الفصل الثاني

الاطار النظري

- الاختبارات معيارية المرجع
- الاختبارات محكية المرجع
- مبادئ في الاختبارات محكية المرجع
- استخدامات الاختبارات محكية المرجع
- أسس الإرجاع إلى المحك
- تفسير الدرجات معيارية المرجع ومحكية المرجع
- الفروق بين الاختبارات معيارية المرجع والاختبارات محكية المرجع
- بناء الاختبارات محكية المرجع
- صدق الاختبارات محكية المرجع
- ثبات الاختبارات محكية المرجع
- الطرق المتبعة في حساب معامل ثبات الاختبارات محكية المرجع
- العوامل المؤثرة في حساب معامل ثبات الاختبارات محكية المرجع
- الخصائص الاحصائية لمفردات الاختبارات محكية المرجع

الفصل الثاني

الإطار النظري

تتكامل المنظومة التعليمية عندما يتم تقويم أداء الطالب لمعرفة مدى تحقيقه للأهداف المرجوة من العملية التعليمية وتعد " الوظيفة الأساسية للاختبارات التحصيلية التى يجريها المعلم فى حجرات الدراسة هي قياس التحصيل الدراسي للمتعلم وبهذا يستمر فى تقويم التقدم التربوي " (نادية بعيبع ، ١٩٩٧ ، ص ١٧٣) .

تعتمد الاختبارات المستخدمة في مؤسساتنا التعليمية في الوقت الحالي على مقارنة الدرجة التي يحصل عليها الطالب في الاختبار بدرجة محددة يتم تعيينها من قبل الجهات المسؤولة عن ذلك قبل تطبيق الاختبار ، تمثل هذه الدرجة أساس الحكم على قدرة الطالب على النجاح والاجتياز وهذه هي الفلسفة التي يقوم عليها القياس محكي المرجع أي المقارنة بمعيار مطلق ، ولا تتم أي محاولة لمقارنة درجة الطالب بدرجات زملائه إلا في حالات تستدعي ذلك مثل القبول في الكليات .

إلا أنه لا يكفي استخدام فكرة القياس محكي المرجع من جانب واحد مع إهمال باقي الجوانب، فلا يجوز أن نركز عند اختيار مفردات الاختبار على التباين الذي ستنتجه هذه المفردات في درجات الطلاب ، لأن الهدف من هذه الاختبارات هو التأكد من وصول الطالب إلى مستوى التمكن المطلوب وليس مقارنته بغيره من الطلاب وبهذا لا يمكن اتباع الخطوات المستخدمة في بناء الاختبارات معيارية المرجع عند تصميم اختبارات محكية المرجع ، كما لا يجوز أيضاً استخدام معاملات الارتباط المعتمدة على التباين لحساب ثبات درجات هذه النوعية من الاختبارات وإنما يجب الاعتماد على الطرق المناسبة لها .

هذا وتشير الباحثة إلى أنه إذا كنا مدركين لأهمية استخدام القياس محكي المرجع فعلينا تطبيقه من كافة الزوايا ، وسيتم فيما يلي التعرف على التغيرات التي طرأت على مفهوم الاختبارات والفلسفة المرتبطة بنوعي القياس والتي أدت إلى تغير في التقنيات المستخدمة معهما ابتداء من طريقة البناء ووصولاً إلى طرق حساب معاملات الثبات والعوامل المؤثرة فيها والتي هي موضوع هذه الدراسة .

الاختبارات معيارية المرجع :

هذه الاختبارات هى الأولى فى الظهور والأشهر فى الاستخدام وقد نشأت فكرتها مرتبطة بالفلسفة التربوية التى كانت سائدة فى الولايات المتحدة الأمريكية أوائل القرن العشرين عندما دعت الحاجة إلى تصنيف الضباط والجنود قبل اشتعال الحرب العالمية الأولى اعتماداً على الفروق بينهم عندما تم تطبيق أول اختبار جمعي عليهم وهو اختبار أرثر أوبتس (صلاح الدين محمود علام، ٢٠٠١، ٢٠-٢١) .

