابوا على البند إجابة خاطئة.

N: عدد الأفراد.

ويتم حساب الحد الأدنى لمعامل التوافق المرجعي من خلال الجدول الثنائي، كما يلي:

الأداء على الاختبار

الأداء على البند		**متقن**	**غير متقن**
	1	*b*	*a*
	0	*d*	*c*

بالاعتماد على الصيغة التالية:

$$HS_{min} = \frac{(a + b)(a + c) + (c + d)(b + d)}{N^2}$$

a : عدد الأفراد المتقنين الذين أجابوا على البند إجابة صحيحة.

b : عدد الأفراد غير المتقنين الذين أجابوا على البند إجابة صحيحة.

c : عدد الأفراد المتقنين الذين أجابوا على البند إجابة خاطئة.

d: عدد الأفراد غير المتقنين الذين أجابوا على البند إجابة خاطئة.

N: عدد الأفراد.

يتراوح معامل التوافق المرجعي بين (0) و(+1)، ويعتبر البند جيّدا إذا كان الفرق بين الحد الأدنى لمعامل التوافق المرجعي ومعامل التوافق المرجعي أكبر أو يساوي 0.05.

نشاط تدريبي:

طُبّق اختبار لقياس مهارة التفكير العلمي على (30) فرداً، حيث تم تقسيمهم بناء نتائج الاختبار إلى مجموعة متقنة ومجموعة غير متقنة وفقاً لدرجة قطع (50)، وتقسيمهم إلى أفراد أجابوا بشكل صحيح (1) وأفراد أجابوا بشكل خاطئ (0) على أحد بنوده.

اختبار مهارة التفكير

درجة البند	متقن	غير متقن
1	12	6
0	4	8

– هل البند مميّز بين الأفراد المتقنين والأفراد غير المتقنين.

الحل:

$$\boldsymbol{HS = \frac{A + D}{N} = \frac{12 + 8}{30} = 0.60}$$

$$\boldsymbol{HS_{min} = \frac{(a + b)(a + c) + (c + d)(b + d)}{N^2}}$$

$$\boldsymbol{= \frac{(12 + 6)(12 + 4) + (4 + 8)(6 + 8)}{30^2} = \frac{456}{900} = 0.51}$$

يتضح الفرق بين معامل التوافق المرجعي (0.60) والحد الأدنى من معامل التوافق المرجعي (0.51) يساوي (0.09) وهو أكبر من (0.05)، وبالتالي فان البند يعتبر جيّدا.

المحاضرة السابعة

ثبات بنود الاختبار

الأهداف:

- يتعرف الطالب على أساليب تقدير معامل ثبات البنود.
- يقدّر الطالب معامل ثبات البند بالطريقة الصحيحة.

يعتمد ثبات الاختبار اعتمادا كليا على ثبات بنوده، ويشير ثبات البند إلى اتساق أو استقرار استجابات الأفراد على نفس البند باختلاف ظروف تطبيقه على نفس الأفراد، وتوجد طرق متعددة في تقدير ثبات بنود الاختبار يمكن تلخيصها فيما يلي:

1. طريقة التطبيق-إعادة التطبيق

1.1. معامل الارتباط "فاي":

أشرنا فيما سبق أنه يمكننا استخدام معامل "فاي" في تحديد درجة استقرار استجابات الأفراد ثنائية التصحيح في البند في موقف آخر، حيث تتطلب هذه الطريقة تطبيق البند على نفس الأفراد في فترتين مختلفتين وفقا للخطوات التالية:

- تطبيق بنود الاختبار على عينة من الأفراد.

- إعادة تطبيق نفس البنود على نفس عينة الأفراد بفاصل زمني.

- رصد استجابات الأفراد على كل بند من بنود الاختبار، حيث تسجل نتائج التطبيق الأول ونتائج التطبيق الثاني في شكل جدول تكراري.

-حساب معامل الارتباط "فاي" بين درجات الاستجابة في التطبيق والتطبيق الثاني.

يتم تقدير معامل ثبات البند وفقا للصيغة التالية:

$$\emptyset = \frac{\boldsymbol{bc - ad}}{\sqrt{(\boldsymbol{a+b})(\boldsymbol{c+d})(\boldsymbol{a+c})(\boldsymbol{b+d})}}$$

حيث ترمز الحروف: **a**، **b** ، **c** ، **d** إلى خلايا الجدول الرباعي، التي توضح توزيع إجابات الأفراد على البند في التطبيق الأول والتطبيق الثاني.

2.1. معامل الارتباط الرباعي:

ويمكن تطبيق معامل الارتباط الرباعي (r_{tet}) Tetrachoric correlation coefficient ويستخدم عندما يكون المتغيران ثنائيان، مثل "فاي" $\boldsymbol{phi}$ لكن يتطلب أن يكون كلا من المتغيرين متصلين فعلياً وموزعين بشكل طبيعي (اعتدالي). وبالتالي يتم تطبيقه على البيانات الترتيبية مقابل البيانات المتصلة لقياس مدى اتفاق المقيم للبيانات الثنائية؛ البيانات الثنائية هي بيانات ذات إجابتين محتملتين – عادة ما تكون صحيحة أو خاطئة.

يتم تقدير معامل الارتباط الرباعي وفق الصيغة التالية:

$$r_{tet} = \cos\left(\frac{180^\circ}{1+\sqrt{\frac{bc}{ad}}}\right)$$

حيث: **a، b ، c ، d** ترمز إلى خلايا الجدول الرباعي.

كلما ارتفعت قيمة معامل الارتباط الرباعي للبند بين الفترتين كلما دل على ثبات درجات البند، والعكس كلما انخفضت قيمة المعامل دل على عدم استقرار أو ثبات درجات البند.

نشاط تدريبي:

تم تطبيق بند على عينة تكونت من (300) فرد خلال فترتين منفصلتين، حيث تكون الاجابة صحيحة (1) أو خاطئة (0)، حيث تم الحصول على البيانات الملخصة في الجدول التالي:

		إعادة التطبيق	
		1	**0**
التطبيق	**1**	32 b	13 a
	0	23 d	17 c

ما مدى ثبات (استقرار) درجات البند بين الفترتين.

الحل:

$$r_{tet} = cos\left(\frac{180^\circ}{1+\sqrt{\frac{32 \times 17}{13 \times 23}}}\right)$$

$$= cos\left(\frac{180^\circ}{2.349}\right) = cos\,(76.628) = 0.23$$

معامل الارتباط الرباعي بين درجات البند في الفترة ودرجات البند في الفترة الثانية منخفض قُدّر بـ (0.23)، وهو منخفض يعكس عدم ثبات (عدم استقرار) درجات البند بين الفترتين.

2. طريقة الاحتمال المنوالي:

تستخدم هذه الطريقة لحساب ثبات بنود الاختبارات الموضوعية التي تعتمد إجاباتها على اختيار بديل واحد من بين بديلين أو من بدائل الاجابة (الحجامي، 2021). ويمكن حساب ثبات البند وفق هذه الطريقة من خلال الصيغة التالية:

$$r_t = \frac{M}{M-1}\,(P - \frac{1}{M})$$

M: عدد بدائل الإجابة.

P : أكبر تكرار نسبي من بين بدائل الاجابة.

حيث كلما زاد معامل الاحتمال المنوالي عن (0.50) يعتبر البند ثابتاً، وكلما انخفض عن القيمة يعتبر البند أقل ثباتاً.

نشاط تدريبي:

نفترض طبقنا اختباراً موضوعياً على عينة تكونت من (200) فرداً، وبعد الاطلاع على إجاباتهم على أحد البنود وجدنا بأنها توزعت على كل بديل من البدائل الأربعة، كما يلي:

(A = 13)، (B = 98)، (C = 38)، (D = 50).

– هل البند ثابت.

الحل:

البدائل	تكرار الاجابة	P
A	13	0.07
B	98	0.49
C	38	0.19
D	50	0.25
المجموع	**200**	**1.00**

$$r_t = r_t = \frac{M}{M-1}\left(P - \frac{1}{M}\right)$$

$$= \frac{4}{4-1}\left(0.49 - \frac{1}{4}\right) = 1.33 \times (0.49 - 0.25) = 0.32$$

معامل ثبات درجات البند منخفض بلغ (0.32)، وهو ضعيف يعتبر البند غير ثابت.

المحاضرة الثامنة

مفاهيم أساسية في ثبات الاختبار

الأهداف:

- يتعرف الطالب على الخطأ المعياري للقياس.
- يميّز الطالب بين مصادر أخطاء القياس.
- يتعرف الطالب على مفهوم الثبات من الناحية النظرية والإحصائية.

تحليل بنود الاختبار مرحلة ضرورية لاختيار البنود الملائمة التي تتمتع بمعاملات صعوبة وتمييز وثبات مقبولة، فإننا ننتقل إلى مرحلة أساسية أخرى تتضمن تصميم وإجراء دراسات الثبات للصيغة النهائية للاختبار، وذلك بالتحقق من مدى اتساق درجات الاختبار في مختلف ظروف تطبيقه أو تصحيحه. يعتبر الثبات Reliability قضية أساسية في القياس النفسي، وتتجلى أهميته بمجرد فهم معناه بشكل جيّد، وكما يوحي المصطلح فان أداة القياس الثابتة هي الأداة التي تعمل بطريقة متسقة ويمكن التنبؤ بها، ولكي يتمتع الاختبار بالثبات يجب أن تمثل الدرجات التي ينتجها حالة حقيقية من المتغير الذي يجري تقييمه (DeVellis, 2016) .

وقبل أن نتناول طرق تقدير ثبات الاختبار يجب أن نوضح مفهوم ذات أهمية في ظهور نظرية القياس، وفي تقدير ثبات درجات الاختبار، وهو خطأ القياس ومصادره.

1. مفهوم الخطأ المعياري للقياس:

تتأثر الدرجات الملاحظة للفرد بنوعين من الأخطاء؛ الأخطاء المنتظمة والأخطاء غير المنتظمة (العشوائية). تساهم الأخطاء المنتظمة بدرجة منتظمة في تباين درجات الاختبار وتكون جزءاً من تباين الدرجة الحقيقية، فدرجة الفرد الحقيقية في الاختبار هي الدرجة الناجمة عن جميع العوامل المنتظمة. أما الأخطاء العشوائية فهي الأخطاء التي لا ترتبط بأداء الفرد في الاختبار، وتؤثر بدرجات متفاوتة في معامل ثبات الاختبار، وهناك مصادر متعددة للأخطاء العشوائية، تتعلق بأداة القياس، وإجراءات تطبيق الاختبار وتصحيحه، والأفراد المختبرين (علام، 2000).

الخطأ المعياري للقياس عبارة عن كيان معياري للدرجات يجب أن يُؤخذ بعين الاعتبار أثناء تقييم جودة أدوات القياس (Cardinet, Sandra, & Pini, 2010) . حيث يعتبر من الجوانب المهمة

في تقدير دقة القياس لأنه يجعل من درجة الفرد الملاحظة تختلف في أغلب الأحيان عن درجته الحقيقية بسبب تأثَر الدرجة الملاحظة بمصادر أخطاء متعددة. فإذا استطعنا تحديد قيمة الأخطاء العشوائية التي أثَرت في الدرجة الملاحظة لكل فرد من الأفراد المختبرين، فانه يمكن إيجاد الانحراف المعياري لدرجات الخطأ، والقيمة الناتجة من تقدير الانحراف المعياري لدرجات الخطأ تسمى الخطأ المعياري للقياس (علام، 2000).

الخطأ المعياري للقياس كيان نظري غير قابل للملاحظة، ونعرف أن متوسطه يُمثَل الخطأ المعياري لمجموعة الأفراد لأننا لا نستطيع في الحقيقة ملاحظة درجة الخطأ لكل فرد من أفراد المجموعة إلا إذا أعيد تطبيق الاختبار على الفرد نفسه عدة مرات، وهذا غير ممكن، ولكن يمكن تقدير قيمته إذا عرفنا قيمة الانحراف المعياري للدرجات الملاحظة وقيمة معامل الثبات. ويرمز للخطأ المعياري للقياس بالرمز $\boldsymbol{SEM}$ أو بالرمز $(\boldsymbol{\sigma_E})$.

يمكن الحصول على قيمة الخطأ المعياري للقياس من قيمة معامل الثبات وفقا للصيغة التالية:

$$\boldsymbol{\sigma_E = \sigma\sqrt{1-\rho_{xx'}}}$$

$\boldsymbol{\sigma_E}$: الخطأ المعياري للقياس.

$\boldsymbol{\sigma}$: الانحراف المعياري لدرجات الاختبار.

$\boldsymbol{\rho_{xx'}}$: معامل الثبات المقدّر.

يعتبر تقدير الخطأ المعياري للقياس من المساهمات الأساسية التي ساعدت على ظهور نظريات القياس، فمن أجل تفسير دقة نتائج القياس فان الخطأ المعياري للقياس يساعد في الحصول على معلومات ملموسة وقابلة للتفسير مباشرة حول ثبات الدرجات المحصل عليها من أداة القياس.

نشاط تدريبي:

نفترض أننا بعد حساب معامل الثبات بطريقة "ألفا كرونباخ" للاتساق الداخلي بين بنود الاختبار، وجدنا بأنّه يقدّر بـ (0.58)، والتباين الكلي لدرجاته قد بلغ (2.57).

– ما مقدار الخطأ المعياري للقياس لدرجات الاختبار.

الحل:

$$\boldsymbol{\sigma = \sqrt{\sigma^2} = \sqrt{2.57} = 1.60}$$

$$\sigma_E = \sigma\sqrt{1-\rho_{xx'}} = 1.60\sqrt{1-0.58} = 1.04$$

بلغ الخطأ المعياري للقياس (1.04)، وهي كبيرة إلى حد ما لأنه كلما زاد الخطأ المعياري للقياس كلما انخفض معامل الثبات.

2. مصادر أخطاء القياس:

تتأثر درجات الأفراد في الاختبار بمصادر متعددة من أخطاء القياس، ونظرا لوجود أخطاء في القياس فانه يجب التعرف على خصائصها والطرق المتوفرة لتقدير مقدارها، وعندما تناقش مسألة ثبات درجات الاختبارات فان التحرر النسبي للدرجات من أخطاء القياس هو ما يجري مناقشته، فنتائج القياس الثابتة تكون متحررة نسبياً من أخطاء القياس، بينما النتائج الأقل ثباتاً تتأثر بدرجة أكبر بأخطاء القياس (رينولدس و لوفنجستون، 2013). ويوجد عدد من العوامل التي يمكن أن تؤثّر درجات الاختبارات، ومن أهم الأخطاء التي تكون راجعة إلى معاينات المحتوى ومعاينات الوقت، بالإضافة إلى مصادر أخرى للأخطاء.

1.2. أخطاء معاينات المحتوى:

من الصعب أن يشتمل الاختبار على جميع البنود أو السلوكات الممكنة، وبدلاً من ذلك يتم تحديد نطاق أو مجموعة شاملة من البنود استناداً إلى محتوى القدرة أو السمة المراد تغطيتها، ويتم سحب عينة من البنود من هذا النطاق، وهذه العينة ربما تكون ممثلة للنطاق المستمدة منه أو لا تكون كذلك. والخطأ الناجم عن الفروق بين عينة بنود الاختبار ونطاق البنود (أي جميع البنود الممكنة) يُشار إليه بأنه خطأ معاينات المحتوى.

ويتم تحديد مقدار أخطاء القياس الذي يرجع إلى معاينات المحتوى بكيفية معاينتنا للنطاق الشامل للبنود، فاذا كانت بنود الاختبار عينة جيدة من النطاق، فان مقدار أخطاء القياس سوف يكون صغيرا نسبيا، وإذا كانت البند عينة غير جيدة من النطاق، فان مقدار أخطاء القياس التي ترجع إلى معاينات المحتوى سوف يكون كبيراً (رينولدس و لوفنجستون، 2013).

2.2. أخطاء معاينات الوقت:

يمكن أن تحدث أخطاء القياس نتيجة اختيار وقت معين لتطبيق الاختبار، ومن المواقف التي تؤدي إلى تغيرات عشوائية في أداء الأفراد المختبرين التعب، والمرض، والقلق أو البيئة الاختبارية كالضوضاء، ودرجة الحرارة ف أداء الاختبار. ويُشار إلى هذا النوع من أخطاء القياس بأنه أخطاء

معاينات الوقت، والذ يعكس التقلبات العشوائية في الأداء من موقف أو وقت لآخر، مما يجعل قدرتنا على تعميم النتائج محدودة عبر مواقف مختلفة.

3.2. مصادر أخرى للأخطاء:

على الرغم أن معاينات المحتوى ومعاينات الوقت يرجع إليها الجزء الأكبر من الأخطاء العشوائية في الاختبارات، إلاّ أن أخطاء التطبيق وتصحيح الدرجات التي لا تؤثر على جميع الأفراد تأثيراً متساوياً تسهم أيضاً في الأخطاء العشوائية الملاحظة في الدرجات (رينولدس و لوفنجستون، 2013). ومن أمثلة مصادر أخطاء القياس، مثلا، أخطاء في التطبيق، وأخطاء في جمع الدرجات، وأخطاء في تقدير الدرجات.

3. مفهوم الثبات:

تعرف وثيقة معايير الاختبارات النفسية والتربوية (2014) Standards for Educational and Psychological Testing الصادرة عن الجمعية الأمريكية للبحث التربوي والجمعية الأمريكية لعلم النفس والمجلس القومي للقياس الثبات على أنه اتساق القياسات من خلال إعادة تطبيق أداة القياس على مجتمع من الأفراد أو المجموعات (APA, AERA, & NCME, 2014) .

