الفصل الثامن عشر

18

إعداد المعايير

الفصل الثامن عشر

إعداد المعايير

يتطلب العديد من المواقف تحديد درجة القطع قبل تفسير الأداء على الاختبار، على سبيل المثال: تقسم بعض البرامج التعليمية إلى جزأين، إذ يتقدم المفحوص لاختبار بعد إنهاء الجزء الأول ليسمح له بعدها بالتقدم إلى الجزء الثاني فقط إذا كانت الدرجة على الاختبار تساوي أو تتجاوز درجة القطع المحددة. وفي حالات أخرى مثل التأهيل لبعض التخصصات التطبيعية والتسكين يطلب من المتقدمين إكمال المعرفة المتخصصة في الاختبار. ويتم الموافقة على التأهيل فقط إذا تساوت درجة المتقدم ودرجة القطع أو تجاوزتها. وتطبق المواد درجات القطع تعرف عادة على إنهاء إعداد المعايير.

وعلى المدى الواسع يمكن القول بأن هناك ثلاثة أساليب لإعداد المعايير. تتطلب الأولى فحص محتوى الاختبار من قبل خبير أو أكثر والذي يصدر حكماً بالاعتماد على الانطباع الكلي عن محتوى الاختبار. ويعتمد الأسلوب الثاني على الحكم على محتوى الفقرات المنفردة. ووجهة نظر تقليدية لهذه الأساليب في إعداد المعايير هو أنها لا تستخدم القضايا السيكولوجية التقنية **(Gallagher, 1978 & Linn, 1978)** والأسلوب الرئيسي الثالث لإعداد المعايير يعتمد على أداء المفحصوين ويعد هذا التوجه سيكومتري نوعاً ما . ومع ذلك فإنه يتطلب عنصر مهم من الأحكام العامة، إذ أن جميع الأساليب التي تصنف تحت هذا الأسلوب تتطلب من معد المعايير اختيار مجموعة مفحوصين ثم اختبار أداءها. وهذه الأساليب البراغماتية الثلاث في إعداد المعايير سيتم وصفها في الأجزاء التالية من هذا الفصل.

والطريقة الأكثر شيوعاً لاستخدام درجة القطع الناتجة عن تطبيق أحد أساليب إعداد المعايير هو تطبيقها مباشرة على الدرجات الملاحظة. بمعنى أن معد المعايير لو أستنتج أن درجة القطع المناسبة كانت 0.69 على سبيل المثال، بعدها على المفحوص أن يجيب على 69% أو أكثر من فقرات الاختبار إجابة صحيحة لاجتياز الاختبار. وطريقة بديلة لاستخدام درجة القطع هي تطبيق درجات النطاق. وفي هذا الأسلوب تعالج فقرات الاختبار على أنها عينة من مجال أكبر من الفقرات، وكل مفحوص يعد حاصلاً على درجة النطاق (أو الحقيقة) والتي تحدد عادة على أنها نسبة الفقرات من النطاق التي يستطيع المفحوص الإجابة عنها إجابة صحيحة. وتستخدم درجة القطع التي يضعها معد المعايير لتجزئة تدريج درجة النطاق إلى منطقتين، وهذه تسمى درجة قطع تدريج النطاق. وفي اختبارات التحصيل تسمى هذه

المناطق حالات السيطرة، فالمفحوصين الذين درجاتهم تساوي درجة القطع أو أعلى منها يطلق عليهم أسم مسيطرين حقيقيين، وأما المفحوصين الآخرين فيطلق عليهم أسم غير مسيطرين حقيقيين. وحالما يتم وضع درجة قطع على متصل النطاق، فإن المشكلة تكمن في تحديد درجة القطع الملاحظة التي تسمح لمستخدم الاختبار استخلاص الاستنتاجات الأكثر ملائمة لحالات السيطرة الحقيقية للمفحوص. ويمكن أن يرافق تحديد درجة القطع الملاحظة الأخذ بعين الاعتبار العلاقة بين الدرجات الملاحظة ودرجات النطاق والتي يرافقها خطأ تصنيف المسيطر الحقيقي على أنه غير مسيطر وغير المسيطر الحقيق على أنه مسيطر.

ويطرح هذا الفصل أساليب تحديد درجة القطع الملاحظة والقضايا المتعلقة بها. تحت عنوان الاعتبارات التقنية في إعداد المعايير.

وقبل الرجوع إلى النطاق الرئيسة في هذا النقاش يجب أن نعرج على نقطتين ثانويتين. الأولى أننا حددنا النقاش في إعداد درجة قطع واحدة تجزئ متصل الدرجة إلى منطقتين اثنتين بهدف الحفاظ على بساطة العرض ومن ناحية أخرى يبدو أن استخدام درجة نجاح واحدة هي الحالة الأكثر شيوعاً في التطبيقات العملية. وبغض النظر عن حاجة مواقف لإعداد معايير بعينها والتي تتطلب استخدام عدة درجات قطع من أجل تقسيم متصل الدرجة إلى عدة مناطق. كمبدأ، فإن جميع الأساس المعروضة في هذا الفصل يمكن تعميمها لتناول مشكلات تتطلب درجات قطع عديدة. والثانية: هي استخدام درجات القطع في اختبارات القبول والتوظيف عندما تستخدم الاختبارات في التنبؤ عن حالات السيطرة المطلقة على محك خارجي. وهذه الموضوعات عرضت في الفصلين الحادي عشر والثاني عشر وان نتعرض إليها مباشرة في هذا الفصل.