تقوم فكرة هذه الاختبارات على مقارنة أداء الفرد بأداء أقرانه وتسمى الجماعة التي يقارن على أساسها الفرد بالجماعة المعيارية ، وتشير تعاريف الاختبارات معيارية المرجع إلى أنها اختبارات أعدت لتحديد الوضع النسبي للفرد وذلك من خلال مقارنة أدائه بأداء الجماعة المعيارية التي ينتمي إليها في المجال الذي يقيسه الاختبار .

حيث يعرفها شحته عبد المولى (١٩٩٩ ، ص ٩) بأنها " الاختبارات التي تستخدم لتقدير أداء الفرد بالنسبة لأداء الأفراد الآخرين في القدرة التي يقيسها الاختبار ومقارنة أداء الفرد بمتوسط (أو نقطة توسط) أداء جماعته المرجعية أو المعيارية " .

أي أنها تحدد موقع الطالب مقارنة بباقي الطلاب وبالتالي تساعدنا على تحديد " موقع " أو " رتبة " الطالب وذلك بمقارنة أدائه بمعدل أداء الآخرين (Kubiszyn & Borich , 2000 , p. 33) ، وبالتالي فهي ببساطة اختبارات تقارن أداء الفرد بمتوسط المجموعة أو المعيار (سعاد حسانين، ٢٠٠٠ ، ص ٩ ؛ Wilson , 1998 , p. 256) ، وهكذا يزود المختبرين بمعايير لتحديد معنى درجة الفرد على اعتبار المعيار مستوى أداء نموذجي لمجموعة محددة ، وبمقارنة الدرجة الخام للفرد بالمعيار نستطيع أن نحدد فيما إذا كانت الدرجة أعلى أو أقل أو حول معدل الجماعة (Woolfolk , 1993 , pp. 506-507) .

هذا وتعرف المجموعة المعيارية للطالب بأنها " طلاب صفه أو من هم في المستوى الأكاديمي نفسه كأن تفسر علامة طالب بأنه أعلى تحصيلاً من ٨٠% من طلاب صفه " (سمير المسلمي ، ١٩٩٦ ، ص ٣١) .

أي تقارن هذه الاختبارات أداء الطالب بأداء عينة ممثلة من الطلاب في نفس العمر والمستوى لتشير فيما إذا كان أداؤه أعلى أو دون المتوسط المتوقع لعمره أو مستواه (Bigge , Stump , & Spagan , 1999 , p. 181) .

ويتم التوكيد في هذا النوع من الاختبارات على الدرجة الكلية للفرد نسبياً حيث تفسر هذه الدرجة بمقارنتها بدرجات عينة ممثلة من الأفراد ، وعلى الباحث أن يقرر ما تعنيه ممثلة فعادة تشير إلى أفراد متشابهين في العمر والجنس (Peers , 1996 , p. 27) .

يتضح من التعاريف السابقة للاختبارات معيارية المرجع أنها تزوّد بمعلومات عامة غير قابلة لتطوير فعاليات التعليم اليومية مباشرة ، كما أنها لا تناسب أهداف التقويم لأن المقارنة بمعيار مسبق الوضع أكثر أهمية من المقارنة بأداء الآخرين ، فليس من المريح لولي الأمر أن يعلم أن ابنه هو أفضل من معظم الطلاب في الصف في القراءة إذا كان كل الطلاب غير قادرين على قراءة مادة مناسبة لمستوى صفهم (Woolfolk , 1993 , p. 508) .

وبحكم الهدف من هذه الاختبارات وهو مقارنة أداء الطالب بأداء غيره من الطلاب فهي تميل لأن تعزّز وتقوّي الاعتقاد بأن مشكلة إختلاف الطالب عن المعيار تكمن في الطالب في حين قد تكون المشكلة الحقيقية في طريقة التدريس أو المنهج (Witt , Elliott , Kramer , & Gresham , 1994 , pp. 28-29) .

كما تشجع هذه الاختبارات على المنافسة حيث يتنافس بعض الطلاب ليكونوا الأفضل في حين يدرك آخرون أن كونهم الأفضل أمر مستحيل لذلك قد يتنافسون ليكونوا الأسوأ ولكلا الهدفين كوارثه (Woolfolk , 1993 , p. 507) .