يقصد بالثبات مدى خلوه من الأخطاء العشوائية التي تشوب القياس، أي مدى قياس الاختبار للمقدار الحقيقي للسمة التي يهدف لقياسها، فدرجات الاختبار تكون ثابتة إذا كان الاختبار يقيس سمة معينة قياسا متسقا في الظروف المتباينة التي قد تؤدي إلى أخطاء القياس (علام، 2000).

بناء على التعريفين السابقين يشير الثبات إلى اتساق درجات الأفراد عبر التطبيقات المتكررة لنفس الاختبار أو الصيغ المتكافئة له، والمصدر الرئيسي لعدم اتساق الأداء على الاختبار هو أخطاء القياس العشوائية. وبتعبير بسيط يشير الثبات إلى دقة درجات القياس عبر مختلف ظروف تطبيق الاختبار (إعادة الاختبار، اختلاف الصيغ، اختلاف البنود، اختلاف التقديرات) على نفس العينة.

تتضمن عملية القياس عدداً من أبعاد القياس مثلا، الفترات، الصيغ، البنود، المصححين، ويشير القياس الثابت إلى اتساق بين مختلف هذه الأبعاد، وعدم الاتساق بين مختلف أبعاد القياس يشير إلى خطأ القياس أو عدم الثبات. فالثبات هو أن تكون درجات الأفراد المختبرين نفسها عبر مختلف الشروط، بمعنى عندما يأخذ الأفراد مجموعة بنود أو أخرى، وتصحح إجاباتهم بمقدر إلى أخر، وعندما يختبرون في فترة أخرى تكون درجاتهم نفسها قدر الامكان.

من الناحية النظرية الثبات هو أن تعكس البيانات المستمدة من الاختبارات والمقاييس الجوانب الحقيقية للسمة أو القدرة المقاسة، أي المصادر المنتظمة للتباين ولا تعكس عوامل الصدفة أو العشوائية أي تباين الخطأ (علام، 2000).

أما من الناحية الاحصائية فالثبات هو نسبة تباين الدرجة الملاحظة إلى تباين الدرجة الحقيقية، أو مربع معامل الارتباط بين الدرجة الملاحظة والدرجة الحقيقية. ويعبر عنه وفق الصيغة التالية:

$$\boldsymbol{\rho^2}_{\boldsymbol{XT}} = \frac{\boldsymbol{\sigma^2(T)}}{\boldsymbol{\sigma^2(T) + \sigma^2(E)}}$$

$\boldsymbol{\sigma^2(T)}$: تباين الدرجة الحقيقية.

$\boldsymbol{\sigma^2(E)}$: تباين درجة الخطأ.

$\boldsymbol{\sigma^2(T) + \sigma^2(E)}$: تباين الدرجة الملاحظة لأن:

الدرجة الملاحظة = الدرجة الحقيقية + درجة الخطأ العشوائي.

بما أن ظروف تطبيق الاختبارات مختلفة فان الأخطاء العشوائية التي تؤثر على الدرجات الملاحظة للأفراد تكون مختلفة، لذلك طورت نظرية القياس الكلاسيكية طرق عديدة لتقدير ثبات درجات الاختبارات تعتمد على إعادة تطبيق الاختبار، والصور المتكافئة للاختبار، والاتساق الداخلي، وتقديرات المصححين.

المحاضرة التاسعة

طرق تقدير ثبات الاختبار

الأهداف:

- يميّز الطالب بين طرق تقدير ثبات درجات الاختبار.
- يقدّر الطالب معامل ثبات درجات الاختبار بطرق الاستقرار، والتكافؤ، والاستقرار–التكافؤ.
- يقدّر الطالب معامل ثبات درجات الاختبار بطرق التجزئة النصفية.
- يقدّر الطالب معامل ثبات درجات الاختبار بطرق الاتساق الداخلي بين البنود.
- يقدّر الطالب معامل ثبات درجات الاختبار بطرق الاتفاق بين تقديرات المحكمين.
- يفسّر الطالب معامل ثبات درجات الاختبار.

توجد عدة طرق لتقدير ثبات الاختبارات والمقاييس التربوية والنفسية تبعا لطبيعة خطأ القياس المراد تقييم أثره على الدرجة الملاحظة للفرد، فإذا أردنا التأكد من مدى استقرار الدرجات عبر الزمن نستخدم طريقة الاستقرار (الاختبار–إعادة الاختبار)، وإذا أردنا التأكد من تكافؤ درجات الاختبار باختبار أخر موازي نستخدم طريقة التكافؤ، وإذا أردنا التأكد استقرار وتكافؤ درجات الاختبار نستخدم طرقة الاستقرار والتكافؤ، وإذا أردنا التحقق من اتساق بنود الاختبار نستخدم طريقة الاتساق الداخلي، وفي حالة ما إذا أردنا التأكد من اتساق تقديرات المصححين نستخدم طريقة الاتساق بين المقدّرين.

1. طرق تتطلب تطبيق اختبارين

1.1. طريقة الاستقرار (الاختبار–إعادة الاختبار):

تعتمد هذه الطريقة على تطبيق الاختبار على عينة من الأفراد ثم يُعاد تطبيق الاختبار نفسه على العينة نفسها في الظروف نفسها بعد فترة زمنية محدّدة. يعتبر الاختبار الأول قياسا موازيا للاختبار الثاني وتسمى هذه الطريقة أيضا بالاستقرار عبر الزمن.

يقدّر الثبات وفق هذه الطريقة بتقدير معامل الارتباط "بيرسون" Pearson بين درجات الاختبارين. يبدو من السهل تطبيق طريقة الاختبار– إعادة الاختبار، ولكن يجب الأخذ بعين الاعتبار بعض المكونات الأساسية للسياق قبل إجراء التجريب الذي يهدف إلى تقدير الاستقرار:

1- يجب ضمان أن طريقة تقدير الثبات ملائمة لطبيعة الاختبار.

2- يجب أن تكون العينة المختارة لهذا التجريب ممثلة للمجتمع المستهدف من الاختبار.

3- يجب أن تعكس ظروف التجريب (حدود الوقت، المكان، الضجيج...) لضمان تطبيق الاختبار بشكل عادي.

4- يجب الأخذ بعين الاعتبار مسافة الوقت بين الاختبارين.

يمكن التذكير بصيغة حساب معامل الارتباط بيرسون:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n : عدد أفراد العينة.

x : درجات الأفراد في الاختبار الأول.

y : درجات الأفراد في الاختبار الثاني.

نشاط تدريبي:

طُبّق اختبار لقياس التفكير المنطقي على عينة من الأفراد، وبعد مدة (20) يوماً أعيد تطبيق الاختبار على العينة نفسها، فتحصلنا على النتائج التالية:

الأفراد	1	2	3	4	5	6	7	8	9	10
الاختبار	3	4	3	2	5	4	5	7	4	5
إعادة الاختبار	1	4	4	3	5	5	4	8	4	6

- قدّر معامل الاستقرار بين درجات الاختبار ودرجات إعادة الاختبار.

الحل:

$$\sum xy = 204 \qquad \sum x = 42 \qquad \sum y = 44$$

$$\sum x^2 = 194 \qquad \sum y^2 = 224$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{10 \times 204 - (42) \times (44)}{\sqrt{[10 \times (194) - (42)^2][10 \times (224) - (44)^2]}} = 0.83$$

معامل الارتباط بين درجات الاختبار ودرجات إعادة الاختبار مرتفع يقدّر بـ (0.83)، فدرجات الاختبار مستقرة بين الفترة الأولى والفترة الثانية، وبناء عليه فان درجات الاختبار ثابتة.

2.1. طريقة الصيغ المتكافئة:

يعتمد تقدير الثبات على الارتباط بين مجموعتين من القياسات، لذلك فان معامل الثبات يعتمد على استخدام صيغتين متكافئتين أو متوازيتين للاختبار على نفس الأفراد. فإذا توفرت صيغتين متكافئتين لاختبار معين مثلا، الصيغة **A** والصيغة **B** فان الثبات يتعلق باتساق درجات الأفراد بين الصيغتين، ويصبح تقدير الثبات ينتج من حساب معامل الارتباط "بيرسون" Pearson بين الصيغة **A** والصيغة **B** للاختبار.

وتتطلب هذه الطريقة أن تكون بنود صيغتي الاختبار من نفس النوع، وذات مستويات صعوبة متساوية، وتنتمي إلى نفس نطاق المحتوى السلوكي الذي تقيسه.

نشاط تدريبي:

تم إعداد صيغة لاختبار لقياس الاستعداد الدراسي، وصيغة مكافئة له وتم تطبيقهما على مجموعة من الأفراد، حيث جاءت النتائج على النحو التالي:

الأفراد	1	2	3	4	5	6	7	8	9	10	11	12
الصيغة A	34	36	36	32	35	31	31	32	32	35	41	40
الصيغة B	34	32	33	30	32	39	33	32	35	37	38	37

– قدّر معامل تكافؤ صيغتي الاختبار **A** و**B**.

الحل:

$$\sum xy = 14285 \qquad \sum x = 415 \qquad \sum y = 412$$

$$\sum x^2 = 14473 \qquad \sum y^2 = 14234$$

$$\mathbf{r} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{12 \times 14285 - (415) \times (412)}{\sqrt{[12 \times (14473) - (415)^2][12 \times (14234) - (412)^2]}} = 0.35$$

يقدّر معامل الارتباط بين درجات الأفراد في الصيغة A ودرجاتهم في الصيغة B بـ (0.35)، فدرجات الاختبار غير متكافئة، مما يؤكد على عدم ثبات درجات الاختبار.

3.1. طريقة الاستقرار والتكافؤ:

في تقدير معامل الاستقرار نسعى إلى تحديد تأثير معاينات مرور الوقت على ثبات الدرجات، وفي تقدير معامل التكافؤ فإننا نسعى إلى قياس تأثير معاينات البنود على الدرجة الكلية للفرد، وعندما نحاول أن نأخذ في الاعتبار هذين المصدرين للتغيرات في الدرجة الكلية فإننا قد انتقلنا إلى معامل تكافؤ الاستقرار الذي يتم تقديم قيمة الثبات من خلال الارتباط بين صيغتي الاختبار في أوقات مختلفة (Laveault & Grégoire, 2014) .

وفي الواقع، يصبح حساب معادلة الاستقرار ضروريًا عندما لا يمكن استخدام نفس الاختبار في حساب الاستقرار. نظرًا لأن مصدرين للتقلب العشوائي سيكونان موجودين في حساب هذا الارتباط، فإن معامل ثبات التكافؤ هو عمومًا أقل تقديراً للثبات بين الثلاثة التي درسناها.

يمكن استخدام هذه الطريقة عندما نريد أن نجمع بين طريقتين؛ طريقة الاستقرار وطريقة التكافؤ، وعندما نريد قياس مفاهيم معينة بمجموعات مختلفة من البنود المستمدة من نطاق شامل أين تكون درجات الأفراد المختبرين بكل مجموعة منها متطابقة أو متقاربة. ويمكن تقدير معامل الاستقرار والتكافؤ Stability and Equivalence بنفس طريقة تقدير معامل التكافؤ، ولكن تطبق أحد صيغ الاختبار ثم بعد فترة زمنية (تكون طويلة نسبيا) تطبق الصيغة الأخرى، ونحسب ارتباط "بيرسون" Pearson بين مجموعة الدرجات (Crocker & Algina, 2006).

وتكون القيمة التقديرية لمعامل الاستقرار والتكافؤ أقل من الطريقتين السابقتين لأنها تجمع بين الأخطاء العشوائية التي تؤثر على كل منهما لأنها تعكس الأخطاء الناتجة عن اختلاف الدرجات الناتجة عن التغيرات التي تحدث للأفراد في السمة التي يقيسها الاختبار بين الفترتين، وكذا اختلاف بنود صيغتي الاختبار.

نشاط تدريبي:

نفترض أننا طبقنا صيغتين متكافئتين (**A**) و(**B**) من اختبار على عينة تكونت من (7) أفراد في فترتين مختلفتين، حيث أن الصيغة (**A**) تم تطبيقها في يوم محدد تلتها الصيغة (**B**) في اليوم نفسه، وبعد فترة زمنية مدتها (20) يوماً أعيد تطبق الصيغة (**A**)، وفيما يلي درجات الأفراد المحصلة.

الأفراد	1	2	3	4	8	9	10
الصيغة A	9	7	7	8	5	4	5
الصيغة B	8	8	7	6	6	4	5
الصيغةA – الفترة2	13	8	8	8	7	5	4

الحل:

لإيجاد القيمة التقديرية لمعامل الاستقرار والتكافؤ نستخدم درجات الأفراد في الصيغة **B** الصيغة **A** في الفترة الثانية، باتباع الخطوات التالية:

$$\sum xy = 354 \qquad \sum x = 44 \qquad \sum y = 53$$

$$\sum x^2 = 290 \qquad \sum y^2 = 451$$

$$\mathbf{r} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{7 \times 354 - (44) \times (53)}{\sqrt{[7 \times (290) - (44)^2]\,[7 \times (451) - (53)^2]}} = 0.81$$

يلاحظ أن قيمة معامل الاستقرار والتكافؤ بلغت (0.81)، وهي قيمة تدل على استقرار وتكافؤ درجات صيغة الاختبار **B** ودرجات الصيغة **A** في الاختبار الفترة الثانية، مما يدل استقرار الصيغتين وتوازيهما (تكافئهما).

2. طرق تتطلب تطبيقاً واحداً للاختبار

تتطلب الطرق الثلاث السابقة التي تناولناها تطبيق اختبار أو صيغ متكافئة له مرتين بفترات زمنية فاصلة، غير أنه في بعض الأحيان من الصعب إعداد صيغتين متكافئتين أو الحصول عليها، أو تطبيق الاختبار مرتين، حينها من الأفضل استخدام اختبار واحد وتطبيقه مرة واحدة.

من بين طرق تقدير كيفية تعميم اتساق درجات أداء الأفراد في الاختبار على مجال بنود يمكن أن تقدّم للأفراد، هو تحديد درجة اتساق أدائهم عبر البنود أو المجموعات الفرعية من البنود في هذه الصيغة الاختبارية الواحدة، والاجراءات المصممة لتقدير الثبات في هذه الظروف تسمى طرق الاتساق الداخلي، ويكون اهتمامنا الرئيس في هذه الطريقة متعلقاً بالأخطاء الناجمة عن معاينات المحتوى (Crocker & Algina, 2006).

سوف نتناول طريقتين واسعتي الانتشار من طرق الاتساق الداخلي تستخدمان في تقدير معامل الثبات تنتجان من تطبيق واحد للاختبار، وتُعرف الأولى بطرق التجزئة النصفية Methods

Split-Half، وتُعرف الثانية بالطرق التي تتطلب تحليل بنية التباين أو التباين المشترك للبنود Item covariances Methods، وكل هذه الطرق تقدّم دليلاً للاتساق الداخلي لاستجابات الأفراد على بنود صيغة واحدة من الاختبار.

1.2. طرق التجزئة النصفية:

في هذه الطريقة يطبق مستخدم الاختبار ومطوره صيغة واحدة من الاختبار على عينة من الأفراد، وبعد تصحيح الاختبار يجزئ بنوده إلى نصفين أو فرعين يتألف كل جزء من نصف الاختبار الأصلي من حيث الطول، ولهذا عند تطبيق اختبار يتألف من (40) بندا فانه يُجزأ إلى نصفين كل منهما يتألف من (20) بندا.

وتوجد أربع طرق شائعة لتجزئة الاختبار إلى نصفين هي: (Crocker & Algina, 2006)

1- تعيين البنود الفردية للاختبار الفرعي الأول والبنود الزوجية للاختبار الفرعي الآخر.

2- ترتيب البنود وفقا لمستوى صعوبتها اعتماداً على استجابات الأفراد، ثم تعيين البنود التي ترتيبها فردي للاختبار الفرعي الأول والبنود التي ترتيبها زوجي للاختبار الفرعي الثاني.

3- التعيين العشوائي للبنود في كلا من النصفين.

4- تجزئة الاختبار إلى جزأين حيث تكون هناك مزاوجة بين بنود كلا الجزأين من حيث المحتوى.

بعدها يجمع كلا من النصفين مستقلاً عن الآخر، ويحسب معامل الارتباط بين درجات المجموعتين الفرعيتين من البنود كمؤشر لثبات درجات الاختبار، وتوجد العديد من صيغ تقدير معامل الثبات بطريقة التجزئة النصفية، منها:

1.1.2. صيغة سبيرمان-براون Spearman-Brown:

تعتمد هذه الصيغة على حساب معامل الارتباط بيرسون بين نصفي الاختبار، ثم تصحيح معامل الثبات باستخدام صيغة سبيرمان-براون للحصول على قيمة معامل ثبات الاختبار الكلي. وعند تطبيق معامل الارتباط بين نصفي الاختبار تكتب الصيغة على النحو التالي:

$$\boldsymbol{\rho_{xx'} = \frac{2\,\rho AB}{1 + \rho AB}}$$

$\boldsymbol{\rho_{xx'}}$: معامل الارتباط المعدل للاختبار الكلي

ρAB: معامل الارتباط بين نصفي الاختبار.

يجب على مستخدمي صيغة "سيبرمان–براون" ملاحظة أنها طريقة تعتمد بالأساس على افتراض توازي نصفي الاختبار، فكلما تم تحرى المبدأ كانت النتائج أكثر دقة، وكلما انحرفت عن هذا الافتراض بشكل أكبر كلما كانت النتائج أقل دقة (Crocker & Algina, 2006).