الأساليب لإعداد المعايير:

مع أن أدب القياس وصف أكثر من 30 منهجية لإعداد المعايير في السنوات الأخيرة **(Behuniak, Archam bault & Gable, 1982)**. إلا انه يمكن تصنيفها ضمن واحدة من فئات ثلاثة رئيسة (ويمكن للقارئ المهتم بمخططات تصنيفية إضافية الرجوع إلى **(Glass, 1978, Hambelton, 1980 , Gaeger, 1979, Meskauskas , 1976 , Millman, 1973).** وتستخدم المنهجيات الرئيسة المعتمدة هنا:

1/ أحكام تعتمد على انطباع شامل عن الاختبار أو ملف الفقرات

2/ أحكام تعتمد على محتوى فقرات الاختبار منفردة.

3/ أحكام تعتمد على أداء المفحوصين في الاختبار.

الأحكام المعتمدة على الانطباع الشامل:

هنا يفحص فريق من الخبراء محتوى الاختبار، واعتماداً على الانطباع الكلي ومجال المحتوى تقترح النسبة التي يجب أن تكون الإجابة عليها صحيحة، وذلك لمن سيكون له أدنى مستوى من الكفاءة ليكون أداؤه ضمن المستوى المطلوب للأداء وكل محكم عادة يضع معيار معين، وفي المعيار النهائي يؤخذ متوسط المعايير جميعها وقد اقترحت شبرد (Shepard, 1976) استخدام مجموعة أحكام متعددة تمثل وحدات متعددة لمن يهتم بنتائج الاختبار. فعلى سبيل المثال في شهادة الكفاءة الجامعية الأولية يمكن ضم فئات من الطلبة والمعلمين وأولياء الأمور وممثلين عن المجتمع المحلي على مدى واسع .

وتعتمد هذه الأحكام في بعض الأحيان على السماح بخطأ، وأشار جلاس (Glass, 1978) إلى هذا الاستخدام على أنه "العد العكسي الذي يبدأ من 100% ، وهو أن معد المعايير يفترض مستوى مرغوب أو مفصل للأداء على الاختبار يجب أن تكون الإجابة على 100% منها إجابة صحيحة ولكنها تسمح ببعض الأخطاء ، الناتجة عن خطأ القراءة أو عدم وضع إشارة أو عدم الانتباه أو أخطاء التصحيح وهكذا.

فقد تستخدم بعض المعرفة عن مجتمع المفحوصين من قبل المحكمين. فعلى سبيل المثال أورد جلاس في تقرير عن طفل في الصفوف الابتدائية يفيد أنه يمكن تدريب هذا الطفل فقط لغاية 70% من الدقة في الجمع لمنزلة واحدة. ومن غير الشائع وجود معلمين يقترحون نسبة ما بين 60–70% على أنها معيار معقول لاختبارات من هذه النوعية، وقد يكون اقتراحهم هذا بسبب خبرتهم مع مفحوصين عند هذا المستوى.

ومع أن الأحكام الشاملة هي واحدة من المنهجيات الأكثر استخداماً في استخداماً في إعداد المعايير، فإنه من الصعب الدفاع عنها عن وجهة نظر منطقية أو سيكومترية ونقد شائع لها هو أن مطور الاختبار لا يمكنه معرفة ما إذا كانت عينة خبراء مختلفين لم تؤسس المعيار عند نقاط مختلفة. واحد الحلول المنطقية لهذه المشكلة تظهر الحاجة لإجراء دراسات مكررة .

وعلى افتراض أن العدد الكلي في الحكام المتوافر والمناسب ثابت، وفي حالة التكرار مرتين فإن عدد الخبراء لكل دراسة سينخفض إلى النصف، وبالتالي فإن مجموعة المعايير تعد من قبل عدد أقل من الخبراء، وهذه قد تؤدي إلى انحرافات من عينة لأخرى. أكثر مما هو للمعايير التي تعد من قبل عدد أكبر من الخبراء . ومشكلة أخرى أنه لا يوجد تأكيد بأن الخبراء المختلفين تعتمد على انطباعاتهم على مظاهر الاختبار نفسها أو أن تتشابه إدراكاتهم لمجال المحتوى، وهذه قد تؤدي إلى انحرافات غير مرغوب بها للمعيار من عينة خبراء لأخرى. وعند استخدام هذه المنهجية يجب على مطور الاختبار أن يوفر على الأقل توثيقاً لعدد الخبراء ومؤهلاتهم وعملية اختبارهم والتعليمات التي زودوا بها وتوزيع استجاباتهم.

الأحكام التي تعتمد على محتوى الفقرة:

تعد هذه المنهجية من منهجيات إعداد المعايير المنهجية التي درست على المدى الأوسع انتشاراً، وهناك ثلاثة طرائق مشهورة في هذه الفئة سيؤخذ بها، وبالترتيب التاريخي لها. فقد اقترح نيدليسكي **(Nedelskyo, 1954)** التقنية الأولى، والتي صححت على وجه الخصوص لفقرات من نوع اختبار من متعدد. وقد اهتم نيدلسكي بتأسيس معايير لأقل كفاءة أو تأهيل للمفحصوين في المستوى الجامعي.

وكانت المعايير التي حددها باستخدام هذه على النحو الآتي:

1/ يعطي كل محكم تعليمات (وعادة يكون خبير مؤهل في مجال المحتوى) لوضع إشارة لكل فقرة يبين عدد الاستجابات التي يمكن للمفحوص الأقل كفاءة (وهي لأقل طالب) أن يحذفها.