ومن هنا بدأت تظهر نقط ضعف هذه الاختباراتْ فإزدادت الأصوات تنادي بنظم أقل تعسفاً وأكثر مناسبة في قياس وتقويم تحصيل الطلاب (نادية عبد السلام ، ١٩٩٦ ، ص ١٧٧) .

وتعقب الباحثة على ما سبق بأن الفلسفة التي تقوم عليها الاختبارات معيارية المرجع وهي اعتبار الاختبارات كأدوات لتصنيف المتعلمين وفقاً لاستعداداتهم العقلية ومستويات تحصيلهم من خلال ترتيبهم وتحديد الوضع النسبي لكل منهم بالنسبة للآخرين لا تتم حالياً في مدارسنا أبداً لأن المقارنة تتم مع الدرجة المحددة من قبل وزارة التعليم وليس مع متوسط درجات الطلاب في مجموعة المختبرين .

ومع ذلك يتم اختيار مفردات هذه الاختبارات بحيث تحدث درجة معقولة من التباين في الاستجابة لذلك تتعرض العناصر الاختبارية التي لا تسهم في إيجاد هذا التباين للحذف ، وهكذا تضيع الدلالة المنهجية التعليمية للاختبار نتيجة لعمليات التحسين (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ص ٢٠٢-٢٠٣) .

الاختبارات محكية المرجع :

ظهرت فكرة الاختبارات محكية المرجع في الولايات المتحدة الأمريكية لمواجهة نواحي القصور التي تعاني منها الاختبارات معيارية المرجع ، وكانت البداية حين قدم روبرت جليزر (١٩٦٣) هذا المفهوم لأول مرة إلى المجتمع التربوي من خلال مقالة وضحت معالمه كونه مصمماً على أساس تحديد المهارات التعليمية المرغوبة ، والتحديد الاجرائي للأهداف التعليمية ، وتحديد مستويات الأداء المقبولة ، وتبني استراتيجيات تدريسية تؤدي إلى وصول المتعلمين إلى هذه المستويات وتحدد ما أتقنه المتعلم وتشخيص مواطن القوة والضعف والإفادة من ذلك في توجيه عملية التعلم (مصطفى محمد كامل ، ١٩٩٩ ، ص ٧) .

ولقد حدد جليزر (١٩٦٣) الفرق الأساسي بين الاختبارات معيارية المرجع المعدة لتحديد الوضع النسبي للطالب (بين زملائه الطلبة) والاختبارات محكية المرجع التي تركز على معيار مسبق التحديد حين قال " ماسوف أطلق عليه قياسات محكية المرجع تعتمد على معيار مطلق للصفة بينما ما أطلق عليه قياسات جماعية المرجع تعتمد على معيار نسبي " (نادية عبد السلام ، ١٩٩٦ ، ص ١٥٦) .

كما طرح في هذا المقال عدداً من الأسئلة تتعلق بصعوبة إعداد عناصر اختبارية تصلح لنوعي القياس في آن واحد ، وأكد على أهمية الوصف الجيد للأنماط السلوكية لدى المتعلم حين قرر أن درجة الطالب توضح مايستطيع عمله وذلك من خلال مقارنة أدائه بمعيار مطلق مسبق التحديد وهذا ما دُعي بدرجة القطع (جابرعبد الحميد جابر ، ١٩٩٦ ، ص ص ٢٣٢-٢٣٣) .

وهكذا فإن فكرة جليزر تعتبر الأولى من بين قائمة الأفكار التي كان لها أثر دائم في التفكير وقضايا القياس التربوي حيث شدَّد في وصف ما يستطيع الفرد عمله وفي تفسير نتائج الاختبار في حدود موضع الفرد على محك الأداء (Linn , 1994 , p. 12) .

ازداد بعد ذلك الاهتمام بهذا المفهوم وتم التركيز على مواصفات الاختبار وكتابة أهداف واضحة ، كما أصبح المعلمون أكثر حساسية لوصف إطارات المعارف والمهارات الضرورية لكتابة مفردات اختبار صادقة تستطيع توجيه التعليم وتسهل تفسير درجات الاختبار (Glaser , 1994 , p. 10) .