2.1.2. صيغة رولون Rulon:

تعتمد هذه الطريقة على تباين درجات نصفي الاختبار عوضا عن معامل الارتباط بيرسون، فهي طريقة تتميز بالسهولة والسرعة في تقدير معامل الثبات، كما أنها طريقة بديلة تتطلب استخدام فرق الدرجات بين نصفي الاختبار أي $D = A - B$ حيث تمثل A درجة الفرد على نصف الاختبار الأول وتمثل B درجة الفرد على نصف الاختبار الثاني، ويستخدم تباين فروق الدرجات على أنه تباين الخطأ في صيغة معامل الثبات:

$$\rho_{xx'} = 1 - \frac{\sigma_D^2}{\sigma_T^2}$$

σ_D^2: تباين الفرق بين الدرجات الفردية والدرجات الزوجية.

σ_T^2: تباين الدرجات الكلية للاختبار.

3.1.2. صيغة قاتمان Guttman:

تستخدم هذه الطريقة أيضا في تقدير معامل ثبات الاختبار بواسطة التجزئة النصفية بين الاختبار الفرعي الأول والاختبار الفرعي الثاني، كما أنها تصلح عندما لا تتساوى الانحرافات المعيارية لنصفي الاختبار، حيث تحسب قيمة الثبات وفق هذه الطريقة بالصيغة التالية:

$$\rho_{xx'} = 2\left[1 - \frac{(\sigma_x^2 + \sigma_{x'}^2)}{\sigma_T^2}\right]$$

σ_x^2: تباين درجات الاختبار الفردي.

$\sigma_{x'}^2$: تباين درجات الاختبار الزوجي.

σ_T^2 : تباين الاختبار الكلي.

نشاط تدريبي:

قام أحد الفاحصين ببناء اختبار لقياس القدرة على حل المشكلات يشتمل على (10) بنود، وتم تطبيقه على عينة تكونت من (8) أفراد، فتحصل على النتائج التالية:

البنود	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	0	1	1	1	1	1
2	0	1	1	0	0	1	0	1	1	1
3	1	1	0	1	0	1	0	0	1	0
4	0	0	1	1	1	0	1	0	0	1
5	0	0	1	1	0	0	0	0	1	1
6	0	1	1	0	1	0	1	1	1	0
7	1	1	1	1	1	0	0	1	1	1
8	1	0	1	0	1	1	0	1	0	1

- قدّر معامل ثبات الاختبار بطرق التجزئة النصفية "سبيرمان–براون" Spearman–Brown، و"قاتمان" Guttman و"رولون" Rulon.

الحل:

1. تقدير الثبات بطريقة "سبيرمان–براون" Spearman–Brown:

n	الاختبار الفرعي بنود فردية (A)	الاختبار الفرعي بنود زوجية (B)	A^2	B^2	AB
1	5	4	25	16	20
2	2	4	4	16	8
3	2	3	4	9	6
4	3	2	9	4	6
5	2	2	4	4	4
6	4	2	16	4	8
7	4	4	16	16	16
8	3	3	9	9	9
Σ	**25**	**24**	**87**	**78**	**77**

قبل تقدير معامل الثبات بطريقة "سبيرمان–براون" يتم تقدير معامل الارتباط "بيرسون" بين درجات الأفراد في الاختبار الفرعي للبنود الفردية والاختبار الفرعي للبنود الزوجية، وذلك كما يلي:

$$r = \frac{8 \times 77 - (25) \times (24)}{\sqrt{[8 \times (87) - (25)^2]\,[8 \times (78) - (24)^2]}} = 0.27$$

بعد حساب معامل الارتباط "بيرسون" بين الدرجات الفردية والدرجات الزوجية الذي قُدّر بـ (0.27)، يتم تصحيح هذا المعامل باستخدام نبوءة (تصحيح) "سبيرمان–براون"، كما يلي:

$$\rho_{xx'} = \frac{2\,\rho AB}{1 + \rho AB} = \frac{2 \times 0.27}{1 + 0.27} = 0.43$$

2. تقدير الثبات بطريقة "قاتمان" Guttman و"رولون" Rulon:

n	x	x'	x^2	x'^2	T	T^2	D	D^2
1	5	4	25	16	9	81	1	1
2	2	4	4	16	6	36	−2	4
3	2	3	4	9	5	25	−1	1
4	3	2	9	4	5	25	1	1
5	2	2	4	4	4	16	0	0
6	4	2	16	4	6	36	2	4
7	4	4	16	16	8	64	0	0
8	3	3	9	9	6	36	0	0
Σ	25	24	87	78	49	319	1	11

– معامل الثبات بطريقة "قاتمان "Guttman":

$$\sigma_x^2 = \frac{8\,(87) - (25)^2}{8\,(8-1)} = 1.27$$

$$\sigma_{x'}^2 = \frac{8\,(78) - (24)^2}{8\,(8-1)} = 0.86$$

$$\sigma_T^2 = \frac{8\,(319) - (49)^2}{8\,(8-1)} = 2.7$$

$$\rho xx' = 2\left[1 - \frac{\sigma_x^2 + \sigma_{x'}^2}{\sigma_T^2}\right] = 2\left[1 - \frac{1.27 + 0.86}{2.7}\right] = 0.42$$

– تقدير الثبات بطريقة "رولون" Rulon:

$$\sigma_D^2 = \frac{8\,(11) - (1)^2}{8\,(8-1)} = 1.55$$

$$\sigma_T^2 = \frac{8\,(319) - (49)^2}{8\,(8-1)} = 2.7$$

$$\rho_{xx'} = 1 - \frac{\sigma_D^2}{\sigma_T^2} = 1 - \frac{1.55}{2.7} = 0.43$$

يُلاحظ أن معاملات الثبات بطرق التجزئة النصفية لـ "سبيرمان–براون"، و"قاتمان" و"رولون" متساوية، والتي بلغت (0.43)، وهذا يدل على أن الاختبار غير ثابت. فعلى الرغم أن طريقة "سبيرمان–براون" تعتمد على تصحيح معامل الارتباط "بيرسون"، وطريقة "قاتمان" تعتمد على تباينات الدرجات الفردية والزوجية، وتقوم طرقة "رولون" على تباين الفرق بين الدرجات الفردية والدرجات الزوجية إلاّ أن معاملات الثبات متساوية.

2.2. طرق التباين المشترك للبنود:

تتميز طرق تقدير الثبات بواسطة التجزئة النصفية بالبساطة والسهولة لأنها تخضع إلى طريقة تقسيم الاختبار، لذلك ابتكرت طرق جديدة بعد تلك الفترة التي ظهرت فيها طرق التجزئة النصفية، حيث تعتمد هذه الطرق بالأساس على مدى اتساق بنود الاختبار أي الارتباطات أو التباينات الداخلية بين البنود.

وتتمثل طرق الثبات التي تعتمد على التباين المشترك للبنود في ثلاثة طرق شائعة، وهي طريقة "ألفا كرونباخ" Cronbach α وطريقتي "كيودر–ريتشاردسون" Kuder–Richardson رقم (20 و21)، وطريقة تحليل التباين لـ "هويت" Hoyt.

1.2.2. صيغة "ألفا كرونباخ" Cronbach α:

يُعدّ معامل "ألفا" من أكثر طرق تقدير الثبات استخداماً في البحوث النفسية والتربوية على الاطلاق، حيث يُستخدم في حساب الاتساق الداخلي للبنود ثنائية التصحيح أو متدرجة التصحيح مثل اختبارات المقال أو مقاييس الاتجاهات، ففي بداية الخمسينات قدم مؤلفات ومناقشة شاملة للطرق المختلفة المستخدمة لتقدير الاتساق الداخلي، وربطها جميعها في صيغة عامة واحدة تُعرف اسم معامل "كرونباخ ألفا" α (Cronbach, 2004) .

ويُقدّر معامل "ألفا" بواسطة الصيغة التالية:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_x^2}\right)$$

k: عدد بنود الاختبار.

$\Sigma\sigma_i^2$: مجموع تباينات بنود الاختبار.

σ_x^2 : التباين الكلي للاختبار.

نشاط تدريبي:

طبق أحد الفاحصين مقياسا للاتجاهات نحو التقويم الجامعي مكوّناً من (6) عبارات متدرجة وفق سلم "ليكرت" الخماسي على عينة من (10) طلبة جامعيين، فتحصل على البيانات التالية:

البنود	1	2	3	4	5	6
1	1	3	2	3	4	1
2	2	3	3	3	3	2
3	3	4	2	2	4	3
4	5	1	5	1	5	1
5	5	5	1	4	4	3
6	4	5	3	5	5	4
7	4	2	2	4	4	4
8	1	3	1	5	4	2
9	3	1	2	3	1	2
10	1	3	2	3	2	3

– قدّر معامل ثبات درجات المقياس باستخدام طريقة "ألفا (α) كرونباخ.

الحل:

$$\sigma_2^2 = \frac{10\,(108) - (30)^2}{10\,(10-1)} = 2.00 \qquad \sigma_2^2 = \frac{10\,(107) - (29)^2}{10\,(10-1)} = 2.54$$

$$\sigma_4^2 = \frac{10\,(123) - (33)^2}{10\,(10-1)} = 1.57 \qquad \sigma_3^2 = \frac{10\,(65) - (23)^2}{10\,(10-1)} = 1.34$$

$$\sigma_6^2 = \frac{10\,(73) - (25)^2}{10\,(10-1)} = 1.17 \qquad \sigma_5^2 = \frac{10\,(144) - (36)^2}{10\,(10-1)} = 1.60$$

Σx^2	Σx	6	5	4	3	2	1	البنود
196	14	1	4	3	2	3	1	**1**
256	16	2	3	3	3	3	2	**2**
324	18	3	4	2	2	4	3	**3**
324	18	1	5	1	5	1	5	**4**
484	22	3	4	4	1	5	5	**5**
676	26	4	5	5	3	5	4	**6**
400	20	4	4	4	2	2	4	**7**
256	16	2	4	5	1	3	1	**8**
144	12	2	1	3	2	1	3	**9**
196	14	3	2	3	2	3	1	**10**
3256	**176**	25	36	33	23	30	29	Σx
		73	144	123	65	108	107	Σx^2
17.60	**10.22**	**1.17**	**1.60**	**1.57**	**1.34**	**2.00**	**2.54**	σ^2

$$\Sigma\sigma_i^2 = 2.54 + 2.00 + 1.34 + 1.57 + 1.60 + 1.17 = 10.22$$

$$\sigma_x^2 = \frac{10\,(3256) - (176)^2}{10\,(10-1)} = 17.60$$

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_x^2}\right) = \frac{6}{6-1}\left(1 - \frac{10.22}{17.60}\right)$$

$$= 1.2 \times 0.42 = 0.50$$

بلغ معامل الثبات "ألفا" لدرجات المقياس (0.50)، وهو معامل ثبات منخفض مما يشير إلى عدم ثبات درجات مقياس الاتجاهات نحو التقويم الجامعي.

2.2.2. صيغتي كيودر–ريتشاردسون 20 و21 Kuder–Richardson:

تعتبر إحدى الطرق المعروفة التي اشتقها كيودر–وريتشاردسون كحل لمشكلة التجزئة النصفية التي فشلت في إعطاء نتيجة واحدة لاختبار معين، فالدراسة التي قدماها كأرضية تتضمن صيغتين تعرف حاليا بصيغة (KR–20) وصيغة (KR–21) واتخذت هذه الأسماء من الخطوات المرقمة في الاشتقاق الذي تم توضيحه في المجلة التي نُشر فيها المقال.

حيث أن صيغة (KR–20) تكتب كما يلي:

$$KR20 = \frac{k}{k-1}\left(1 - \frac{\sum p_i q_i}{\sigma_x^2}\right)$$

k: عدد بنود الاختبار.

$\sum p_i q_i$: مجموع نسب ضرب معامل صعوبة في معامل سهولة كل بند.

σ_x^2: تباين الاختبار الكلي.

هذه الصيغة مكافئة لصيغة معامل ألفا كرونباخ عندما يتم تعويض $\sum p_i q_i$ بـ σ_i^2 نلاحظ أن المجموع يشير إلى أن تباين كل بند ينبغي حسابه أولا ثم جمع البنود جميعها.

بافتراض تساوي صعوبة البنود جميعها اشتق Kuder–Richardson صيغة أبسط منها لا تتطلب حساب صعوبة وسهولة كل بند، حيث يمكن كتابة صيغة (KR–21) على النحو التالي:

$$KR21 = \frac{k}{k-1}\left(1 - \frac{\mu(k-\mu)}{k\,\sigma_x^2}\right)$$

μ : متوسط الدرجة الكلية.

k : عدد بنود الاختبار.

σ_x^2 : تباين الدرجة الكلية للاختبار.

عندما تتساوى صعوبة بنود الاختبار فان تقديرات الثبات في كلا من الصيغتين (KR–20) و(KR–21) تكون متساوية، في حين أنه عندما تختلف صعوبة البنود فان تقدير الثبات عند استخدام صيغة (KR–21) يكون أقل من القيمة المحسوبة باستخدام الصيغة (KR–20).

نشاط تدريبي:

بعد تطبيق اختبار من طرف أحد الأخصائيين التربويين لتقييم القدرة على التفكير الابتكاري على عينة من الطلاب، تحصل على النتائج التالية:

10	9	8	7	6	5	4	3	2	1	البنود
1	1	1	1	0	1	1	1	1	1	**1**
1	1	1	0	1	0	0	0	0	0	**2**
0	1	0	0	1	0	0	0	0	0	**3**
1	0	0	1	0	1	1	1	0	0	**4**
1	1	0	0	0	0	1	1	0	0	**5**
0	1	1	1	0		0	1	1	0	**6**
1	1	1	0	0	1	1	1	1	1	**7**
1	0	1	0	1	1	0	1	0	1	**8**
1	0	1	1	1	1	1	0	1	0	**9**
0	1	1	1	1	0	1	0	1	0	**10**

– أحسب ثبات الاختبار بطريقة **KR20** وطريقة **KR21**، وماذا تستنتج من معامل الثبات المحصل عليه في كلا من الطريقتين.

الحل:

Σx^2	Σx	10	9	8	7	6	5	4	3	2	1	البنود
81	9	1	1	1	1	0	1	1	1	1	1	**1**
16	4	1	1	1	0	1	0	0	0	0	0	**2**
4	2	0	1	0	0	1	0	0	0	0	0	**3**
25	5	1	0	0	1	0	1	1	1	0	0	**4**
16	4	1	1	0	0	0	0	1	1	0	0	**5**
36	6	0	1	1	1	0	1	0	1	1	0	**6**
64	8	1	1	1	0	0	1	1	1	1	1	**7**
36	6	1	0	1	0	1	1	0	1	0	1	**8**
49	7	1	0	1	1	1	1	1	0	1	0	**9**
36	6	0	1	1	1	1	0	1	0	1	0	**10**
363	57	0.7	0.7	0.7	0.5	0.5	0.6	0.6	0.6	0.5	0.3	p_i
$\sigma_x^2 = 4.23$		0.3	0.3	0.3	0.5	0.5	0.4	0.4	0.4	0.5	0.7	q_i
$\Sigma p_i q_i = 4.23$		**0.21**	**0.21**	**0.21**	**0.25**	**0.25**	**0.24**	**0.24**	**0.24**	**0.25**	**0.21**	$p_i \times q_i$

1. حساب معامل الثبات بطريقة KR20

$$\Sigma p_i q_i = 0.21 + 0.25 + 0.24 + 0.24 + 0.24 + 0.25 + 0.25 + 0.21 + 0.21 + 0.21 = 2.31$$

$$\sigma_x^2 = \frac{10\,(363) - (57)^2}{10\,(10-1)} = 4.23$$

$$\mathbf{KR20} = \frac{k}{k-1}\left(1 - \frac{\sum p_i q_i}{\sigma_x^2}\right) = \frac{10}{10-1}\left(1 - \frac{2.31}{4.23}\right)$$

$$= 1.11 \times 0.45 = 0.50$$

معامل ثبات درجات الاختبار منخفض بلغ (0.50)، وعليه فان الاختبار غير ثابت.

2. حساب معامل الثبات بطريقة KR21:

$$\mu = \frac{\Sigma x_i}{n} = \frac{57}{10} = 5.7$$

$$\mathbf{KR21} = \frac{k}{k-1}\left(1 - \frac{\mu(k-\mu)}{k\,\sigma_x^2}\right) = \frac{10}{10-1}\left(1 - \frac{5.7(10-5.7)}{10 \times 4.23}\right)$$

$$= 1.11 \times 0.42 = 0.47$$

بلغ معامل ثبات درجات الاختبار (0.47)، وهو منخفض يدل على عدم ثبات الاختبار.

يتضح بأن معامل الثبات المقدّر بطريقة KR20 قد بلغ (0.50) وهو مساوي تقريبا لمعامل الثبات المقدّر بطريقة KR21 الذي بلغ (0.47)، ويمكن أن يرجع ذلك إلى تقارب مستويات صعوبة بنود الاختبار، وجاءت قيمة KR21 أقل من قيمة KR20 لأن هناك اختلافات طفيفة بين معاملات صعوبة بنود الاختبار.

3.2.2. طريقة هويت Hoyt:

استخدم "هويت" Hoyt أسلوباً جديداً يختلف عن الأساليب السابقة المذكورة في تقدير معامل الثبات يؤدي إلى نتائج متماثلة لتلك الناتجة من معامل "ألفا كرونباخ"، ويعتمد على تحليل التباين بمعالجة الأفراد والبنود على أنها مصدرين للتباين، وتتطلب طريقة "هويت" معرفة طرق إحصائية أكثر تعقيدا من الطرق الأخرى لتقدير معامل الثبات، وباستخدام رموز تحليل التباين المعياري فان معامل الثبات يحدد على النحو التالي:

$$\rho_{xx'} = \frac{MS_P - MS_e}{MS_P}$$

MS_P: متوسط مربعات الأفراد المأخوذة من جدول تحليل التباين.