2/ يسجل الحكام ولكل فقرة مقلوب عدد الاستجابات . على سبيل المثال لفقرة من خمس بدائل، فلو استخرج بديلين فإن تسجيل قيمة المقلوب تكون $1/3$.

3/ إيجاد M وهي مجموعة المقلوب عبر الفقرات جميعها، ويمكن عدّها الدرجة المحتملة للمفحوص الأقل تأهيلات، كما حددت من قبل المقدرين لذلك الحكم المفرد.

4/ تؤخذ متوسطات قيم M عبر جميع المتعلمين (σ_M) . وقد اقترح ندلسكي أما الدرجات الناجحة في الاختبار يجب أن تساوي $\sigma_M + \mu_M$ حيث k قيمة اصطلاحية تختار بشكل مناسب، وتتراوح قيمتها بين 0.5 و .10 وقد تم نقد الاقتراض المنطقي المستخدم في اختبار K وتعديل قيمة μ_K ، لذلك فإن بعض مستخدمي هذه التقنية يفضلونه البساطة ووضع أقل نقطة نجاح عند μ_M . **(Meskauskas,1976)**. واقترحت المنهجية الثانية من قبل انغوف **(Angoff, 1971)** ووصفها بالأساس في ملاحظة تذيلية وذلك في توضيحه لكيفية إجراء تحويلات للتداريج المختلفة بدون بيانات معيارية. وتعطي هنا للمحكمين تعليمات بالأساس للتفكير في المجموعة الأقل قبولاً، وكذلك تقدير نسبة المجموعة الأقل قبولاً، وكذلك تقدير نسبة المجموعة الأقل قبولاً (ولكل فقرة) والتي تستطيع الإجابة على الفقرة إجابة صحيحة (ويمكن التفكير بهذا على أنه احتمالية الإجابة الصحيحة على الفقرة لمن لديه الحد الأدنى المقبول من الكفاءة) وثم تجمع هذه الاحتمالات عبر الفقرات جميعها للحصول على أدنى درجة نجاح وضعت من قبل المحكم المفرد . والقيمة المتفق عليها لتقديرات المحكمين جمعيهم تعد أقل درجة نجاح.

واقترح أيبل **(Ebel, 1971)** نظام مشابه لا نخوض ولكنه ميز أن كل محتوى الفقرة المناسب ومستوى صعوبتها قد تؤثر على المحكمين حول الأداء المتوقع للمفحوص الأقل كفاءة على الفقرة. وتستخدم هذه التقنية شبكة ذات بعدين في تصنيف الفقرات (رباعية الخلايا) .

أحد أبعاد الشبكة هو مستوى قبول الفقرة، أما البعد الثاني فهو للصعوبة وعادة ما يكون ذو ثلاثة مستويات انظر الجدول (18–1) . وفي البداية تصنف الفقرات في خلايا الشبكة، وبعدها يؤشر المحكم النسبة المئوية للفقرات في الخلية ويجب أن يجيب عليها المفحوص الأقل كفاءة إجابة صحيحة. ويمثل جدول (18–1) نسب مئوية افتراضية في كل خلية والتي يجب الحصول عليها من قبل كل محكم. وكذلك مبين عدد الفقرات لكل خلية في الاقواس أسفل النسبة المئوية. وتحسب أدنى درجة نجاح موصى بها من قبل محكم واحد بالمعادلة:

$$Xc = P\ 3\ (M) \quad \text{.....} \quad (18\text{-}1)$$

جدول (18–1) : جدول توضيحي ونسب الفقرات المصنفة من قبل محكم واحد باستخدام منهجية ايبل في إعداد المعايير لاختبار مؤلف من 200 فقرة

مستوى القبول	مستوى الصعوبة		
	سهل	متوسط	صعب
اساسي	%90	%50	%10
	20 فقرة	(25)	(5)
مهم	%60	%30	%20
	(35)	(22)	10
مقبول	%40	%20	%10
	(19)	(12)	(15)
عليه استفسار	%25	-	-
	(7)	(20)	(10)

$$X = 0.90 + (20)0.50 + (25)0.1 + (5)0.6 + (35)0.30 + (22)0.20 + (10)0.40 + (19)0.20$$
$$= 0(10) + 0(20) + 0.25(7) + 0.10(15) + (12) = 73.85.$$

حيث Xc درجة القطع بدلالة الدرجات الخام، و p نسبة الفقرات في الخلية التي يجيب عنها المفحوصين الأقل كفاءة.

وM هي عدد الفقرات في الخلية

والمجموع هو المجموع الكلي للخلايا الاثني عشر

ويمكن الحصول على درجة النجاح النهائية بحساب متوسط قيم Xc لجميع المحكمين. ويمكن تحويل هذه القيمة إلى نسب الإجابات الصحيحة لدرجة النجاح وذلك بالقسمة البسيطة على العدد الكلي للفقرات في الاختبار.

الأحكام المعتمدة على أداء المفحوصين:

ينادي العديد من أنصار الاختبارات محكية المرجع بإعداد معايير للاختبار اعتماداً على المعرفة المبينة على الأداء وذلك أثناء محاولاتهم في تطبيق الاختبار، لذلك لا بد من اختيار المحكمين الذين سيشاركون في إعداد المعايير بدقة وذلك لأن لديهم الخبرة لما سيكون عليه أداء المفحوصين النموذجي على مثل هذا الاختبار.