تعريف الاختبارات محكية المرجع :

يذكر بابام Popham أن بعض المربين قد حاولوا تعريف الاختبارات محكية المرجع على أنها أي اختبارات غير معيارية المرجع (نادية عبد السلام ، ١٩٨٦ ، ص ٥٦٢) ، في حين يشير فؤاد أبو حطب وسيد أحمد عثمان وآمال صادق (١٩٩٧ ، ص ص ٤١٣-٤١٤) إلى أننا لا نستطيع اعتبار الاختبار محكي المرجع لمجرد أنه ليس معياري المرجع وذلك لأن عملية الإرجاع إلى المحك تعني أن الاختبار مصمم ومنشأ على أساس أن الأداء متصل بمراجع السلوك ، أي أن الاختبار بحكم طريقة تصميمه يجب أن يمدنا بمعلومات عن قدرة الطالب على القيام بأداءات معينة بصورة مطلقة .

ومن خلال التعاريف الكثيرة التي وردت في هذا المجال رأت الباحثة تصنيف تعاريف الاختبارات محكية المرجع في محورين أساسيين وهما :

- الاختبارات التي تقارن أداء الفرد بمستوى أداء مطلق .
- الاختبارات التي تقيس أداء الفرد في ضوء أهداف سلوكية .

أولاً : الاختبارات التي تقارن أداء الفرد بمستوى أداء مطلق :

عرّف جليزر ونتكو (١٩٧١) الاختبار محكي المرجع على أنه الاختبار الذي يبنى على قصد لإعطاء قياسات قادرة على التفسير المباشر في ضوء معايير محددة (Crehan , 1974 , p. 255).

وتعرّفها نادية عبْد السلام (١٩٨١ ، ص ١٧٨) بأنها اختبارات " تستخدم لتحديد مستوى الفرد بالنسبة لمحك ثابت ، بمعنى أداء مقنن محدد " ، ويتفق مع هذا التعريف ما ورد عن (Bigge et al. , 1999 , p. 181 ; Woolfolk , 1993 , p. 507) .

أي تستخدم هذه الاختبارات " لتقدير أداء الفرد بالنسبة إلى ميزان أو مستوى مطلق للأداء بدون الحاجة إلى موازنة أدائه بأداء الأفراد الآخرين ، ويتطلب هذا تحديد مستوى مسبقاً للأداء وهو يمدنا بمعلومات محددة ومفصلة عن تحصيل الطلاب في موضوع دراسي معين ويستخدم عادة لوصف تقدمهم " (سعاد حسانين ، ٢٠٠٠ ، ص ٩) .

وبالتالي فهي أدوات بنيت خصيصاً لتقويم مستوى أداء الفرد في علاقته ببعض المحكات (Wilson , 1998 , p. 253) .

ويشير كبيسزين وبوريش (Kubiszyn & Borich , 2000 , p. 33) إلى أنها نوع من الاختبارات تخبرنا عن إتقان الطالب لمجموعة من المهارات وذلك بمقارنة أدائه بمحك أو معيار مطلق ، مثل هذه المعلومات تساعدنا على أن نقرر فيما إذا كان الطالب يحتاج

عوناً أقل أو أكثر في مجموعة من المهارات بغض النظرعن تحديد مركزه أو رتبته مقارنة مع باقي الطلاب .

هذا ويمكن تعريف المحك على أنه : " السلوك الذي يحدد كل نقطة على امتداد متصل التحصيل " (نادية عبد السلام ، ١٩٩٦ ، ص ١٦٠) ، أو أنه أساس مستقل للحكم على الأداء في الاختبار (فؤاد أبو حطب وآخرون ، ١٩٩٧ ، ص ١٩٣) .

ويمكن تعريفه على أنه مستوى الأداء المتوقع وهذا بالطبع مهمة صعبة ولكنها هامة للمعلم وذلك لتعيين مستوى يعتقد أنه دال على السلوك المقبول (Jacobsen , Eggen , Kauchak , & Dulaney , 1985 , p. 69) .

وعلى هذا فإن " المحك الذي ينسب إليه الاختبار أو يرجع إليه قد يكون هدفاً تعليمياً أو مستوى كفاءة المتعلم أو نتيجة مرغوبة التحقق بعد التعلم " (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ٢٣٥ ؛ مجدي حبيب ، ١٩٩٦ ، ص ٣٦٠) .