MS_e: متوسط مربعات البواقي المأخوذة من جدول تحليل التباين.

وقد ربط هويت هذه الصيغة بالتعريف النظري لمعامل الثبات، وذلك بملاحظة أن متوسط مجموع المربعات للأفراد MS_P تمثل تباين الدرجة الملاحظة، ومتوسط مجموع مربعات البواقي MS_e تمثل تباين الخطأ في التعريف النظري التالي للثبات:

$$\rho_{xx'} = \frac{\sigma_X^2 + \sigma_E^2}{\sigma_X^2}$$

نشاط تدريبي:

تم تطبيق اختبار يتكون من 3 مهمات معقدة على عينة مكونة من 4 طلاب في المرحلة الثانوية، وقد توزعت درجاتهم على النحو التالي:

	المهمات (البنود)		
الطلاب	**1**	**2**	**3**
1	1	2	3
2	0	1	2
3	1	2	3
4	0	1	2

– قدّر معامل الثبات بطريقة الاتساق الداخلي لـ "هويت" Hoyt.

الحل:

	المهمات (البنود)			متوسط درجات الأفراد في
الطلاب	**1**	**2**	**3**	**كل بند**
1	1	2	3	**2.00**
2	0	1	2	**1.00**
3	1	2	3	**2.00**
4	1	1	2	**1.33**
متوسط البند	**0.75**	**1.5**	**2.5**	**المتوسط الكلي 1.583**

لحساب الثبات بطريقة "هويت" يجب اتباع سبع خطوات وهي:

1-حساب مجموع المربعات الكلي SS_t بناء على درجة البند الفردية في نص الجدول على النحو التالي، بالتذكير فان المتوسط الكلي 1.583:

$$SS_t = (1-1.583)^2 + (2-1.583)^2 + (3-1.583)^2 + (0-1.583)^2 + (1-1.583)^2 + (2-1.583)^2 + (1-1.583)^2 + (2-1.583)^2 + (3-1.583)^2 + (1-1.583)^2 + (1-1.583)^2 + (2-1.583)^2$$

$$= 0.340 + 0.174 + 2.008 + 2.506 + 0.340 + 0.174 + 0.340 + 0.174 + 2.008 + 0.340 + 0.340 + 0.174$$

$$= 8.918$$

ويشير إلى التباين الكلي في البيانات.

2-حساب مجموع المربعات الراجع إلى أداء الطالب SS_p باستبدال كل درجة البند في كل صف بمتوسط درجة الطالب المقابلة، وحسابه على النحو التالي:

$$SS_p = (2-1.583)^2 \times 3 + (1-1.583)^2 \times 3 + (2-1.583)^2 \times 3 + (1.33-1.583)^2 \times 3 = 0.522 + 1.020 + 0.522 + 0.188$$

$$= 2.252$$

ويشير إلى التباين المرتبط بأداء الطلاب.

3-حساب مجموع المربعات المنسوب إلى الفروق بين البنود SS_i من خلال استبدال كل درجة العمود بمتوسط البند المعطى في العمود الأخير من الجدول، بالقيام بما يلي:

$$SS_i = (0.75-1.583)^2 \times 4 + (1.5-1.583)^2 \times 4 + (2.5-1.583)^2 \times 4$$

$$= 2.776 + 0.028 + 3.364$$

$$= 6.168$$

4-حساب التباين المرتبط بالخطأ $_e$ ، فكما ذكرنا التباين الكلي يتكون من ثلاث مكونات تسمى تباين الطالب، وتباين درجة البند، وتباين الخطأ، فتباين الخطأ ينتج من خلال طرح تباين الأفراد وتباين البنود من التباين الكلي، على النحو التالي:

$$SS_e = SS_t - SS_p - SS_i$$

$$SS_e = 8.918 - 2.252 - 6.168 = 0.498$$

يتم الحصول على متوسط المربعات للأفراد MS_p باستخدام الصيغة:

$$MS_? = \frac{SS_?}{df_?}$$

على النحو التالي:

5- حيث أن $df_p = 4 - 1 = 3$ لأن عدد الطلاب هو 4 وبالتالي فان النتيجة تكون:

$$MS_p = \frac{SS_p}{df_p} = \frac{2.252}{3} = 0.751$$

6- حيث أن $df_e = 6$ نلاحظ أن df_e هو نتاج الفرق بين df_t و df_i :

$$df_e = df_t - df_i = 8 - 2 = 6$$

$$MS_e = \frac{0.498}{6} = 0.083$$

7-حساب معامل الثبات:

$$\rho_{xx'} = \frac{MS_P - MS_e}{MS_P} = \frac{0.751 - 0.083}{0.751} = 0.889$$

معامل ثبات درجات الاختبار مرتفع بلغ تقريباً (0.89)، وبالتالي فان الاختبار ثابت.

3. طرق الاتساق بين تقديرات المحكمين:

إذا اعتمد في تقدير درجات اختبار على أحكام ذاتية، فمن المهم تقدير درجة الاتفاق عندما يقوم أكثر من فرد بتقدير الدرجات، ويُطلق عليه "ثبات تقديرات المحكمين" Interater reliability، وفي هذه الحالة يُطبق الاختبار مرة واحدة على عينة من الأفراد، ويقدّر فردان (أو أكثر) على حدة درجات الاختبار (راينولدس ولفنستون، 2013).

وتوجد عدة مؤشرات لتقدير معامل الاتفاق أو التجانس بين المقدرين، ومن أمثلة الاختبارات التي تستخدم في تقدير ثبات تقديرات المصححين أو الملاحظين، اختبار كاندل Kandall لاتساق التقديرات واختبار Cooper للاتفاق بين التقديرات، واختبار "كابا" لـ "كوهين" Cohen's Kappa في حالة تقديرات ثنائية (نجاح-فشل)، ومعامل إمكانية التعميمGeneralizability coefficient.

تسمح هذه الأساليب في تحديد الأخطاء الراجعة إلى عدم اتساق تقديرات الملاحظين والمصححين الناتجة عن الذاتية من خلال التشدد أو التساهل أثناء عملية التصحيح وأثر الهالة، وعدم وضوح موازين التقدير أو الملاحظة، أو فترة ملاحظة أداء الأفراد. وكلما ازدادت قيمة الخطأ الراجعة لعدم

اتساق التقديرات بين المصحّحين انخفض معامل ثبات درجات الاختبار، وكلما قلت قيمة الخطأ ارتفع معامل ثبات درجات الاختبار.

ويمكن تقدير اتفاق المحكمين بحساب النسبة المئوية من المرات التي يعيّن فيها محكمان نفس الدرجات لأداءات الأفراد، ويُطلق عادة على هذا المدخل "اتفاق المحكمين"، أو "النسبة المئوية للاتفاق"، وطريقة حساب ذلك يكون بالصيغة التالية:

الاتفاق بين المحكمين=عدد الحالات التي عُيّن لها نفس الدرجات/العدد الكلي للحالات×100

نشاط تدريبي:

قام محكميْن بتقدير مقاطع موسيقية عزفها 25 طالباً، وتم تقدير درجات القطع من 1 إلى 5 استناداً إلى محكات محددة في ميزان تقدير وصفي، حيث تدل الدرجة 1 على أدنى أداء وتدل 5 على أعلى أداء، والموضحة في الجدول التالي:

	تقديرات المحكم 2				
تقديرات المحكم 2	1	2	3	4	5
5	0	0	1	2	4
4	0	0	2	3	2
3	0	2	3	1	0
2	1	1	1	0	0
1	1	1	0	0	0

أحسب النسبة المئوية للاتفاق بين درجات المحكميْن.

الحل:

$$\text{الاتفاق بين المحكمين} = \frac{12}{25} \times 100 = 48\%$$

يتضح بأن درجة الاتفاق بين المحكمين بلغت (48%) وهي منخفضة تدل على عدم اتساق تقديرات المحكمين.

بعدما تناولنا مصادر أخطاء القياس التي تمحورت حول معاينة الوقت، ومعاينة محتوى الصيغة، معاينة محتوى البنود، ومعاينة المقدرين التي انبثقت منها أربعة طرق لتقدير الثبات تمثلت في

الاستقرار، والتكافؤ، والاتساق الداخلي، والاتساق بين التقديرات، والمعالجة الاحصائية للبيانات المحصلة من تصميمات طرق الثبات.

جدول (3): ملخص طرق تقدير الثبات

المصدر الرئيسي للخطأ	معامل الثبات	إجراءات جمع البيانات	المعالجة الاحصائية للبيانات
1. التغير في المفحوصين عبر الزمن	1. معامل الاستقرار	1. تطبيق اختبار، انتظار، إعادة الاختبار	1. حساب معامل الارتباط "بيرسون"
2. معاينة المحتوى من صيغة لأخرى	2. معامل التكافؤ	2. تطبيق الصيغة1، ثم الصيغة2	2. حساب معامل الارتباط "بيرسون"
3. معاينة المحتوى أو بنود معينة	3. معامل الاتساق الداخلي	3. تطبيق صيغة احد في موقف واحد	3.أ. تجزئة الاختبار إلى نصفين، ايجاد معامل الارتباط بينهما؛ استخدام تصحيح سبيرمان–براون. 3. ب. تجزئة الاختبار إلى نصفين؛ استخدام معادلة "قاتمان" أو "رولون" 3. ج. حساب تباين البنود؛ حساب "معامل ألفا" 3. د. حساب تباين البنود؛ حساب معامل "KR20، أو KR21" 3. هـ. حساب تباين الأفراد، وتباين الباقي؛ حساب الثبات بطريقة "هويت"
4. التغير بين تقديرات الملاحظين (المقيّمين)	4. معامل الاتساق بين التقديرات	4. تطبيق الاختبار، اختيار مقدّرين للتقييم	4. حساب معامل الاتفاق (أو الاتساق)؛ "نسبة الاتفاق"، معامل "كاندل"، أو "كابا"...

المصدر: (Crocker & Algina, 2006) مع تعديلات من المؤلف

المحاضرة العاشرة

العوامل المؤثرة على ثبات الاختبار

الأهداف:

- يحدد الطالب العوامل المؤثرة على ثبات درجات الاختبار.
- يقدّر الطالب طول الاختبار ومعامل الثبات المرغوب.

معاملات ثبات درجات الاختبار ليست قيما مطلقة، وإنما تعد قيما تقديرية تؤثر فيها عوامل متعددة يجب مراعاتها أثناء تصميم أدوات القياس، وعند انتقاء هذه الأدوات، واستخدامها، وتفسير نتائجها. حيث أنه بالإضافة إلى مصادر الأخطاء العشوائية المؤثرة على درجات الثبات يجب الأخذ بعين الاعتبار بعض العوامل الأخرى، والتي تتعلق بطول الاختبار، وتجانس عينة المختبرين، حدود الزمن، وخصائص بنود الاختبار، موضوعية التصحيح.

1. طول الاختبار:

درجات الاختبار الأطول أكثر ثباتاً من درجات الاختبار الأقصر المؤلف من فقرات متشابهة، لذلك فطول الاختبار أحد المظاهر التي تؤثر بالتأكيد على تباين الدرجة الحقيقية والدرجة الملاحظة، وقد اشتق Spearman–Brown علاقة بين طول الاختبار وتقدير الثبات، والتي عبّر عنها بالصيغة التالية:

$$\rho\, jj' = \frac{k \times \rho xx'}{1 + (k-1)\, \rho xx'}$$

ρ xx′: الثبات المرغوب للاختبار المعدل.

ρjj′ : الثبات الأصلي للاختبار.

***k*:** عدد مرات زيادة عدد البنود.

– تساعدنا صيغة العلاقة بين طول الاختبار وتقدير الثبات في التعرف على درجة الدقة التي سوف يصل إليها الاختبار عندما نعدل عدد البنود بنسبة **k**، فعند زيادة طول الاختبار فانّ قيمة **k** تصبح أكبر من (1)، وعندما يقل طول الاختبار فان قيمة **k** تصبح أقل من 1.

– يجب ملاحظة أن زيادة قيمة الثبات الناتجة عن زيادة طول الاختبار تتبع قانون الغلة (ليس بالضرورة زيادة طول الاختبار الى أبعد حدّ ممكن سوف يزيد بشكل مناسب في الثبات)، كما أن زيادة طول الاختبار سوف تترتب عليه زيادة في تكلفة كتابة البنود الاضافية والوقت المستهلك على تطبيق وتصحيح الاختبار، كما يشترط زيادة بنود الاختبار أن تكون موازية في المحتوى والصعوبة لبنود الاختبار الأصلي.

نشاط تدريبي:

في المثال السابق عندما تم تقدير معامل ثبات درجات الاختبار الذي تكوّن من (6) بنود باستخدام طريقة "ألفا كرونباخ" تحصلنا على معامل ثبات قُدّر بـ (0.50)، وفرضاً أردنا رفع عدد بنود الاختبار إلى (10) بنود. فكم يصبح معامل ثباته الجديد.

الحل:

$$\boldsymbol{k = \frac{10}{6} = 1.67}$$

$$\boldsymbol{\rho\, jj' = \frac{k \times \rho xx'}{1 + (k-1)\,\rho xx'}}$$

$$\boldsymbol{= \frac{1.67 \times 0.50}{1 + (1.67 - 1)\,0.50} = \frac{0.835}{1.335}}$$

$$\boldsymbol{= 0.63}$$

يصبح معامل ثبات درجات الاختبار (0.63) بعد رفع عدد بنوده إلى (10) بنود.

– تسمح لنا صيغة Spearman–Brown بتحديد نسبة طول الاختبار الذي يجب زيادته لبلوغ درجة مستهدفة من الثبات، وبتعديل المعادلة السابقة يمكن تعويض قيمة **k** لتصبح:

$$\boldsymbol{k = \frac{\rho xx'(1 - \rho jj')}{\rho jj'(1 - \rho xx')}}$$

$\boldsymbol{\rho xx'}$: معامل الثبات الأصلي.

$\boldsymbol{\rho jj'}$: معامل الثبات المرغوب.

– تجدر الاشارة إلى أن صيغة Spearman–Brown تشترط أن تكون البنود التي تُضاف إلى الاختبار متكافئة مع بنود الاختبار الأصلي، بمعنى لها نفس المحتوى ومستويات صعوبة متساوية.

نشاط تدريبي:

فرضاً أننا تحصلنا على معامل ثبات درجات اختبار معين قُدّر بـ (0.64) وعدد بنود (12)، وأردنا بلوغ معامل ثبات يقدّر بـ (0.80). كم يصبح عدد بنود الاختبار للحصول على معامل الثبات المطلوب بلوغه.

الحل:

$$\boldsymbol{k = \frac{\rho xx'(1-\rho jj')}{\rho jj'(1-\rho xx')}}$$

$$\boldsymbol{= \frac{0.80\ (1-0.64)}{0.64\ (1-\ 0.80)} = \frac{0.288}{0.128}}$$

$$\boldsymbol{= 2.25}$$

نسبة عدد البنود المطلوب إضافتها إلى الاختبار هي 2.25.

$$\boldsymbol{2.25 \times 12 = 27}$$

عدد بنود الاختبار المطلوبة للحصول على معامل ثبات يقدّر بـ (0.80) هو 27 بنداً.

2. تجانس عينة المختبرين:

بما أن قيمة معامل الثبات تعتمد على تباين الأفراد في درجاتهم الحقيقية ودرجات الخطأ، لذا فتجانس عينة المختبرين يعد مهما في تطوير الاختبار واختياره. فاذا كانت عينة ناشر الاختبار غير متجانسة في السمة المقاسة بدرجة كبيرة فان هذا يؤدي حتما إلى خفض معامل الثبات عند تطبيق الاختبار على عينة أكثر تجانسا.

قدم Gullikson (1950) خلاصة للجدل السيكومتري لعدم تجانس عينة المختبرين وأثره على ثبات الاختبار، كما قدم Magnuson (1967) الصيغة التالية للتنبؤ بتغير الثبات نتيجة تغير في تباين العينة:

$$\boldsymbol{\rho jj' = 1 - \frac{\sigma_x^2\ (1-\rho xx')}{\sigma_u^2}}$$

$\boldsymbol{\sigma_x^2}$: تباين العينة الأصلية.

$\boldsymbol{\sigma_u^2}$: تباين العينة الجديدة.

$\boldsymbol{\rho xx'}$: تقدير ثبات العينة الأصلية.

$\rho_{jj'}$: الثبات المتنبأ به للعينة الجديدة.

على سبيل المثال أثناء دراسة ثبات درجات أداة قياس معينة فان العديد من الوضعيات التي يمكن أن تساهم في خفض الفروق الفردية، فتطبيق اختبار على عينة تمتلك تباينا أقل من المجتمع الأصلي أي عينة منحدرة من وسط مفضل يمكن أن تؤدي إلى شك في تباين النتائج المحصلة فتكون أقل من المحصلة بواسطة عينة ممثلة.

تفترض المعادلة تساوي تباينات الخطأ لكلا المجموعتين، وأن التغير في الدرجات الملاحظة يعود إلى الفروق في توزيع الدرجات الحقيقية للمجموعة، حيث أن مستخدم الاختبار لا يمكنه الجزم بتحقيق هذا الافتراض عندما تكون العينة المستخدمة مختلفة بشكل ملحوظ عن العينة الأصلية، ومن الأنسب هنا التحقق من ثبات العينة الجديدة. إذا كانت عينة المفحوصين متجانسة بشكل كبير في السمة المقاسة فان الثبات المحسوب يكون أقل مما لو كانت العينة غير متجانسة.