على سبيل المثال ففي إعداد معايير الأداء لاختبار في الجغرافية الجوية في المدرسة الثانوية فنحن نفضل استخدام معلمي المدارس الثانوية كمحكمين بدلاً من اساتذة الجامعات في الطبوغرافيا. لتمييز هذا فقد اشار شبرد **(Shepard, 1979)** إلى ان معايير المحكمين هذه تتأثر بشكل ملحوظ بادراكهم لمعرفة كيفية اداء المفحوصين على الاختبار. ومن الأنسب درجة أكثر استخدام بيانات حقيقية لعينة مفحوصين مختارة بشكل مناسب بدلاً من الاستناد إلى محكمين مناسبين معتمدين بدرجة أكبر على انطباع خاص لمدركاتهم حول قدرات المفحوصين في الأداء على اختبار معين.

وأسلوب مبسط يستخدم البيانات المعيارية هو تطبيق الاختبار على مجموعة مفحوصين قدرتهم أقل من قدرة الفئة التي يستهدفها الاختبار، ومنها إعداد معايير لأقل كفاءة استناداً إلى وسط أداء المجموعة أو وسيطها على سبيل المثال، وصف جلاس **(Glass, 1978)** إعداد معايير أدنى كفاءة لامتحان شهادة الدراسة الثانوية مستنداً على وسيد أداء طلبة الصف التاسع. ومثال آخر لهذا الأسلوب هو الاختبار المعياري المستخدم في غربلة الأطفال ذوي صعوبات اللغة والذي يحدد درجة قطع لكل مستوى صفي بحيث تصلى بمقدار انحراف معياري واحد عن متوسط طلبة المستوى الصفي نفسه.

وأسلوب أخر يستخدم بيانات مجموعات مفحوصين مختلفين اختلافاً واضحاً بمستويات الكفاءة على المادة التي سيختبرون بها، وتحدد درجة القطع بحيث تعظم التمييز بين هاتين

المجموعتين وواصف بيرك **(Berk, 1976)** منهجية تستخدم مجموعات مفحوصين احداها درست المادة التعليمية . والأخرى لم تدرسها ووصف نبدلسكي **(Nedlsky, 1954)**.

منهجية مجموعات المقارنة الإعداد المعايير والتي تتألف من ستة خطوات وعلى النحو الآتي:

1/ اختيار محكمين مؤهلين لديهم مألوفية بمجتمع المفحوصين

2/ السماح للمحكمين بمناقشة – إن أمكن– والموافقة على مكونات أداء ذوي أقل كفاءة مطلوبة.

3/ استخدام الأحكام لتحديد المفحوصين الكفؤين وغير الكفؤين، واستثناء أي واحد يظهر في منطقة الوسط.

4/ اختبار مجموعتي المفحوصين.

5/ رسم توزيع درجات المفحوصين لكلا المجموعتين على المتصل نفسه.

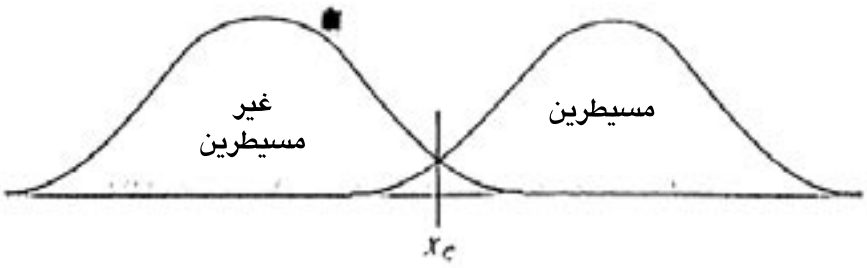
6/ إعداد معيار الأداء عند نقطة تقاطع منحنى التوزيعين (أنظر شكل (18–1)).

وطرائق أخرى لإعداد المعايير لبيانات مجموعتي مقارنة وصفها كلوفر**(Kloffer, 1980)** وزيكي وليفنجستون **(Zieky, Livingston, 1977)**.

وطريقة أخرى، اقترحت من قبل نيدلسكي تسمى منهجية مجموعة الخط الفاصل (Borderline group method) وفيها يتم اختيار المحكمين وإعطاء المعلومات كما في الخطة (2،1) اعلاه، ولكن يسأل المحكمين فيما بعد لتحديد المفحوصين الذين يظهروا في منطقة الخط الفاصل في الكفاءة . وتختبر هذه المجموعة بعد ذلك ويستخدم وسيط توزيع درجاتهم في تحديد نقطة مستوى أقل كفاءة مطلوبة.

ويشير نقد هذه الأساليب إلى أن المعايير المعدة بناءاً على أداء المفحوصين تتناقض مع الهدف الأساسي للاختبارات محكية المرجع لأنه لا يوجد أية إشارة لمحتوى الاختبار في علامة أو دلالة معايير الأداء. وأوصى جلاس **(Glass, 1978)** بالإضافة إلى أن استخدام مجموعة درست المادة التعليمية والأخرى لم تدرسها. يبدو أنه محاولة ضعيفة لتعويض وقت الدارسة Seat time لاثبات الكفاءة. ونقد مشابه يطبق عندما تتألف المجموعات المحكية من اعضاء ممارسين للتخصص أو الوظيفة وطلبة طموحين يبحثون عن الحق في ممارسة الوظيفة أو التخصص. وفيما لو أعد معيار يميز بين مثل هاتين المجموعتين، فإن هذا يعد افتراض واضح لصدق شهادة الخبرة السابقة والتي تستدعي منطقياً مسألة الحاجة لاختبارات حالية التي وضعت المعايير من أجلها ومع أن هذه الانتقادات تستحق أن تؤخذ بعين الاعتبار عند تأسيس المعايير على أساس وحيد لأداء المفحوصين ويوجد دعم أكثر لاستخدام بيانات

الأداء في عملية إعداد المعايير كجزء من المعلومات المدعمة والتي يمكن استخدامها في مناقشة نتائج معايير الأداء التي تم التوصل إليها من خلال منهجيات وطرائق أخرى .