وبالتالي فإنه على عكس التقويم معياري المرجع يحاول التقويم محكي المرجع أن يحدد أي الأفراد قد وصلوا إلى محك الأداء (Witt et al. , 1994 , p. 29) .

هذا ويمكن للباحثة القول أن الاختبارات المستخدمة في مدارسنا الآن تتطابق مع ما أجمعت عليه التعاريف السابقة حيث تقوم على مقارنة أداء الطالب بمعيار مسبق التحديد يمثل الحد الأدنى للأداء المتوقع لاعتبار الطالب قد وصل إلى حد التمكن وهو الدرجة المحددة من قبل وزارة التعليم دون أية محاولة إلى موازنة أدائه بأداء الأفراد الآخرين .

ثانياً : الاختبارات التي تقيس أداء الفرد في ضوء أهداف سلوكية :

يعرّف بابام Popham هذه الاختبارات بأنها اختبارات تستخدم للتأكد من مركز الفرد بالنسبة إلى مجال سلوكي محدد تحديداً جيداً ، ويتكون مجال السلوك من مجموعة من المهارات التي يؤديها المختبرون في موقف الاختبار (نادية عبد السلام ، ١٩٨٦ ، ص ٥٦٣) ، ويذكر هامبلتون (Hambelton , 1988 , pp. 277-278) أن هناك عدة نقاط تستحق التعليق حول تعريف بابام Popham وهي :

١- تستخدم المصطلحات أهداف ، كفاءات ، ومهارات بمعنى واحد في هذا المجال .

٢- يجب تعريف الأهداف المقاسة بشكل جيد مما يجعل عملية كتابة المفردة أسهل وأكثر صدقاً ويحسن تفسيرات درجة الاختبار حيث يتحسن وصف تفسيرات الدرجة بسبب وضوح المجالات السلوكية أو المحتوى الذي ترجع إليه درجات الاختبار .

٣- عندما تقاس أكثر من كفاءة في اختبار واحد يجب تسجيل أداء المختبر في كل كفاءة .

٤- لا يتضمن التعريف تحديد درجة قطع ، ومن الشائع وضع حد أدنى كمعيار أداء لكل كفاءة تقاس في اختبار محكي المرجع لتفسير أداء الفرد بالنسبة لها .

يضاف إلى ذلك اختلاف عدد الأهداف المقاسة من اختبار لآخر واختلاف عدد مفردات الاختبار التي تقيس كل هدف .

وهذا يتفق مع تعريف إسماعيل الوليلي (١٩٩٦ ، ص ١٥) في أن الاختبار محكي المرجع هو " الاختبار الذي يستخدم في تقدير أداء الفرد بالنسبة إلى مجال من السلوك محدَّد تحديداً جيداً أي أن عبارة " التحديد الجيد للمجال السلوكي " هي الركيزة الأساسية لهذا المفهوم " .

ويشير شحته عبد المولى (١٩٩٩ ، ص ١٠) إلى أن هذه الاختبارات " تبنى على أساس أهداف محددة وموصفة وصفاً دقيقاً والتي تحدد حالة (مستوى) الفرد كمتقن أو غير متقن لهذه الأهداف والخاصة بمحتوى محدَّد من المعارف والمهارات وذلك على أساس درجة القطع أو مستوى قدرة محدَّد مسبقاً " .

في حين يعرفها مصطفى محمد كامل (١٩٩٩ ، ص ١١) على أنها الاختبارات التي تستخدم في البرامج التعليمية القائمة على الأهداف للتحقق من مدى إتقان الطلاب للمخرجات المعبرة عن الأهداف التعليمية بالنسبة لمجال سلوكي محدَّد جيداً ، ويتكون هذا المجال من مجموعة من المهارات أو الميول أو الاستعدادات التي يؤديها المختبرون في مواقف الاختبار وذلك في ضوء مستوى إتقان يشكل الحد الأدنى الذي ينبغي على الفرد الوصول إليه لكي يعتبر متقناً .