3. حدود الزمن:

عندما يكون الاختبار موقوتا فان العديد من المختبرين لا يصلون إلى الاجابة على كل البنود، لذا فمدى مجهود المختبر يؤثر بانتظام في أدائه على بنود الاختبار لأن البنود المتروكة تؤثر على درجاته والتي نجدها في آخر الاختبار وتصحح عموما بالدرجة (0). هذا الاجراء سوف يخلق نوعا من التضخم في الارتباط بين البنود الأخيرة، مما يؤدي إلى اتساق أكبر بين البنود، والتي في الواقع غير ذلك لأن هذا الاتساق لا يكون راجعا إلى مؤشر أن البنود تقيس نفس الشيء ولكن لأنها متروكة من طرف الأفراد المختبرين.

تقدير ثبات اختبارات السرعة أو الموقوتة يجب تفسيره بحذر عندما تتطلب المهام الاختبارية أكثر من قدرة أداء مهام بسيطة بسرعة كبيرة، فتقدير الثبات يحدث مخاطرة نتيجة تزييف الارتباط بين البنود في حالة التجزئة النصفية أو حتى الاتساق الداخلي، ويفضل في هذه الحالة استخدام طريقة الاختبار وإعادة الاختبار التي لا تتأثر بعامل الزمن المحدد للإجابة.

4. خصائص بنود الاختبار:

تؤثر بنود الاختبار في ثبات درجات الاختبار ككل، فخلو الاختبار من الخطأ يعتمد على كيفية بناء البنود، فبعضها يشتمل على مؤثرات الاجابة مما يساعد على التخمين، وبعض البنود تتميز بدرجة عالية من الصعوبة أو السهولة مما يعمل على خفض قيمة معامل الثبات. كذلك البنود

الغامضة أو غير محددة الهدف أو التي تكون تعليماتها أو صياغتها غير دقيقة تؤثر على ثبات درجات الاختبار.

تؤدي صعوبة أو سهولة البنود بشكل كبير إلى التواء الدرجات سواء سلبا في حالة اختبار سهل أو التواء موجب في حالة اختبار صعب، فمعامل الارتباط بيرسون لا يمكن بلوغ قيمته القصوى (1) بتوزيع المتغيرين للارتباط معتدلة أو نفس الالتواء، فاختلاف توزيع البيانات في الاختبار عنها في إعادة الاختبار (تكون معتدلة في التطبيق الأول وملتوية في التطبيق الثاني) سوف يؤثر على استقرار الدرجات.

5. موضوعية التصحيح:

تؤثر أحكام المصححين على ثبات درجات الاختبار خاصة بالعوامل الذاتية أو عوامل التحيز، حيث تنخفض قيمة معامل الثبات نتيجة لذلك، فتصحيح الاختبارات الموضوعية مثل الاختيار من متعدد، والصواب والخطأ، والتكملة...وغيرها لا تطرح مشكلة لأنه عادة ما يكون التصحيح موضوعيا، ولكن المشكلة تبدو واضحة في تقدير درجات اختبارات المقال، وبعض اختبارات الأداء أو مقاييس الشخصية لأن التصحيح فيها يتطلب أحكاما ذاتية حول استجابات المختبرين مما يؤثر تأثيرا كبيرا على ثبات التقديرات.

المحاضرة الحادية عشرة

صدق الاختبار: مفهومه وأنواعه

الأهداف:

- يتعرّف الطالب على مفهوم صدق الاختبار.
- يميّز الطالب بين أنواع أدلة صدق درجات الاختبار (المحتوى، المحك، البناء).
- يقدّر الطالب صدق محتوى الاختبار باستخدام الاتفاق بين المحكمين واتساق البند بالأهداف.
- يقدّر الطالب الصدق المرتبط بمحك باستخدام طرق الارتباط والتنبؤ.
- يقدَر الطالب صدق التكوين الفرضي باستخدام طرق التمييز والارتباط والتجريب والتحليل المنطقي.
- يفسّر الطالب أدلة الصدق المجمعة عن الاختبار.

يعتبر الصدق من المسائل والاعتبارات أكثر أهمية في بناء وتقييم الاختبارات، وقد تطوَر سريعا في العشريتين الماضيتين نتيجة التطوَرات السريعة التي حدثت في مجال التقويم النفسي والتربوي، ويُعد مسألة أساسية وهامة فيما يتعلق بتفسير واستخدامات درجات الاختبار، ويتحدّد في الأدلة التي يمكن أن نقدمها لتدعيم القرارات التي يمكن اتخاذها على أساس درجات الاختبار.

وقد تغيَرت النظرة إلى الصدق الذي يُعدّ خاصية استدلالية للدرجات وليس خاصية من خصائص الاختبار، وقد تطوّر مفهوم الصدق خلال العقدين السابقين بعدما اعتبر كلاسيكيا مدى قياس الاختبار للسمة المطلوبة ليتحوّل إلى السعي نحو تأسيس مدى أهمية ملاءمة تفسير واستخدام درجات الاختبارات.

1. مفهوم الصدق:

يعدّ الصدق من معايير الجودة الفنية للاختبارات ذات أهمية في بناء وتقييم الاختبارات، وخاصية هامة من خصائص درجات الاختبارات، حيث يشير تقليديا إلى أن يقيس الاختبار أو أي أداة أخرى فعلاً ما أُعدّ لقياسه. ولكن هذه النظرة للصدق كلاسيكية لا تعبّر بدقة عن التعريف الاجرائي للمفهوم، فقد تطورت النظرة إلى الصدق كما وصفه Cronbach (1971) على أنه العملية التي من خلالها يجمع فيها مطوّر الاختبار أو مستخدمه الأدلة التي تدعم الاستنتاجات التي استخلصها من درجات الاختبار.

وتعرّفه وثيقة معايير العملية الاختبارية النفسية والتربوية American Educational Research Association. American Psychological Association & National Council on Measurement in Education (2014) على أنه الدرجة التي تؤيد بها الأدلة والنظرية تفسير درجات الاختبارات التي تتطلبها استخدامات مقترحة للاختبار، ولذلك فان الصدق يُعدّ من أكثر الاعتبارات أهمية عند بناء وتقييم الاختبارات.

وبالتالي يشير الصدق إلى الأدلة المستمدة من درجات الاختبارات لهدف معين ضمن مجموعة شروط وضعت مسبقا، لذلك فالصدق ضمن المفهوم الحديث ليس خاصية من خصائص الاختبارات وإنما خاصية من خصائص درجات الاختبارات.

فالصدق يتعلق بمدى فائدة أداة القياس في اتخاذ قرارات مرتبطة بغرض أو أغراض معينة، وليس خاصية من خصائص الأداة ذاتها، فالصدق يعدَ خاصية استدلالية (Messick. 1995). إضافة إلى أن الصدق لا يتضمن فقط تلاؤم الهدف مع السمة المطلوب قياسها وإنما في مختلف الأدلة المطلوب جمعها لاستخدام وتفسير درجات الاختبارات.

أكدت وثيقة معايير العملية الاختبارية (2014) AERA. APA. NCME بأن تأييد الصدق مسؤولية مشتركة بين مطوَر الاختبار الذي يقدَم أدلة وأسس منطقية للاستخدام المرجو للاختبار، ومستخدم الاختبار الذي يقيَم الأدلة الموجودة ضمن السياق الذي استخدم فيه الاختبار (p. 11).

2. أنواع الصدق

1.2. صدق المرتبط بالمحتوى:

يهدف صدق المحتوى إلى تقييم ما إذا كانت البنود تمثّل نطاق الأداء أو البناء المستهدف بشكل مناسب، ويجب أن يمثّل المحتوى تمثيلا جيدا لنطاق البنود الذي يتم تحديده مسبقا، ونقصد بنطاق البنود المعارف والمهارات والعمليات التي يتم معاينتها بواسطة بنود الاختبار. وحسب وثيقة "معايير العمليات الاختبارية التربوية والنفسية" يتم الحصول على أدلة المحتوى من خلال الارتباط بين محتوى الاختبار والبناء المستهدف قياسه (AERA. APA & NCME. 2014. p. 11).

ويتطلب أن يكون النطاق الشامل للبنود معرفا تعريفا دقيقا وإجرائيا للسمة المراد قياسها، فتحديد الأهداف السلوكية تعدّ خطوة ضرورية في قياس التحصيل، مثلا في اختبار البنود نادرا ما يهدف مستخدم الاختبار التعرف ما إذا كان المفحوص يعرف معاني كلمات محددة فقط، لكنه يهتم بمعرفة المفحوص بالكلمات المشابهة لها. عادة ما يستخدم في صدق المحتوى طريقة تحكيم

مجموعة من الخبراء المستقلين للحكم ما إذا كانت عينة البنود مناسبة لنطاق القياس، ويتضمن صدق المحتوى على الأقل سلسلة من الخطوات:

1- تحديد نطاق الأداء المستهدف.

2- اختيار فريق من الخبراء المؤهلين في مجال النطاق.

3- التزويد بهيكل بنائي لعملية مطابقة البنود لنطاق الأداء.

4- جمع وتلخيص البيانات الناتجة عن عملية المطابقة (Crocker & Algina, 2006) .

معظم أساليب تقدير صدق المحتوى تعتمد على الأحكام التقييمية لخبراء المواد الدراسية أو المهتمين بتنمية المهارات والكفاءات التعليمية والمهنية والفنية، وتتعلق هذه الأحكام بتقدير مدى تناظر بين بنود الاختبار والنطاق السلوكي الذي تمثّله هذه البنود.

- يمكن للتأكد من صدق المحتوى إعداد استمارة تشتمل على ميزان تقدير تتضمن الأبعاد الرئيسية المتعلقة بالنطاق السلوكي، مثل محتوى البنود، نوع المهارات ومجالها، ومصادر أو مواد ذات أهمية، ونوع البنود وملاءمتها للمحتوى والمهارة المرجوة. حينها يقوم المحكم بتقييم كل بند من بنود الاختبار في هذه الأبعاد على ميزان التقدير. ويُفضَل الاعتماد على أكثر من محكم للحصول على تقديرات أكثر اتساقا، ويمكن التحقق من ذلك بتحليل التقديرات إحصائياً باستخدام مؤشرات إحصائية تزوّد بدرجة اتساق التقديرات مثل تحليل التباين، معاملات الاتفاق.

- يمكن أيضا التحقق من الصدق امبريقياً بتطبيق الاختبار على عينة من الطلاب قبل بدء عملية التعليم، ثم إعادة تطبيقه بعد نهايتها وفحص نتائج الاختبار في المرتين للتعرف على ما إذا كان الاختبار يقيس بالفعل المجال الذي اهتمت به عملية التعليم.

يستند تقدير صدق المحتوى إلى ثلاثة فروض صاغها Lenon كالتالي:

1- يجب أن يكون المجال الذي يُختبر فيه الأفراد محدوداً بنطاق شامل للبنود الذي تبدو أهميته لهم، ويمكن تعريفه تعريفا دقيقا.

2- يمكن انتقاء عينة من البنود من هذا النطاق بطريقة هادفة ومناسبة.

3- يمكن تحديد عينة البنود وأسلوب المعاينات المستخدم وتعريفها بدقة كافية لكي يتمكن مستخدم الاختبار الحكم على مدى تمثيل عينة البنود للنطاق السلوكي الشامل الذي يقيسه.

– توجد مجموعة من الطرق المستخدم في تقدير معاملات صدق المحتوى تتمثل في:

1– نسبة البنود المزاوجة للأهداف السلوكية.

2– نسبة البنود المزاوجة للأهداف بتقديرات عالية الأهمية.

3– الارتباط بين الأوزان النسبية للأهداف والبنود التي تقيس هذه الأهداف.

4– معامل التوافق بين البند والهدف.

5– نسبة الأهداف التي لم تقيَم بأي بند في الاختبار (Crocker & Algina, 2006)

اقترح المعامل الرابع (4) من طرف Hambelton & Revenelli (1977) و Hambelton (1980) الذي يستخدم في تقييم إلى أي مدى تكون درجة صدق محتوى بند معين بالنسبة لمجموعة الأهداف. وتعتمد هذه الصيغة على افتراض أنه في الحالة المثالية يجب أن يزاوج البند هدفا واحدا فقط من مجموعة الأهداف، وتعكس طريقة جمع البيانات هذا الافتراض لأنه تم توجيه الخبراء لمزاوجة البند لكل هدف، وأن يعطوا القيمة (1) فيما لو كانت هناك مزاوجة والقيمة صفر (0) إن لم يكن متأكدا من المزاوجة والقيمة (−1) فيما لو لم يطابق البند مع الهدف بوضوح.

ويمكن حساب معامل توافق البند مع الهدف **K** بالصيغة المبسطة:

$$I_C = \frac{N}{2N-2}(\bar{X}_i - \bar{X})$$

N: عدد الأهداف.

$\bar{X}_i$: متوسط تقدير الأحكام على البند i للهدف k.

$\bar{X}$: متوسط تقدير الأحكام على البند i في جميع الأهداف.

أكبر قيمة محتملة لتوافق البند مع الهدف تساوي (1.00) عندما يتطابق البند مع هدف واحد من قبل المحكمين جميعا، وفي حالة تطابق بند واحد مع أكثر من هدف فان معامل التوافق يكون أقل من (1.00). وفي الحالة المثالية يجب أن تكون قيم c لكل بند في الاختبار عالية تقيس هدفاً صُمّم لقياسه، ومنخفض مع الأهداف الأخرى.

نشاط تدريبي:

يتضمن الجدول بيانات لتقديرات (03) محكمين حول درجة تطابق 5 بنود مع هدفين مختلفين، حيث يشير التقدير (−1) إلى درجة مطابقة منخفضة، ويشير التقدير (1) إلى مطابقة مرتفعة.

البنود	الهدف1			الهدف2		
	المحكم1	المحكم2	المحكم3	المحكم1	المحكم2	المحكم3
1	−1	0	0	1	1	1
2	1	1	1	−1	**−1**	**−1**
3	0	1	1	0	1	1
4	0	**−1**	**−1**	0	**−1**	**−1**

– أحسب معامل توافق البند مع الهدف لكل بند مع كل هدف.

الحل:

البنود	الهدف1 $\bar{X}_i$	$\bar{X}$	الهدف1 I_C	التوافق مع الهدف1	الهدف2 $\bar{X}_i$	$\bar{X}$	الهدف2 I_C	التوافق مع الهدف2
1	−0.33	0.33	−0.66	غير متوافق	1	0.33	**0.66**	**متوافق**
2	1	0.00	1	**متوافق**	−1	0.00	−1	غير متوافق
3	0.66	0.66	0	غير متوافق	0.66	0.66	0	غير متوافق
4	0.66	−0.66	0	غير متوافق	−0.66	−0.66	0	غير متوافق

إجراءات حساب معامل توافق البنود مع كل هدف:

البنود	الهدف1	الهدف2
1	$I_{C1} = \frac{2}{2\times 2-2}(-0.33) - 0.33) = -0.66$	$I_{C1} = \frac{2}{2\times 2-2}(1-0.33) = 0.66$
2	$I_{C2} = \frac{2}{2\times 2-2}(1-0) = 1$	$I_{C2} = \frac{2}{2\times 2-2}(0-(-1) = -1$
3	$I_{C3} = \frac{2}{2\times 2-2}(0.66-0.66) = 0$	$I_{C3} = \frac{2}{2\times 2-2}(0.66-0.66) = 0$
4	$I_{C4} = \frac{2}{2\times 2-2}(0.66-(-0.66) = 0$	$I_{C4} = \frac{2}{2\times 2-2}(-0.66 - (-0.66) = 0$

يتضح من خلال معاملات التوافق فان البنود التي تقيس الأهداف هي البند1 الذي يتطابق مع الهدف2 والبند 2 الذي يتطابق مع الهدف1، أما باقي البنود الأخرى فلا تتوافق مع الأهداف.

2.2. الصدق المرتبط بالمحك:

في العديد من الحالات يهدف مستخدم الاختبار لاستخلاص نتائج من درجات الاختبار لفحص السلوك على محك أداء معين لا يمكن قياسه مباشرة بالاختبار، ويناسب الصدق المرتبط بمحك المواقف التي نود فيها استخدام أداة قياس في تقدير سلوك معين ذي معنى، وهذا السلوك خارج نطاق الاختبار ذاته ويُعدّ بمثابة المحك.

فالصدق المرتبط بمحك يستند إلى الأسلوب الامبريقي في دراسة العلاقة بين درجات اختبارات أو مقاييس معينة تعدّ بمثابة منبّئات ودرجات مقاييس خارجية مستقلة تعدَ بمثابة محكات أداء عملية. ويتم تصميم دراسة الصدق المرتبط بمحك وفقا للخطوات التالية:

1- تحديد سلوك المحك المناسب وطريقة قياسه.

2- تحديد عينة مناسبة من ممثلة من المفحوصين الذين سيستخدم الاختبار لفئتهم.

3- تطبيق الاختبار والاحتفاظ بدرجة كل مفحوص.

4- عندما تكون بيانات المحك مناسبة، يُحصّل على قياس للأداء على المحك لكل مفحوص.

5- تحديد قوة العلاقة بين درجات الاختبار والأداء على محك.

تميّز أدبيات القياس والتقويم بين نوعين من الصدق المرتبط بمحك أحدهما يسمى **الصدق التنبؤي** والآخر يسمى **الصدق التلازمي**. يشير الصدق التنبؤي إلى تقدير مدى صلاحية الاختبار في التنبؤ بالأداء المستقبلي للفرد الذي يقاس باختبار محك باستخدام درجات اختبار يطبق عليه في الوقت الحاضر. تتعلق هذه الأدلة بدرجة العلاقة بين درجات اختبار ونمط معين من السلوك المستقبلي مما يمكن من التنبؤ بهذا السلوك، وذلك يتطلب مرور مدة زمنية بين الحصول على الدرجات في الاختبار التنبؤي، ودرجات الاختبار المحك الذي نقدّر في ضوئه صدق القرار.