شكل (18–1) اعداد درجة قطع عند تقاطع منحنيات التوزيع التكراري مجموعتي المقارنة

افترض على سبيل المثال أن معيار 90% صح ثم اقتراحه لمستوى الدخول في تخصص الصحة ، ولكن أشار الاختبار الميداني التمهيدي إلى أن أقل من 15% من الخريجين الحديثين في برنامج البكالوريوس لهذا التخصص يمكن أن ينطبق عليهم هذا المعيار. وتوحي هذه المعلومات بأن رخصة الممارسة تواجه معضلة جدية، فقد تكون البرامج الأكاديمية الحالية لا تعد الأغلبية العظمى من طلبتها للحياة العملية الوظيفية، أو أن عملية بناء الاختبار وصدق المحتوى النهائي معرض للمساءلة، أو أن المعيار الذي أعد كان غير منطقي. وتستحق الاحتمالات الثلاثة المذكورة جميعها أن تؤخذ بعين الاعتبار، وذلك قبل تأسس البرنامج الاختباري والتوصية بمعايير الكفاءة.

اقترح جايجر **(Jaeger, 1982)** عملية تجمع ما بين ظواهر الحكم الشامل على أداء المفحوصين وطريقة الحكم على الفقرات ويطلق على هذه الطريقة اسم عملية التحكيم المركب المتكرر على الفقرة، وتم إنشاء تطبيقاتها في تأسس معايير اختبارات كفاءة المدرسة الثانوية في كارولينا.

وقد استخدمت هذه الطريقة ثلاث مجموعات من المحكمين: معلمة المدارس الثانوية ومدراء المدارس والمرشدين ومواطنين في الولاية وسئل كل محكم لاكمال لاختبار وتدريج كل فقرة على مقياس نعم – لا ليجيب على لاسؤال:

" هل نستطيع كل منتظم في الدراسات الجامعية الأولية في شمال كارولينا أن يجيب على هذه الفقرة إجابة صحيحة؟" اعتقد جايجر أن هذه الكلمات اللطيفة أظهرت الحاجة لتعريف

البناء مثل ادنى كفاءة لازمة والسماح للمحكم ليركز على الأفعال الأكثر قابلية للملاحظة للحصول على الدبلوما.

ويتم تحديد درجة النجاح الموصى بها من قبل كل محكم بالجمع عبر الفقرات (بإعطاء درجة واحدة لكل إجابة نعم وصفر لكل إجابة لا) ويحدد توزيع درجة النجاح الموصى بها لكل مجموعة، وكذلك توزيع استجابات نعم – لا لكل مجموعة وعلى كل فقرة تم تسجيلها. ومن ثم يشخص المحكمين الأداء الفعلي لعينة مفحوصين في الصف الحادي عشر على الفقرة (قيمة p)، وكيفية تقدير المحكمين الآخرين للفقرة نفسها. ويسمح للمحكمين بتعديل أحكامهم الأصلية في ضوء هذه المعلومات وفي دورتين متتابعتين. ويستخدم وسيط القيم لتوزيع درجات النجاح الموصى بها من قبل كل مجموعة الأكبر للمعايير الموصى بها قد تظهر بالدورات المتتابعة.

وهذا ما لم يظهر بالفعل في دراسة جايجر. ومن الواضح أن المحكمين المختلفين يستخدمون المعلومات الإضافية بطرائق مختلفة عندما يعيدون تقويم الفقرات.

الأبحاث التجريبية في منهجيات إعداد المعايير:

من خلال الأساليب المتعددة، من الطبيعي أن يطرح جماعة القياس اسئلة من مثل: هل تؤدي الطرائق المختلفة في إعداد المعايير لنتائج ملحوظة مختلفة؟ وهل تعد مجموعات المحكمين المختلفين معايير متشابهة باستخدام التقنية نفسها؟ وما هي العوامل المصاحبة أو المرافقة للتباين في إعداد المعايير من قبل محكمين مختلفين؟ لغاية الآن، أجريت دراسات تجريبية عدة للإجابة عن هذه الأسئلة، ولكن لم يصل إلى الإجابات المحددة. وسنناقش هنا بعض النتائج التوضيحية للاثبات.

دعنا أولاً نجري مقارنة عبر طرائق أحكام الفقرة، ففي إحدى الدراسات وجد أندروز وهيشت؟ **(Andrews & Hecht 1971)** أن محكمين ثمانية (كمجموعة) استخدموا طريقة ايبل وإعداد معايير الكفاءة الأقل المطلوبة عند 68% إجابات صحيحة، ولكن معيار 49% إجابات صحيحة للمحكمين أنفسهم حصلوا عليها باستخدام طريقة نيدلسكي .