فعلى سبيل المثال يمكن أن ترتبط درجات اختبار الاستعداد الدراسي بدرجات الطالب بالكلية بمقدار (0.60)، وبالتالي فان درجات اختبار الاستعداد الدراسي لها درجة صدق تنبؤي بالنسبة لدرجة الطالب بالكلية.

تتعدد طرق تقدير الصدق التنبؤي اعتمادا على مجالات استخدام الاختبارات والمقاييس ونوع القرارات التي تسترشد بالبيانات المستمدة من دراسات الصدق، من بينها:

1.2.2. طريقة الارتباط بين الاختبار والمحك:

تعتمد الطريقة على تطبيق الاختبار التنبؤي ثم الانتظار إلى حين حدوث السلوك المتنبأ به (المحك) للحصول على درجات الأفراد في اختبار المحك ثم إيجاد معامل الارتباط بين درجات الاختبار التنبؤي ودرجات الاختبار المحك. فإذا كان معامل الصدق التنبؤي مرتفع بدرجة كافية يمكن اعتبار القرارات المتخذة صادقة.

يجب التنويه إلى أن معامل الارتباط بين درجات الاختبار المنبئ ودرجات اختبار المحك لا يستخدم فقط "معامل بيرسون"، حيث عندما يتم تصنيف أداء الأفراد على كل من الاختبار المتنبئ والمحك (مثل، ناجح–راسب على المتنبئ، والنجاح–الفشل على المحك) يمكن تقدير مثلا معامل فاي" Phi أو "كابا" Kappa أو أي طريقة ارتباطية مناسبة للاستخدام مع البيانات التصنيفية.

نشاط تدريبي:

طُبّق اختبار لقياس الاستعداد الدراسي SAT على عينة من الطلبة المترشحين لشهادة البكالوريا، وفي نهاية السنة تم الحصول على معدّلاتهم في البكالوريا، والنتائج التالية توضح درجاتهم في اختبار الاستعداد الدراسي ومعدلاتهم في البكالوريا.

الأفراد	1	2	3	4	5	6	7	8	9	10
اختبار الاستعداد	670	300	430	320	550	520	670	280	750	600
معدل البكالوريا	12.86	11.23	13.79	10.05	13.35	14.56	16.12	12.20	15.65	14.75

– هل اختبار الاستعداد الدراسي صادق.

الحل:

للتحقق من الصدق المحكي لاختبار الاستعداد الدراسي تم استخدام معادلة "بيرسون" الخطي بين درجات الاختبار التنبؤي (نتائج اختبار الاستعداد) ودرجات المحك (معدل البكالوريا).

$$\mathbf{r} = \frac{\boldsymbol{n\sum xy - (\sum x)(\sum y)}}{\sqrt{[\boldsymbol{n\sum x^2 - (\sum x)^2}]\,[\boldsymbol{n\sum y^2 - (\sum y)^2}]}}$$

$$= \frac{\mathbf{10 \times 7084.5 - (5090)(134.56)}}{\sqrt{[\mathbf{10 \times 2848900^2 - (5090)^2}]\,[\mathbf{10 \times \sum 1844.06^2 - (134.56)^2}]}}$$

$$= \mathbf{0.80}$$

n	x	y	xy	x^2	y^2
1	670	12.86	8616.2	448900	165.38
2	300	11.23	3369	90000	126.11
3	430	13.79	5929.7	184900	190.16
4	320	10.05	3216	102400	101.00
5	550	13.35	7342.5	302500	178.22
6	520	14.56	7571.2	270400	211.99
7	670	16.12	10800.4	448900	259.85
8	280	12.2	3416	78400	148.84
9	750	15.65	11737.5	562500	244.92
10	600	14.75	8850	360000	217.56
Σ	**5090**	**134.56**	**70848.5**	**2848900**	**1844.06**

بلغت قيمة معامل الصدق التنبؤي بين درجات كل من الاختبار التنبؤي والمحك (0.80)، وبالتالي فان اختبار الاستعداد الدراسي صادق.

2.2.2. طريقة الانحدار للتنبؤ بدرجات المحك:

يمكن استخدام معادلة الانحدار في التنبؤ بدرجة الفرد في اختبار المحك بمعلومة درجته في الاختبار التنبؤي اعتمادا على قيمة معامل **الصدق التنبؤي** لكن يتطلب أن تكون العلاقة بين درجات الاختبارين خطية (مستقيمة).

ويمكن التوصل إلى معادلة خط الانحدار في مرحلة دراسة صدق الاختبار المستخدم في التنبؤ، إذ يمكن استخدام قيمة معامل الارتباط بين هذا الاختبار والاختبار المحك، وكذلك قيمة المتوسط والانحراف المعياري لدرجات كل منهما في التوصل إلى معادلة الانحدار:

$$Y' = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

حيث تتضمن هذه المعادلة متغيرين أحدهما المتغيَر المتنبئ x والآخر المتغير المتنبأ به Y'، أما بقية رموز المعادلة $\bar{x}$، σ_y ، $\bar{y}$ ، σ_x يتم حسابها في مرحلة دراسة الصدق. ويلاحظ أن:

$\bar{x}$ و σ_x : المتوسط الحسابي والانحراف المعياري لدرجات عينة الأفراد في المتغيَر المتنبئ.

$\bar{y}$ و σ_y : المتوسط الحسابي والانحراف المعياري لاختبار المحك المتنبأ بها في العينة التي تكون درجات أفرادها في الاختبار التنبؤي x.

نشاط تدريبي:

فرضاً أننا طبقنا مقياس لقياس الفعالية الذاتية الأكاديمية للتحقق من صدقه، حيث أن نظرياً يمكن أن نتنبأ بالفعالية الذاتية الأكاديمية من خلال درجاتهم في نهاي السنة الدراسية. وبعد إجراء مقياس الفعالية الذاتية الأكاديمية في بداية السنة، تم الحصول في نهاية السنة على درجاتهم في العلوم الطبيعية، والنتائج موضحة في الجدول التالي:

الأفراد	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
الفعالية الذاتية الأكاديمية	25	14	17	29	22	16	30	28	12	29
تحصيل العلوم الطبيعية	16	7	9	17	14	11	18	15	5	14

– ما مدى قدرة درجات العلوم الطبيعية في التنبؤ بالفعالية الذاتية الأكاديمية (معامل الصدق التنبؤي).

الحل:

n	x	y	xy	x^2	y^2
1	25	16	400	625	256
2	14	7	98	196	49
3	17	9	153	289	81
4	29	17	493	841	289
5	22	14	308	484	196
6	16	11	176	256	121
7	30	18	540	900	324
8	28	15	420	784	225
9	12	5	60	144	25
10	29	14	406	841	196
Σ	**222**	**126**	**3054**	**5360**	**1762**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{222}{10} = 22.2 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{126}{10} = 12.6$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{10 \times 3054 - (222)(126)}{\sqrt{[10 \times 5360 - (222)^2][10 \times 1762 - (126)^2]}}$$

$$= 0.936$$

معامل الارتباط بين المقياس المنبئ (الفعالية الذاتية الأكاديمية) ودرجات المتنبأ به (درجات العلوم الطبيعية) موجب قوي بلغ (0.936).

حساب معامل التحديد الذي يساوي:

$$R = r^2 = (0.936)^2 = 0.876$$

يعني معامل التحديد بأن المتغير المنبئ (درجات مقياس الفعالية الذاتية الأكاديمية) يفسّر نسبة (87.6%) من التباين في المتغير المتنبأ به (درجات تحصيل العلوم الطبيعية).

يتم حساب معادلة خط الانحدار y على x كما يلي:

$$\hat{y} = a + bx$$

$$b = \frac{\sum xy - n\bar{x}\,\bar{y}}{\sum x^2 - n\,\bar{x}^2}$$

$$= \frac{3054 - 10 \times 22.2 \times 12.6}{5360 - 10 \times 22.2^2} = \frac{256.8}{431.6} = 0.595$$

$$a = \bar{y} - b\,\bar{x} = 12.6 - 0.595 \times 22.2$$

$$= -0.609$$

معادلة الانحدار الخطي تساوي:

$$\hat{y} = (-0.609) + 0.595\,x$$

يتبيّن من خلال معادلة الانحدار الخطي أن الزيادة وحدة واحدة في المتغير المنبئ (الفعالية الذاتية الأكاديمية) تصحبها زيادة في المتغير المتنبأ به (درجات تحصيل العلوم الطبيعية)، بمقدار (0.595) أي بنسبة (59.5%). وهذا ما يؤكد على توفر مقياس الفعالية الذاتية الأكاديمية على دليل صدق تنبؤي.

أما **الصدق التلازمي** فيشير إلى العلاقة بين درجات اختبار وقياس محكي طُبَق الاثنان في الوقت نفسه، مثلا إذا تقدم المتكوّن (المعلم) لاختبار الورقة والقلم في المعرفة التدريسية ومن ثم أجريت عليه ملاحظة لتقدير أدائه أثناء عملية التعليم.

فالعلاقة الايجابية يمكن أن تكون مؤشرا للصدق التلازمي لاختبار المعرفة التدريسية، حينها يتعلق بدرجة اقتران تباين درجات الاختبار بتباين اختبار آخر يطبق في الوقت نفسه تقريبا وبالتالي يهتم الصدق التلازمي بالوصف مقارنة الصدق التنبؤي الذي يهتم بالتنبؤ.

على سبيل المثال يمكن للمعلم أن يقارن درجات اختبار تحصيلي مقنن في مجال دراسي معين بدرجات اختبار تحصيلي يعده بنفسه لطلابه في مجال معين، تعتمد خطوات تقدير الصدق التلازمي على تطبيق الاختبار المراد التحقق من صدقه التلازمي، ثم الحصول على درجات الأفراد في المحك بعدها إيجاد معامل الارتباط بين مجموعتي الدرجات.

نشاط تدريبي:

لفحص صدق مقياس المهارات الاجتماعية للأطفال تم تقييم مهاراتهم الاجتماعية باستخدام مقياس تقرير ذاتي لعينة من (12) طفلاً من قبل الوالدين (الأم خاصة)، وتقييم المهارات الاجتماعية للأطفال بالاعتماد على ميزان تقدير السلوكات الملاحظة من قبل المعلمين في الوقت نفسه كمحك. وبعد جمع البيانات تم عرضها في الجدول الآتي:

الأطفال	1	2	3	4	5	6	7	8	9	10	11	12
التقرير الذاتي	129	111	90	107	105	83	89	114	123	78	103	95
السلوكات الملاحظة	80	71	55	76	87	51	66	69	82	64	83	69

– قدّر معامل الصدق التلازمي لمقياس المهارات الاجتماعية للأطفال.

الحل:

$$\sum x = 1227 \qquad \sum y = 853 \qquad \sum xy = 88573$$

$$\sum x^2 = 128189 \qquad \sum y^2 = 61999$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{12 \times 88573 - (1227)(853)}{\sqrt{[12 \times 128189 - (1227)^2][12 \times 61999 - (853)^2]}}$$

$$= 0.702$$

معامل الارتباط بين درجات مقياس المهارات الاجتماعية للأطفال ودرجات المحك (درجات تقدير السلوكات الملاحظة من قبل المعلمين) بلغ (0.702)، وهو مرتفع يشير إلى تمتع مقياس المهارات الاجتماعية للأطفال بالصدق التلازمي.

تتأثر قيم معامل الصدق المرتبط بمحك بعوامل متعددة يجب على مستخدمي الاختبارات والمقاييس ومتخذي القرارات مراعاتها عند تفسير البيانات المتعلقة بهذا النوع من الصدق، وهي: مدى تجانس عينة الأفراد، وثبات درجات المحك، تأثر المحك بمتغيرات أخرى مثل، التحيَز، وعدد بنود الاختبار التنبؤي، والمدة الزمنية الفاصلة بين تطبيق الاختبار التنبؤي واختبار المحك، وعدد أفراد عينة المختبرين.

3.2. الصدق المرتبط بالمفهوم:

مصطلح صدق المفهوم قدمه (Cronbach & Meehl, 1955) ونال اهتماما وقبولا متزايدا من جانب الباحثين في القياس خاصة في السنوات الأخيرة، وذلك نتيجة غموض في كثير من المفاهيم النفسية الذي نتج عنه صعوبة في تطوير اختبارات ومقاييس أكثر صدقا. يتناول صدق المفهوم العلاقة بين نتائج الاختبارات وبين المفهوم النظري الذي يهدف الاختبار لقياسه، مثل مفاهيم الذكاء، القلق، الدافعية للإنجاز، الإبداع، الانبساط–الانطواء...وغيرها.

يركز صدق المفهوم على ثلاثة عناصر؛ الاختبار، والسمات المراد قياسها، وماذا يقيس الاختبار من وجهة نظر القائم بإعداده، أي أن العناصر تتعلق بالمفهوم والتفسير والنظرية التي يستند إليها التفسير. ويتضمن صدق المفهوم تجميع أدلة من سلسة من الدراسات تتضمن الخطوات التالية:

1–صياغة فرضية أو أكثر تبيّن الاختلافات المتوقعة في الخصائص الديموغرافية أو محكات الأداء أو مقاييس الأبنية الأخرى ذات العلاقة بمحك أداء تم تصديقه، هذه الفرضيات يجب أن تعتمد على نظرية مصاغة بوضوح يقع البناء ضمنها، وتزوّدنا بتعريف خاص للبناء.

2–اختيار أو تطوير أداة تتألَف من بنود تمثّل سلوكات مخصصة وواضحة بشكل ملموس للبناء.

3– جمع بيانات تجريبية تسمح باختبار العلاقة الافتراضية.

4– تحديد ما إذا كانت البيانات مطابقة للفرضية مع الأخذ بعين الاعتبار مدى إمكانية تفسير النتاجات الملاحظة بواسطة النظريات المنافسة أو التفسيرات البديلة.

من الواضح من الخطوات السابقة أن صدق الاختبار وصدق نظرية المفهوم الذي نهتم به لا يمكن الفصل بينهما، فإذا كانت العلاقة المفترضة مثبتة كما تنبأت به النظرية فان كلا من المفهوم والاختبار الذي يقيسه مفيداً، وإن لم تثبت الفرضية بواسطة دراسة الصدق فان مطوّر الاختبار لا يستطيع معرفة ما إذا كان هناك قصور في البناء النظري أو الاختبار الذي يقيسه أو كلاهما.

يتطلب جمع أدلة عن صدق المفهوم من مصادر متعددة أكثر من أي أسلوب صدق آخر، وتعدد أساليب جمع الأدلة منها: **دراسات تعتمد على الارتباطات، دراسات تعتمد على التجريب، دراسات تعتمد على التحليل المنطقي، دراسات تعتمد على تباين طرق القياس.**

تهتم الدراسات الارتباطية بتحديد طبيعة الفروق بين الأفراد الذين يحصلون على درجات مرتفعة والأفراد الذين يحصلون على منخفضة في الاختبار، وتحليل العلاقات بين درجات مجموعة من الاختبارات للكشف عن السمات التي تشترك في قياسها واختلافها عن غيرها من السمات. أما الدراسات التجريبية فتهدف إلى التعرف على التغيّرات التي تحدث في أداء الفرد في اختبار معين نتيجة التأثير التجريبي لأحد المتغيرات المتعلقة بالفرد للتحقق من صحة فرض معين يتعلق بما يقيسه الاختبار.

وتركز الدراسات المنطقية على محتوى الاختبار وطريقة تصحيحه أو تقدير درجاته، والعوامل التي تؤثر على الدرجات، وتركز أخرى على تباين ملاحظات الفرد في البناء عبر طرائق قياس مختلفة من خلال عناصر تحليل التباين الناتجة عن تطبيق نظرية إمكانية التعميم.

1.3.2. طرق تعتمد على الارتباطات:

تتعد الطرق المستخدمة في دراسات صدق المفهوم اعتمادا على الارتباطات، من بين الأساليب:

التمييز بين المجموعات: تهدف هذه الطريقة إلى المقارنة بين مجموعتين مختلفتين من الأفراد في ضوء الدرجات التي يحصلون عليها في اختبار معين، حيث يُتوقع اختلاف درجات الأفراد في الاختبار بين المجموعتين (Crocker & Algina, 2006).

فإذا أردنا التحقق مثلا من صدق المفهوم لاختبار يقيس جودة الحياة عند المصابين بالسرطان، فانه يمكن اختيار مجموعتين من الأفراد استنادا إلى تشخيص الأخصائيين النفسيين بحيث تكون هذه السمة مرتفعة لدى إحدى المجموعتين ومنخفضة لدى الأخرى، حينها يطبق الاختبار على المجموعتين. فإذا تبين وجود فروق جوهرية بين درجات كل من المجموعتين، فانه يمكن اعتبار

ذلك أحد أدلة صدق المفهوم للاختبار، ويُحتفظ حينها بالبنود التي ميّزت بدرجة أكبر بين المجموعتين.

نشاط تدريبي:

تم إعداد مقياس لتقييم الاكتئاب لدى مرضى المصابين بالسرطان، ولغرض التحقق من صدقه تم تطبيقه على مجموعة من الأفراد الذين تم تشخيصهم من قبل الاخصائيين النفسانيين بأنهم مكتئبين، ومجموعة من الأفراد الذين شخصوا بأنهم غير مكتئبين. وتحصلنا على النتائج من تطبيق المقياس، كما يلي:

الأفراد	1	2	3	4	5	6	7	8	9	10
مكتئبين	27	28	25	26	34	30	31	28	33	26
غير مكتئبين	20	18	25	19	18	17	15	22	16	18

– تحقّق من صدق مقياس الاكتئاب من خلال الفروق بين المجموعتين.

الحل:

للتحقق من الصدق التمييزي يتم المقارنة بين مجموعة الأفراد المكتئبين ومجموعة الأفراد غير المكتئبين، من خلال حساب اختبار "ت" لدلالة الفروق بين عينتين مستقلتين.

التحقق من تجانس المجموعتين:

$$S_1^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n\,(n-1)} = \frac{10 \times (8380) - (288)^2}{10\,(10-1)} = 9.51$$

$$S_2^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n\,(n-1)} = \frac{10 \times (3612) - (188)^2}{10\,(10-1)} = 8.62$$

يتضح أن هناك تقارب في تباين المجموعتين، مما يؤكد على تجانسهما. وهذا يسمح باستخدام اختبار "ت" في حالة التجانس، والتي يُعبّر عن معادلته كالآتي:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{[(n_1-1)s_1^2 + (n_2-1)s_2^2]}{n_1 + n_2 - 2}\right)\left(\frac{n_1 + n_2}{n_1 \times n_2}\right)}}$$

$$\bar{x}_1 = \frac{288}{10} = 28.8$$

$$\bar{x}_1 = \frac{188}{10} = 18.8$$

$$t = \frac{28.8 - 18.8}{\sqrt{\left(\frac{[(10-1)9.51 + (10-1)8.62]}{10+10-2}\right)\left(\frac{10+10}{10\times 10}\right)}} = \frac{10}{1.81} = 5.53$$

بالرجوع إلى القيمة الجدولية عند درجة حرية $(df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18)$ ومستوى الدلالة (0.05) التي تساوي (2.101) فإنها أصغر من قيمة "ت" المحسوبة (5.53).

وبالتالي توجد فروق دالة عند مستوى 0.05 بين مجموعة الأفراد المكتئبين ومجموعة الأفراد غير المكتئبين على درجات مقياس الاكتئاب. وهذا ما يؤكد على تمتع مقياس الاكتئاب الذي تم إعداده بصدق تمييزي.

الارتباط بين الطرق والسمات: يمكن للنتائج التي نحصل عليها في الاختبارات أن ترجع أيضاً إلى السمة المقاسة أو طريقة القياس (استبيان، قائمة ملاحظة، بطاقة تقرير ذاتي، مقابلة...). فإذا درسنا السمة نفسها أو سمات متباينة بطرق مختلفة فانه يمكن أن نحصل على نتائج مختلفة، لذا يجب دراسة أكثر من سمة وأكثر من طريقة معاً، أي ندرس ما يسمى بالصدق التقاربي والتمايزي.

في **الصدق التقاربي** ننظر إلى الارتباطات بين السمات نفسها إذا أجري قياسها بطرق مختلفة، وفي **الصدق التمايزي** ننظر إلى الارتباطات بين سمات متباينة إذا أجري قياسها بطريقة واحدة. فالأسلوب الأول **(الصدق التقاربي)** يمكن استخدام مكونات تحليل التباين للتعرف ما إذا كانت درجات الفرد على البناء لا تتباين عبر طرائق القياس المختلفة، فقد اقترح Kane (1982) استخدام هذه الطريقة من خلال تحليل مكونات التباين الناتجة عن تطبيق "نظرية إمكانية التعميم". ويعبّر عنه بالصيغة التالية:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2}$$

$\hat{\sigma}_p^2$: التباين الراجع للأفراد

$\hat{\sigma}_e^2$: تباين الباقي

على سبيل المثال أثناء تقييم الأداء في العلوم باستخدام طرق قياس مختلفة (الملاحظة، دفاتر الكتابة، المحاكاة بواسطة الكمبيوتر، الإجابة القصيرة)، حيث سوف نعتبر أن البناء أكثر أهمية (أو قابلية للتعميم) فيما لو كانت درجة المفحوص النسبية نفسها إذا استخدمت طرق قياس مختلفة.

ويشير Kane (1982) إلى أن معامل إمكانية التعميم الذي نحصل عليه من نظرية إمكانية التعميم هو معامل للصدق لأنه يمكن تفسيره على أنه متوسط معامل الصدق التجميعي الناتج عن الاختيار العشوائي لطرائق مختلفة تقيس السمة نفسها من نطاق الطرائق المختلفة.

نشاط تدريبي:

نفترض أننا نريد التحقق من صدق اختبار أداء الطلاب في العلوم الطبيعية باستخدام ثلاث طرق (الاجابة القصيرة، المعالجة اليدوية، المحاكاة بالكمبيوتر) طُبّقت على عينة من (10) طلاب في المتوسطة. وبعد تقييم أدائهم بواسطة طرق القياس الثلاث، تحصلنا على البيانات التالية:

الأفراد	1	2	3	4	5	6	7	8	9	10
الاجابة القصيرة	2	8	4	4	8	8	6	4	3	1
المعالجة اليدوية	3	5	2	3	5	5	4	3	2	2
المحاكاة بالكمبيوتر	2	7	2	6	5	7	5	3	2	3

الحل:

للتحقق من الصدق التقاربي باستخدام نظرية إمكانية التعميم، نتبع الخطوات التالية:

المتوسط الحسابي	طرق القياس			
(X_{pi})	المحاكاة بالكمبيوتر	المعالجة اليدوية	الاجابة القصيرة	الأفراد
2.33	2	3	2	1
6.67	7	5	8	2
2.67	2	2	4	3
4.33	6	3	4	4
6.00	5	5	8	5
6.67	7	5	8	6
5.00	5	4	6	7
3.33	3	3	4	8
2.33	2	2	3	9
2.00	3	2	1	10
(X_{PI})				المتوسط الحسابي
4.13	**4.2**	**3.4**	**4.8**	(X_{pi})

1- حساب مجموع مربعات الأفراد SS_p كما يلي:

$$SS_p = (2.33 - 4.13)^2 \times 3 + (6.67 - 4.13)^2 \times 3 + (2.67 - 4.13)^2 \times 3 + (4.33 - 4.13)^2 \times 3 + (6 - 4.13)^2 \times 3 + (6.67 - 4.13)^2 \times 3 + (5 - 4.13)^2 \times 3 + (3.33 - 4.13)^2 \times 3 + (2.33 - 4.13)^2 \times 3 + (2 - 4.13)^2 \times 3$$

$$= 9.72 + 19.253 + 6.453 + 0.12 + 10.453 + 19.253 + 2.253 + 1.92 + 9.72 + 13.653 = 92.80$$

2-حساب مجموع مربعات طرق القياس SS_m كما يلي:

$$SS_m = (4.18 - 4.13)^2 \times 10 + (3.4 - 4.13)^2 \times 10 + (4.2 - 4.13)^2 \times 10$$

$$= 4.444 + 5.378 + 0.044 = 9.867$$

3-حساب مجموع مربعات الباقي SS_r كما يلي:

$$SS_r = \Sigma(X_{pm} - X_{PM})^2 - SS_p - SS_m$$

$$\Sigma(X_{pm} - X_{PM})^2 = (2 - 4.13)^2 + (3 - 4.13)^2 + (2 - 4.13)^2 + (8 - 4.13)^2 + (5 - 4.13)^2 + (7 - 4.13)^2 + (4 - 4.13)^2 + (2 - 4.13)^2 + (2 - 4.13)^2 + (4 - 4.13)^2 + (3 - 4.13)^2 + (6 - 4.13)^2 + (8 - 4.13)^2 + (5 - 4.13)^2 + (5 - 4.13)^2 + (8 - 4.13)^2 + (5 - 4.13)^2 + (7 - 4.13)^2 + (6 - 4.13)^2 + (4 - 4.13)^2 + (5 - 4.13)^2 + (4 - 4.13)^2 + (3 - 4.13)^2 + (3 - 4.13)^2 + (3 - 4.13)^2 + (2 - 4.13)^2 + (2 - 4.13)^2 + (1 - 4.13)^2 + (2 - 4.13)^2 + (3 - 4.13)^2$$

$$= 4.551 + 14.951 + 0.018 + 0.018 + 14.951 + 14.951 + 3.484 + 0.018 + 1.284 + 9.818 + 1.284 + 0.751 + 4.551 + 1.284 + 0.751 + 0.751 + 0.018 + 1.284 + 4.551 + 4.551 + 4.551 + 8.218 + 4.551 + 3.484 + 0.751 + 8.218 + 0.751 + 1.284 + 4.551 + 1.284 = 121.467$$

$$SS_r = 121.467 - 92.800 - 9.867 = 18.80$$

4-حساب درجات الحرية للأفراد df_p والطرق df_m والباقي df_r، كما يلي:

$$df_p = n_p - 1 = 10 - 1 = 9$$

$$df_m = n_m - 1 = 3 - 1 = 2$$

$$df_r = (n_p - 1)(n_m - 1) = (10 - 1)(3 - 1) = 18$$

5-حساب متوسطات المربعات للأفراد MS_p والطرق MS_m والباقي MS_r، كما يلي:

$$MS_p = \frac{SS_p}{n_p - 1} = \frac{92.80}{10 - 1} = 10.311$$

$$MS_m = \frac{SS_m}{n_m - 1} = \frac{9.867}{3 - 1} = 4.933$$

$$MS_r = \frac{SS_r}{(n_p - 1)(n_m - 1)} = \frac{18.80}{(10 - 1)(3 - 1)} = 1.044$$

6-تلخيص النتائج في جدول تحليل التباين:

مصدر التباين SV	مجموع المربعات SS	درجات الحرية df	متوسط المربعات MS
الأفراد P	92.80	9	10.311
الطرق M	9.867	2	4.933
الباقي R	18.80	18	1.044

7-حساب مكونات تباين الأفراد $\hat{\sigma}^2$ وتباين الخطأ $\hat{\sigma}_e^2$ المتوقعة كما يلي:

$$\hat{\sigma}_p^2 = \frac{(MS_p - MS_r)}{n_m} = \frac{(10.311 - 1.044)}{3} = 3.089$$

$$\hat{\sigma}_e^2 = MS_r = 1.044$$

8- بعد حساب كل من مكون تباين الأفراد $\hat{\sigma}_p^2$ ومكون تباين الخطأ $\hat{\sigma}^2$ يمكن تقدير معامل إمكانية التعميم المتوقع، كما يلي:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2} = \frac{3.089}{3.089 + 1.044}$$

$$= 0.747 \cong 0.75$$

بلغ معامل إمكانية التعميم (0.75) وهو معامل الصدق التقاربي، حيث أن درجات الأفراد لا تتباين كثيراً عبر طرائق القياس المستخدمة (الاجابة القصيرة، المعالجة اليدوية، المحاكاة بالكمبيوتر). وبالتالي فالاختبار صادق لأنه يقيس القدرة نفسها عبر نطاق الطرق الأخرى الممكنة.

أما **الأسلوب الثاني (الصدق التمايزي)** يستند إلى الارتباطات التي تعتمد على إيجاد معامل الارتباط بين درجات اختبار ويفترض أن يقيس تكوينا فرضيا معينا، ودرجات اختبار آخر بيّنت الأدلة المتعددة أنه يقيس المفهوم ذاته. مثلا، اختبارات الذكاء التي يقوم الباحثون ببنائها تعتمد على إيجاد ارتباط درجات الاختبار الجديد بدرجات اختبار شائع الاستخدام ونال كثيرا من دراسات

الصدق، مثل اختبار Sanford–Binet أو اختبار Wechsler للذكاء، فإذا وجد قيمة معامل الارتباط مرتفعة فانه يُعدّ دليلا على صدق المفهوم.

نشاط تدريبي:

تم إعداد اختبار جديد لقياس الذكاء لدى الأطفال، وتم تطبيقه على عينة من الأطفال الذي طُبّق عليهم أيضاً اختبار الذكاء "وكسلر" Wechsler الشائع الاستخدام لغرض التحقق من صدقه، والنتائج المحصلة تم عرضها في الجدول الآتي:

الأطفال	1	2	3	4	5	6	7	8	9	10	11	12
الاختبار	125	140	130	105	90	110	110	120	105	95	90	70
اختبار "وكسلر"	55	80	100	65	50	80	75	80	55	60	90	70

– تأكد من الصدق التمايزي لاختبار الذكاء الذي تم إعداده، علماً أن البيانات لا تتوزع اعتدالياً.

الحل:

بما أن بيانات الاختبار الذكاء المعدّ وبيانات اختبار "وكسلر" لا تتوزع اعتدالياً فانه يتم تقدير معامل الارتباط "سبيرمان" Spearman للرتب عوضاً عن معامل "بيرسون".

الأفراد	x	y	R_x	R_y	D	D^2
1	125	55	**9**	**2.5**	**−6.5**	**42.25**
2	140	80	**12**	**9**	**−3**	**9**
3	130	100	**10**	**12**	**2**	**4**
4	105	65	**4.5**	**5**	**0.5**	**0.25**
5	90	50	**1.5**	**1**	**−0.5**	**0.25**
6	110	80	**6.5**	**9**	**2.5**	**6.25**
7	110	75	**6.5**	**7**	**0.5**	**0.25**
8	120	80	**8**	**9**	**1**	**1**
9	105	55	**4.5**	**2.5**	**−2**	**4**
10	95	60	**3**	**4**	**1**	**1**
11	90	90	**1.5**	**11**	**9.5**	**90.25**
12	135	70	**11**	**6**	**−5**	**25**
Σ	/	/	/	/	/	**183.5**

$$r_s = 1 - \frac{6 \sum D^2}{n\,(n^2 - 1)}$$

$$= 1 - \frac{6\,(183.5)}{12\,(12^2 - 1)} = 0.35$$

معامل الارتباط بين درجات اختبار ذكاء الأطفال الذي تم إعداده واختبار "وكسلر" لذكاء الأطفال منخفض بلغ (0.35)، حيث يفسّر التباين المشترك بينهما قيمة ضئيلة بلغت (12%) أي ($r_s{}^2 = 0.35^2 = 0.12$). وبالتالي فان الصدق التمايزي لاختبار ذكاء الأطفال ضعيف، فالاختبار غير صادق.

التحليل العاملي: يتناول أسلوب التحليل العاملي العلاقات القائمة بين الاختبار ومجموعة أخرى من الاختبارات بدلاً من اختبار واحد، ويستخدم في ذلك أحد الأساليب الإحصائية متعددة المتغيرات يسمى "التحليل العاملي"، والذي يهدف إلى تحليل العلاقات بتكوين مصفوفة لقيم معاملات الارتباط بين درجات الاختبار، ودرجات اختبارات أخرى مناسبة، وإجراء تحليل إحصائي لهذه المصفوفة لتحديد أقل عدد من العوامل (المفاهيم) التي تسهم في تفسير التباين في قيم معاملات الارتباط داخل المصفوفة، وكذلك العوامل التي تسهم في الأداء في كل هذه الاختبارات.

يوجد تطبيقين عامين في صدق المفهوم باستخدام التحليل العاملي:

1)- يتم معالجة مصفوفة الارتباطات الداخلية للبنود لتحديد ما إذا كانت الاستجابات على البنود تتجمّع معا في نمط معين يمكن التنبؤ به أو منطقي في ضوء التركيب النظري للبناء الذي نهتم به، وتحدَد ما إذا كانت الأبنية المحددة تجريبيا من خلال التحليل العاملي مناظرة للأبنية النظرية التي افترضها مطوَر الاختبار.

2)- يمكن معالجة مصفوفة ارتباطات مجموعة اختبارات أو قياسات مختلفة لتحديد مدى ارتباط الدرجات الملاحظة الذي يُعزى إلى تباين عامل مشترك واحد أو أكثر، وبالتالي التعرف ما إذا كانت الاختبارات الفرعية أو الاختبارات التي يُفترض أن تقيس البناء نفسه ثم تحديدها على أنها تقيس عاملا مشتركا.

مصفوفة السمات-الطرق المتعددة: ليس من الضروري الاقتصار على التعرف على العلاقات الموجبة بين الاختبار الجديد الذي يفترض أن يقيس تكوينا فرضيا معينا وغيره من الاختبارات

المشابهة التي تقيس المفهوم نفسه، وإنما أيضا التعرف على علاقته ببعض الاختبارات أو المقاييس التي لا تقيس هذا المفهوم.

طُوّر هذا الأسلوب من طرف Campell & Fiske(1959) للتحقق من ذلك اعتمدا فيه على مصفوفة تسمى "مصفوفة السمات–الطرق المتعددة" التي يمكن باستخدامها الكشف عن علاقات الاختبار باختبارات مشابهة للتوصل إلى الصدق التقاربي، وعلاقته باختبارات مختلفة عنه للتوصل إلى الصدق التمايزي.

يجب على الباحث في هذه الطريقة أن يفكر في طريقتين أو أكثر لقياس البناء الذي يهتم به، إضافة إلى تحديد بناءات أخرى مختلفة تماما يمكن قياسه بشكل مناسب بالطرائق نفسها المطبقة على البناء، وباستخدام عينة واحدة من الأفراد يتم الحصول على قياسات لكل بناء بكل طريقة، ثم تُحسب الارتباطات بين كل زوج من القياسات، وكل معامل ارتباط يعرف على أنه أحد الأنواع الثلاثة الآتية:

1)– معاملات الثبات: هي معاملات ارتباط بين قياسات البناء نفسه باستخدام طريقة القياس نفسها، ويجب أن تكون عالية.

2)– معاملات الصدق التجميعي: هي معاملات ارتباط قياسات البناء نفسه باستخدام طرائق قياس مختلفة، ويجب أن تكون عالية لكن احتمالية الضعف الناجمة عنه عدم ثبات طرائق القياس يجب أخذها بعين الاعتبار.