ولاحظ جلاس **(Glass , 1978)** من خلال النتائج غير المعروفة أن 95% من المفحوصين حققوا محك نيدلسكي، ولكن 50% فقط تطابقوا مع محك ايبل. ودراسة أكثر حداثة لبيونياك وزملاؤه **(Behuniak , eta 1982)** أجروا خلالها مقارنة بين طريقتي نغوف ونيدلسكي في إعداد معايير اختبار تحصيلي محكي المرجع في القراءة والرياضيات وأسس محكمين عددهم ستة لاختبار القراءة متوسط نسبة إجابة صحيحة 56.7% كمعيار بطريقة انغوف و 43.2% بطريقة نيدلسكي.

وأما في اختبار الرياضيات فكانت النتائج عكس ذلك، وإذ كانت النسبة 70.2% بطريقة انغوف و 77.1% بطريقة نيدلسكي . ومع أن دراسات المقارنة هذه لا توفر أدلة لأفضلية طريقة على غيرها، إلا أنها أثبتت بوضوح أن طرائق تحكيم الفقرة المختلفة لا يمكن عدّها توائم في إعداد المعايير . وكانت النتائج غير متطابقة في الدراسة الأخيرة وذلك لأن مجموعة المحكمين قسمت إلى نصفين، وهنالك فروق أساسية بين المعايير الموصى بها من قبل مجموعات المحكمين المختلفة باستخدام الطريقة نفسها.

وقارن كل من كوفلر (Koffler, 1980) وميلز (Mills , 1983) نتائج إعداد المعايير لطرائق الحكم على محتوى الفقرات والطرائق التي تستخدم بيانات المفحوصين. وعلى وجه الخصوص فحص كوفلر إعداد المعايير بطريقة نيدلسكي وأسلوب مجموعات المقارنة بينما استخدم ميلز اسلوب انغوف لمجموعات المقارنة وتقنياتمجموعة الخط الفاصل، ومع ذلك فقد كانت النتائج مختلطة (ممزوجة) نوعاً ما. فقد أظهرت دراسة ميلز أنه عند استخدام ثلاثة طرائق أو أكثر فمن الممكن الحصول على اتفاق بين طريقتين منهما على الأقل.

فعلى سبيل المثال كانت معايير طريقة انغوف وطريقة مجموعات المقارنة أكثر تطابقاً مما هي مع معايير طريقة مجموعة الخط الفاصل في معظم الحالات المدونة في الدراسات ولتوضيح التنوع في المعايير التي قد تنتج في الطرائق المختلفة، يبين جدول (18–2أ) بعض نتائج دراسة ميلز، إذ بدون الاختبارات معينة قائمة معايير النسبة الصحيحة، كذلك يبين جدول (18–2ب) نسبة المفحوصين الذين حققوا كل معيار.

وبينت دراسات أخرى تجريبية لإعداد المعايير أن المحكمين بخصائصهم المختلفة قد يصلوا إلى توصيات مختلفة تماماً. فعلى سبيل المثال دون جايجر ((Jager, 1982 فوارق مهمة في المعايير الناتجة عن مجموعات مؤلفة من معلمين ومواطنين عندما راجعوا فقرات اختبارات كفاءة المدرسة الثانوية شمال كاليفورنيا إضافة إلى أن ساوندرز وزملاؤه (Saunders, etal, 1981) وجدوا علاقة بين مستوى معرفة المحكمين للمحتوى ودرجة النجاح التي حصل عليها من المحكمين بطريقة نيدلسكي ومن مراجعة مثل هذه النتائج، ولاي دراسة إعداد معايير من الأنسب الحساب التجريبي لمدى مساهمة كل من الفروق بين الأحكام وطرائق إعداد المعايير النهائية الموصى بها. وقد اكتشف كل من برينان ولوك دور (Brenan & Lookwood, 1980) قابلية نظرية إمكانية التعميم للتطبيق في هذا السياق، كذلك فإن هذا المجال يحتاج للمزيد من البحث والعمل.

جدول (18–2أ) : أدنى معايير نجاح (من خلال النسبة الصحيحة) المحصلة من طرائق مختلفة من دراسة دونها ميلز 1983.

صيغة الاختبار	طريقة انغوف	مجموعة المقاونة	مجموعة الخط الفاصل
G	0.68	0.68	0.87
H	0.68	.75	0.88
I	0.62	.65	0.80
J	0.78	.62	0.90
K	0.68	.70	0.85
L	0.68	.60	0.83

جدول (18–2ب) : نسب المفحصوين الذين فشلوا في النجاح في المعايير الناتجة عن طرائق مختلفة.

صيغة الاختبار	طريقة انغوف	مجموعة المقاونة	مجموعة الخط الفاصل
G	7.06	6.30	29.75
H	6.03	8.19	22.09
I	7.60	9.42	26.66
J	9.10	4.07	23.71
K	7.82	9.32	25.40
L	7.95	5.38	23.11

أخيراً، فبالإضافة إلى احتمالية الحصول على نتائج مختلفة من قبل عينات محكمين مختلفة، وتقنيات مختلفة ، فقد أشار فان دير لیندن (Vander Linden, 1982) أن عدم توافق الأحكام الداخلية قد يكون مشكلة ويظهر هذا عندما يؤشر محكم احتمالية منخفضة لمفحوص يمتلك أقل كفاءة مقبولة للنجاح في اختبار فقرات سهلة، واحتمالية نجاح المفحوص نفسه في اختبار فقراته صعبة، فقد اقترح معامل تناقض يعتمد على نظرية الاستجابة للفقرة، وأسس تطبيقاتها مع الأحكام التي حصل عليها من طريقتي نيدلسكي وانغوف. وعند الأخذ بعين الاعتبار الحاجة إلى بذل جهود في هذا المجال، اقترح شيبرد (Shepard, 1980) أبحاث تطبيقية إضافية لإعداد المعايير يجب أن تسوّغ لبرامج اختبارات التأهيل غير المتميزة.