3)– معاملات الصدق التمايزي: هي معاملات ارتباط بين مقاييس بناءات مختلفة باستخدام طريقة القياس نفسها، ويُطلق عليها اسم معاملات الطريقة الواحدة–السمات المتجانسة، وهذه يجب أن تكون أقل بصورة أساسية من كلاً من معاملات الثبات أو معاملات الصدق التجميعي.

من أجل تسهيل المقارنة بين الأنواع المختلفة من المعاملات المحصلة يتم ترتيب المعاملات في مصفوفة السمات–الطرق المتعددة.

2.3.2. طرق تعتمد على التجريب:

تعتمد هذه الطرق على التدخل التجريبي لإحداث تغييرات في درجات الأفراد في اختبار معين كوسيلة للتعرف على مدى تأثَر الأداء بمعالجات أو متغيرات معينة مما يساعد في تأكيد بعض

التفسيرات المتعلقة بنتائج الاختبار أو رفضها، فإذا تأثّرت درجات الاختبار بمؤثرات معينة يؤدي ربما إلى الحد من تفسير الدرجات مما يضفي بعض الشك على صدق المفهوم للاختبار.

على سبيل المثال التحقق من صدق المفهوم لمقياس "القلق" من المهم تحديد مدى قدرة المقياس على تزويدنا بنتائج تتفق مع نظرية معينة تتعلق بالقلق، فإذا أوضحت النظرية أن القلق يزداد في موقف معينة تثير الإحباط فانه يمكن إجراء تجربة لتحديد ما إذا كان الأفراد الذين يواجهون أحد هذه المواقف يحصلون على درجات مرتفعة في المقياس أكثر من الذين لم يواجهوا هذا الموقف.

هذه الدراسات عبارة عن دراسات تجريبية من حيث التصميم، ويهدف التطبيق على الأفراد الذين تم إخضاعهم لمعالجة معينة صُمَمت لتغيير موقفهم على البناء عن أولئك الذين لم يتعرضوا للمعالجة، فان لم توجد فروقات بين المجموعتين فان التفسيرات المحتملة هي فشل في النظرية أو البناء أو كلاهما، أو عدم ملاءمة الأداة في قياس البناء أو فشل في المعالجة أو كلاهما.

3.3.2. طرق تعتمد على التحليل المنطقي:

تعتمد هذه الطرق على الفحص الدقيق للاختبار والأداء المطلوب، وإحداث تكامل بين نتائج هذا الفحص وبين النظرية التي يستند إليها الاختبار وآراء المختبرين الذين سبق أن اختبروا باختبارات مشابهة، ويعتبر Cronbach أن التحليل المنطقي من أهم مصادر التوصل إلى فروض بديلة فيما يتعلق بالأداء في الاختبارات، فالمحكم الذي لديه خبرة سابقة بالأخطاء التي شابت الاختبارات السابقة يمكنه أن يكتشف جوانب الضعف في أداة القياس الجديدة.

كما يمكن الاستعانة بأسلوب تحليل العمليات التي يستخدمها الأفراد في أدائهم أو في التوصل إلى إجاباتهم عن بنود الاختبار، لذلك عادة ما يرفق على سبيل المثال في اختبارات تقييم الأداء بتعليقات أو أسئلة تطلب من الأفراد تبرير اختياره للإجابة أو تبرير إجابته عن المهمات المطلوبة.

يساعد المفهوم في تحديد السمات التي يهدف الاختبار لقياسها، كما يساعد في بناء وتمحيص النظريات التربوية والسيكولوجية، وذلك عن طريق جمع أدلة ومعلومات متعددة لتغطية التكوينات الفرضية باستخدام الأساليب السابقة. ولأن صدق المفهوم يمكن تطبيقه على جميع أنواع الاختبارات وعلى نطاق واسع من استخدامات درجات الاختبارات فان التمييز بينه وبين الأسلوبين الآخرين للصدق قد يكون مصطنعا، لأن النوع أكثر ملاءمة للصدق يبقى محكوما بأنواع الاستدلالات المراد استنباطها من درجات الاختبار.

يتضح بأن أنواع الصدق الثلاثة؛ صدق المحتوى والصدق المرتبط بمحك وصدق المفهوم ليست متباينة وإنما جوانب متكاملة يتم التحقق بواسطتها من تأييد تفسيرات واستخدامات الدرجات المحصلة من أدوات القياس، لذلك فان الفصل بين هذه الجوانب أي دراسة أحدهما بدون أخرى يؤدي إلى أدلة غير كافية حول الصدق.

المحاضرة الثانية عشرة

العوامل المؤثّرة على صدق الاختبار، وعلاقة الصدق بالثبات

الأهداف:

- يتعرف الطالب على العوامل المؤثرة على صدق الاختبار.
- يميّز الطالب بين مفهومي الثبات والصدق.

1. العوامل المؤثرة على الصدق:

أشرنا في البداية إلى أن ثبات درجات الاختبارات تتأثر بعوامل متعددة، وهي: طول الاختبار، وتجانس عينة المفحوصين، وحدود الزمن، وخصائص بنود الاختبار، وموضوعية التصحيح. يمكن لهذه العوامل أيضا أن تؤثر على درجات صدق الاختبارات لأن معظم طرق تقدير الصدق تعتمد على التجريب الذي من خلاله نتحصل على درجات الاختبار.

كما يتأثَر الثبات بمصادر خطأ متعدَدة يمكن للصدق أن يتأثَر بمصادر خطأ استدلالاتنا وتفسيراتنا أيضا، حيث حدّد (Messick, 1995) مصدرين أو تهديدين أساسيين رئيسيين هما: ضعف تمثيل المفهوم والتباين غير الملائم للتكوين الفرضي. يتعلّق التهديد الأول للصدق بمحدودية وضعف شمولية محتوى الاختبار، والأبعاد، أو الأوجه المتصلة بالمفهوم موضع الاهتمام، ويتعلّق التهديد الثاني باتساع إضافي للتباين غير الملائم وغير المرتبط بتفسير المفهوم.

وقد تناولت وثيقة "معايير العملية الاختبارية التربوية والنفسية (2014) المصدرين وأشارت إلى ضعف تمثيل المفهوم إلى الدرجة التي يفشل فيها الاختبار في قياس المفهوم، وأشارت إلى التباين غير الملائم للتكوين الفرضي إلى الدرجة التي تتأثر بها درجات الاختبار بالعمليات غير جوهرية في المفهوم المطلوب(APA, AERA, & NCME, 2014) .

إضافة إلى هذين العاملين توجد عوامل أخرى يمكن أن تحد أو تخفَض من صدق التفسيرات منها ما يرتبط بخصائص الأفراد المختبرين (القلق، انخفاض الدافعية، تزييف الإجابة)، ومنها ما يرتبط بإجراءات تطبيق الاختبار وإجراءات تقدير درجاته (تقديم تعليمات غير مناسبة، محدودية وقت الاجابة، عدم اتساق أو تحيّز تقديرات المصححين)، ومنها ما يرتبط بالتعليم والتدريب الخاص للأفراد المختبرين.

2. العلاقة بين الصدق والثبات:

يهتم الثبات بالدقة التي يقيس بها الاختبار خصائص معينة، لذلك فهو مرتبط بخطأ القياس ويتم تعريفه أساساً كما أشرنا على أنه نسبة التباين الحقيقي إلى التباين الكلي للاختبار. ومن ناحية أخرى يتم تعريف الصدق على أنه جودة ما يتم قياسه؛ أي التشابه (أو التماثل) بين ما نريد قياسه وما يقيسه الاختبار فعلياً. وبشكل أساسي يعتبر الصدق إذن جزءاً من التباين الحقيقي ذات الصلة بأغراض استخدام الاختبار.

لتلخيص هذا التمييز بين الثبات والصدق، نلقي نظرة على تركيبة درجات الاختبار حيث أن الدرجة الملاحظة للفرد تساوي الدرجة الحقيقية مضافاً إليها درجة الخطأ، وأن الدرجة الحقيقية يمكن أن تجزئ إلى تأثير البُعد ذات الصلة أي الذي نريد قياسه، وتأثير الأبعاد غير ذي صلة ولكنه يتوافق مع التباين الحقيقي على القياس (ليس عشوائيا)، الاختبار يمكن للاختبار بالفعل قياس عدة أشياء.

وفي هذا الاطار تنقسم الدرجة الحقيقية إلى جزأين التباين ذي الصلة بالبُعد الذي يقيسه والتباين غير ذي صلة بما يقيسه، فاذا أخذنا تعريفنا للثبات والصدق فان الثبات هو نسبة التباين الحقيقي (ذي الصلة وغير ذي الصلة) إلى التباين الكلي، بينما الصدق هو نسبة التباين ذي الصلة إلى التباين الكلي.

يمكننا أن نستنتج ما يلي:

– إذا كان الاختبار ثابت، فإنه ليس بالضرورة صادقاً وأنه لكي يكون الاختبار صادقاً من الضروري أن يكون ثابتاً، أي أن التباين الكلي ليس مجرد خطأ القياس.

– ضمان صدق الاختبار لا يؤدي مع ذلك، فيما يتعلق بالثبات إلى واحد أو أكثر من المؤشرات التي يوجد إجماع بشأنها، فالتحقق من صدق الاختبار هو عملية تقدمية تبدأ بمجرد بناء الاختبار (التحقق من صحة المحتوى).

– التمييز بين الصدق–الثبات يعكس غياب الدقة خطأ غير ثابت أو عشوائي حول نقطة متوسطة (والتي قد تكون أو لا تكون الهدف). يعكس غياب الصدق خطأ ثابتًا يُبعد النتيجة عن الهدف المقصود. فالصدق والثبات مفهومان متميزان مرتبطان بعلاقة ضمنية.

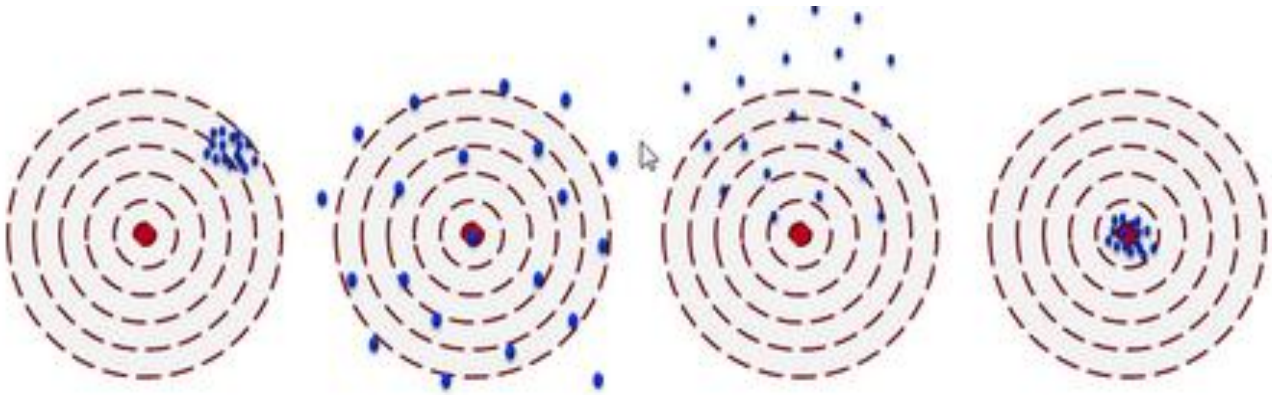
– العلاقة الضمنية **[الصدق ⟸ الثبات]**: الثبات شرط ضروري ولكنه غير كافٍ للصدق.

(1) الاختبار غير الثابت هو بالضرورة غير صادق.

(2) الاختبار الصادق هو بالضرورة، على الأقل، ثابت إلى حد ما.

(3) الاختبار الثابت ليس بالضرورة ثابتاً.

شكل (2): التمثيل البياني للعلاقة بين الصدق والثبات



صادق وثابت غير صادق وغير ثابت أقل ثباتاً وصدقاً ثابت وغير صادق

وفي السياق نفسه أشارت (Gipps, 1994) إلى الصدق تقليديا يعتبر أكثر أهمية من الثبات، فالاختبار عالي الثبات يكون استخدامه نادرا إذا لم يكن صادقا، والاختبار لا يكون صادقا –وفق النظرية الكلاسيكية للاختبارات– إذا لم يكن ذات مستوى أدنى من الثبات، فالكتابات في القياس تتجه نحو الاحتفاظ بأن الصدق أكثر أهمية من الثبات.

يرتبط مفهوم الصدق بمفهوم الثبات ارتباطا وثيقا، حيث لا يمكن أن تتصف درجات الاختبار بالصدق في استخدامات معينة دون أن تكون درجاته ثابتة، كما أن مؤشر الثبات يعدّ الحد الأقصى لمعامل صدق الاختبار. وتؤثر قيم معاملات الثبات في درجة العلاقة بين الاختبار التنبؤي واختبار المحك، كما تسهم قيمة معامل استقرار درجات الاختبار في زيادة الثقة بصدق المفهوم للاختبار لأنه يهتم بالثبات النسبي للسمات.

ويرى كرونباخ وزملائه Cronbach. Glaser & Rajaratnam أن الصدق والثبات مفهومان مترابطان ويمكن أن يندرجا تحت اسم "مقاييس إمكانية التعميم" فالفرق الرئيسي بينهما يكمن في الأبعاد التي نريد التعميم عليها، لذلك تعدّ نظرية إمكانية التعميم التي صاغها كرونباخ وزملائه من التطورات المعاصرة التي ساهمت في إبراز التكامل بين مفهومي الصدق والثبات (علام، 2000).

وبالتالي مفهوم إمكانية التعميم يربط بين الصدق والثبات في إطار النظرية الكلاسيكية للاختبار التي تعتمد على عينة من السلوك والهدف هو التعميم من العينة إلى نطاق ذلك السلوك، فإمكانية التعميم ضرورية لأنه لا يمكننا تقييم أداء الأفراد في كل المجال وإنما في عينة من الأداء، ومن أجل الوصول إلى التعميم بأي ثقة يجب الايفاء بالعديد من الشروط مثل التحديد بحذر السلوك (صدق المفهوم) والتقييم في حد ذاته يجب أن يكون ثابتا، وإن لم يكن التقييم ثابتا حينها التعميم عبر بعض الأبعاد كالبنود، والمقيّمين، والقياسات (الفترات) غير موثوق منه.

قائمة المراجع

التقي، أ. (2009). *النظرية الحديثة في القياس* .عمان: دار المسيرة.

الحجامي، ب. (2021). دراسة مقارنة في حساب الثبات بطريقة الاحتمال المنوالي لاختبار القدرة العقلية ثلاثية ورباعية البدائل. *مجلة كلية التربية الأساسية للعلوم التربوية والانسانية، 13(4)*، 25-61.

رينولدس، س ولوفنجستون، ر. (2013). *إتقان القياس النفسي الحديث: النظريات والطرق (ترجمة علام، م)*. عمان: دار الفكر.

علام، م. (2000) .*القياس والتقويم التربوي والنفسي*. القاهرة: دار الفكر العربي.

American Psychological Association. (n. d.). *APA dictionary of psychology*. Consulted 1 12, 2023. At: https://dictionary.apa.org/scale

APA, AERA, & NCME. (2014). *Standards for educational and psychological testing*. Washington: DC: Author.

Bertrand, R., & Blais, J. G. (2004). *Modèles de mesure: L'apport de la théorie des réponses aux items.* Québec: Presses de l'université de Québec.

Briesch, A., Swaminathan, H., Welsh, M., & Chafouleas, S. (2014). Generalizability theory: A practical guide to study design implementation, and interpretation. *Journal of School Psychology, 52*(1), 13–35.

Cardinet, J., Sandra, J., & Pini, G. (2010). *Applying generalizability theory using EDUG*. New York: Routledge.

Chadha, N. (2009). *Applied psychometry*. India Pvt Ltd : SAGE Publications.

Cohen, S. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7 Ed). USA: McGraw-Hill.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory* (2 Ed). Ohio: Cengage Learning.

Cronbach, L. (2004). My current thoughts on coefficient Alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 25*(4), 281–302.

de Gruijter, D., & van der Kamp, L. (2008). *Statistical test theory for the behavioral sciences.* London: Taylor & Francis Group.

De Ketele, J. (1993). L'évaluation conjuguée en paradigmes. *Revue Française de Pédagogie*(3), 59-80.

DeVellis, R. (2016). *Scale development: Theory and applications* (4 Ed). Los Angeles: Sage Publications.

Ebel, R., & Frisbie, D. (1991). *Essentials of educational Measurement.* New Delhi: Prentice-Hall of India.

Finch, W. & French, B. (2019). *Educational and psychological measurement.* New York: Routledge.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment.* London: Routledge-Falmer.

Hubley, A., & Zumbo, D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. Geisinger, et al., *APA handbook of testing and assessment in psychology* (pp. 3-19). Washington, DC: American Psychological Association.

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3 Ed.). Bruxelles: De Boeck.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Meyer, P. (2010). *Reliability.* New York: Oxford University Press.

Meyer, P. (2014). *Applied measurement with jMetrik.* New York: Routledge.

Miller, M., Linn, R., & Gronlund, N. (2009). *Measurement and assessment in teaching* (10 Ed). Upper Saddle River: Merrill/Pearson.

Niko, A., & Brookhart, S. (2014). *Educational Assessment of Students* (6 Ed). USA: Pearson New International Edition.

Popham, W. (2011). *Classroom assessment: What teachers need to know.* San Francisco: Pearson Education, Inc.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* California: Sage Publications.

Urbina, S. (2004). *Essentials of psychological testing.* New Jersey: John Wiley & Sons, Inc.