وللتعلم أكثر عن أثر التعليمات المختلفة المقدمة للمحكمين . ومزايا وعيوب عمل المحكمين المستقل أوضحت مجموعات، وكذلك الوقت الأكثر مناسبة والصيغ التي يفضل بوساطتها عرض البيانات المعيارية على المحكمين.

اعتبارات عملية في إعداد المعايير:

بغض النظر عن الطريقة المستخدمة فإنه لا يمكن الاستغناء عن الحاجة إلى أحكام في تأسيس معايير الأداء الاختباري.

ومن المهم تمييز أن إعداد المعايير عبارة عن مشكلة سيكومترية مهمة. فهي ليست قضية تقنية فحسب، بل أن توابع ملائمة المعايير أو عدمه للأفراد والمعاهدة والمجتمع يجب اخذها بعين الاعتبار. وللبعض مثل جلاس **(Glass, 1972)** فإن النتائج. النظرية والأسس التجريبية للمعرفة التي تخضع لها أساليب إعداد المعايير إن كانت غير ملائمة فإن التطبيق نفسه يبدو غير قابل للتعديل ولاخرين (مثل Hambelton, Hambelton 1978, 1980 popham, 1978, Scriven, 1978, & Shepard, 1976, 1979).

فإن الحاجة إلى المعايير موجه لاتخاذ قرارات في سياقات تربوية وتحليليه، وهذا واقع لا يمكن الغاؤه او إهماله حتى يتم حل المشكلات الفلسفية والمنهجية المتعلقة بالتطبيقات العملية. وباعطاء هذه الحالة من المضاهاة وصولاً إلى مشكلة اعداد المعايير، فإن الخطوة الأولى هي الاستفسار عما إذا كانت هنالك حاجة ملحة ومنطقية لتأسيس معايير أداء لتفسير درجات الاختبار لا زالت قيد المساءلة . فعلى سبيل المثال ففي مجال مثل الدراسات الاجتماعية في المدارس الابتدائية فليس من الضروري أن يسيطر الطلبة سيطرة تامة على المادة في وحدة ما قبل تدريس وحدة أخرى. ومع أن تقييم تحصيل الطلبة يكون مناسباً إلا أن الحاجة إلى تحديد درجات قطع كمؤشر لمستويات السيطرة لمثل هذا الاختبار هو موضوع مطروح للمساءلة . في حين أن الحاجة واضحة لمعايير أداء لتفسير درجات اختبار مستخدم في اتخاذ قرار. وينصح هنا بمراجعة الأساليب الشائعة لإعداد المعايير وتحديد تلك التي يبدو أنه يمكن الدفاع عنها في الموقف المعطى.

بعد ذلك، فمن المهم تحديد المهددات المشابهة لعدم صدق الاستنتاجات التي أخذت من درجات الاختبار. وقد حدد جايجر **(Jaeger, 1979)** عدداً من مهددات صدق الاستنتاجات المأخوذة في درجات الاختبار إلى درجات النطاق. وربط المشكلات المختلفة بطريقة معينة في إعداد المعايير. وتلك التي تطبق على طرائق التحكيم جميعها والتي تعتمد على محتوى الفقرة تكون متميزة، وذلك لأسباب منها لتحديد غير المناسب للنطاق، والخطأ العشوائي المؤثر على قرارات المحكمين المختلفة، والطرائق غير المناسبة أو غير الملائمة لمعاينة الفقرات للتأكد من تمثيلها للنطاق وتحيز الأحكام في مراجعة الفقرات منفردة . لاحظ أن

المهددات الأولى والثالثة والرابعة ظهرت بالفعل في طرائق بناء الاختبار الخاطئة أو على الأقل غير ملائمة لتؤكد الدرجة اللازمة في صدق المحتوى لأنواع الاستنتاجات المرغوب بها بوساطة درجات الاختبار) ومن المحتمل أن التهديد الثاني والخامس التي يمكن ضبطها أو خفضها، أثناء إعداد المعايير، بافتراض أن المعايير تأسست لدرجات طوّر اختبارها مسبقاً. لذا يبدو أن استخدام عدد كبير من المحكمين (ويفضل أكثر من 6 غلى 8 الذين استخدموا في دراسات منشورة سابقاً) واختيارهم عشوائياً من مجموعة محددة بعناية من المحكمين المؤهلين وتزويدهم بمعلومات واضحة عن السياق التي ستستخدم فيه درجات الاختبار، وتدريبهم على أداء المهمة لتقليل المشكلات المحتملة. وقد زودنا كل من زيكي وليفنجستون (Zieky & Livingston, 1977) وهاملتون (Hambelton, 1978) بإرشادات إضافية ومقترحات تحقق طرائق عدة في إعداد المعايير التي نوقشت هنا.

إضافة إلى ذلك ، ينصح باستخدام اثنين أو أكثر من أساليب إعداد المعايير وعينات محكمين عديدة تختار عند الضرورة لتمثيل وحدات محكمين مختلفة وملائمة. وفيما لو فشلت الطرائق المتنوعة في إنتاج معيار مفرد موصم (كما يجب أن تظهر) فإن السؤال عن تلك التي يجب اختيارها أو كيفية المواءمة لتكوين حل مركب مناسب فهذا يتطلب في النهاية كلماً قيمياً.

وينصح أيضاً فحص الأدلة التجريبية حول كيفية اداء المفحوصين النموذجي في الاختبار واستخدام هذه المعلومات في تقويم نتائج وضح معايير معينة. افترضعلى سيبل المثال حالة تأسس الاداء فيها باحدى طرائق تحكيم الفقرات عن درجة قطع 87% عن اختبار تعلم للاتقان في الرياضيات لأطفال مرحلة الطفولة المتاخرة. ولكن نتائج التجريب الميداني بينت أن نسبة بسيطة من الطلبة قد حققوا هذا المحكم بعد التعليم. أن نتائج وضع هذا المعيار يجب أن تؤخذ بعين الاعتبار من خلال عدة وجهات نظر. الأول، أنه يجب أن يؤخذ بعين الاعتبار توابع التعلم المتتابع، فإن تقدم الأطفال دون أن يحققوا هذا المعيار فماذا سيكون تأثير هذا على التعلم المستقبلي ؟ بالإضافة إلى أن أثر الدافعية من مواجهة إعادة التعليم يجب أن تؤخذ بعين الاعتبار. والفرض الضائعة (من خلال تقليل الوقت لتعليم الموضوعات الأخرى) قد يكون عاملاً مشابهاً في مجالات مثل شهادة التخصص، فإن نتائج (عواقب) تأهيل المتقدمين الذين هم أقل من أدنى كفاءة مطلوبة يجب أن توازن مقابل خسارة المجتمع للخدمات التي يقدمها أصحاب الشهادات المؤهلين تأهيلاً حقيقياً، والذين فشلوا في تحقيق معيار متشدد. ومن الواضح أن تأسيس معايير أداء مفيدة تتطلب موازنة جيدة لتقنية ملائمة ورؤية واضحة لأهداف وغايات البرنامج الذي يتطلب قرارات عن المفحوصين تعتمد على أدائهم في الاختبار. إحدى الأساليب تسمح لمعد المعايير أن يأخذ هذه التوابع بانتظام تتطلب تطبيق نظرية القرار، والتي ستناقش في الأجزاء الآتية.

اعتبارات تقنية في إعداد المعايير:

يركز الأدب التقني في إعداد المعايير على إحدى القضايا الأساسية، هي إعداد درجة قطع على التدريج الملاحظ تسمح لمستخدم الاختبار تكوين خلاصات عن حالة السيطرة الحقيقية للمفحوصين والتي تم تحديدها على متصل درجات النطاق. وكما سنرى، فإن المعنى الملائم للخلاصات المناسبة يعتمد على الطريقة المستخدمة في إعداد درجة القطع على التدريج الملاحظ . وفي بعض الطرائق تحدد الملائمة من خلال احتمالية التصنيف الخاطئ. وستصف في هذا الجزء أساليب عدة لهذه المشكلة، وكل أسلوب موصوف هنا هو تطبيق لنظرية القرار والتي تم عرضها في الفصل الثاني عشر. كذلك فإن الأدب التقني طرح قضية تقنية أخرى ثانوية نوعاً ما، ويعتمد تطوير معامل لتكميم نوعية القرارات المأخوذة على درجة القطع على التدريج الملاحظ. والأبحاث التي تعالج هذه المشكلة تجدها في (Wilcox, 1977) (Mellen berg, 1977) و (Hunyh, 1980) يمكن للقارئ المهتم بالموضوع الرجوع إليها.

إعداد درجة قطع على التدريج الملاحظ:

تقليل احتمالية التصنيف الخاطئ: تقسم درجة قطع تدريج النطاق درجة النطاق إلى منطقتين: مفحوصين عند درجة القطع أو أعلى منها وهؤلاء مسيطرين حقيقيين، ومفحوصين درجاتهم أدنى من درجة القطع وهؤلاء غير مسيطرين حقيقيين وتفسر درجة النطاق للمفحوص على أنها نسبة الفقرات من النطاق التي يستطيع المفحوص أن يجيب عنها إجابة صحيحة، كذلك فإن مدى تدريج درجة النطاق يمتد من صفر إلى 1.

وسنرمز إلى درجات النطاق بالرمز t، ودرجة القطع على تدريج النطاق وكما لاحظنا من البداية أن $t0$ قد تستخدم في أي من طرائق إعداد المعايير الموصوفة في الأجزاء السابقة. وبما أن تدريج درجة النطاق هو تدريج لدرجات النسبة الصحيحة وطرائق إعداد المعايير ينتج عنها درجات إما على شكل نسبة مئوية أو تدريج عدد الإجابات الصحيحة فمن الضروري تحويل درجات القطع هذه قبل تفسيرها على أنها $t0$. وفيما لو تم التعبير عن درجة القطع على متصل النسبة الصحيحة فإنه يجب تقسيمها على 100 للحصول على $t0$: وفيما لو تم التعبير عنها على تدريج العدد الصحيح فيجب تقسيمها على عدد الفقرات المستخدمة في إعداد درجة القطع من أجل الحصول على $t0$.

وحالما تتحدد $t0$، فإن المشكلة تكمن في تحديد درجة القطع على تدريج الدرجات الملاحظة والتي تسمح لمستخدم الاختبار عمل استنتاجات مناسبة حول حالات السيطرة للمفحوصين. وعادة يرمز لدرجة القطع الجديدة بالرمز X_0 . ويعبر عنها على تدريج العدد