أي أن هذه الاختبارات " تصف أداء الطالب بدلالة أنماط محددة من المهارات أو المهام التعليمية التي يتضمنها الاختبار " (صلاح الدين أبو ناهية ، ١٩٩٤ ، ص ٥٢) .

ويشير بيندر (Bender , 1998 , p. 145) إلى أنه بسبب الحاجة إلى معلومات أكثر كمالية عن أداء الطلاب طورت اختبارات تقارن أداء الطالب بمجموعة أهداف سلوكية أكثر من المقارنة بأداء باقي الطلبة .

وهكذا " تبنى الاختبارات محكية المرجع لغرض تفسير أداء الفرد في ضوء مجموعة من الكفايات الأدائية (Cometencies) " (إبراهيم مبارك الدوسري ، ٢٠٠٠ ، ص ١٢٥)

ويمكن إيجاز ما سبق في أن الاختبار محكي المرجع يستخدم لتحديد مكانة الفرد بالنسبة لمجال سلوكي حسن تعريفه (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ٢٣٣) .

يلاحظ من عرض التعاريف السابقة للاختبارات محكية المرجع أنها اتفقت على الاهتمام بمدى إتقان الطالب للأهداف التعليمية التي تغطي مجال سلوكي معين ، مع عدم التركيز على إرجاع درجة الطالب لدرجة قطع كما في الفئة الأولى من التعاريف .

مبادئ في الاختبارات محكية المرجع :

يتطلب الاختبار محكي المرجع :

١- تحديد المجال المفروض تغطيته بالاختبار بشكل واضح ويعد هذا أمر سهل في مجالات المهارة الأساسية كالحساب وأكثر صعوبة في مجالات أخرى كالدراسات الاجتماعية .

٢- تعريف مجموعة من الأهداف التعليمية تصف بشكل دقيق الاستجابة التي يتوقع أن يقدمها الطالب في نهاية التعلم .

٣- تحديد معايير أداء ملائمة للطالب هو مفتاح الاستخدام الفعَّال للاختبار محكي المرجع في التعليم الصفي .

٤- اختيار مفردات الاختبار على أساس كيفية عكسها للسلوك المحدد في الأهداف التعليمية لقياس مخرجات التعلم المتوقعة .

٦- تقديم تقرير يصف أداء الطالب وذلك لتحديد ما يستطيع أن يؤديه الطالب ومستوى الأداء الذي يستطيع أن يصل إليه بوضوح (Gronlund , 1973 , pp. 3-6) .

استخدامات الاختبارات محكية المرجع :

١- تتجنب المقارنات بين الأفراد وتفسَّر درجاتها في ضوء سلوك محدد وبهذا فهو يسهل التعليم المطوَّع ، كما تزود المدرسين بمعلومات تتعلق بالتحسن الأكاديمي لتلاميذهم مما يساعدهم على استخلاص خطط تعليمية فعالة (نادية عبد السلام ، ١٩٩٦ ، ص ١٥٧) .

٢- تناسب المواقف التعليمية التي تتطلب قياس أداء الطالب لمجموعة من المعارف والمهارات المرتبطة بمحتوى مادة دراسية معينة (إسماعيل الوليلي ، ١٩٩٦ ، ص ١٢ ؛ Peers , 1996 , p. 27) .

٣- لا تحدِّد الأعمال التي أتقنت فقط لكنها تحدِّد أيضاً أنواع الأعمال التي تحتاج والتي لا تحتاج لانتباه تربوي مما يساعد على تقويم الفعالية التعليمية وتحسين التكنيك التعليمي (نادية عبد السلام ، ١٩٩٦ ، ص ص ١٥٨ – ١٥٩) .

٤- تعيين المختبرين في واحد من مستويات الإتقان لكل واحد من الأهداف المغطاة بالمفردات في الاختبار (Swaminathan , Hambelton , & Algina , 1974 , p. 263) .

٥- تساعد على تحديد استعداد الشخص للانتقال إلى مستوى أعلى ، ولا تتم أية محاولة لتحديد كم أدى الطالب أفضل أو أقل من المحك ولكن يقيَّم بطريقة النجاح / الرسوب فقط (Witt et al. , 1994 , p. 29) .

٦- تصف السلوك وصفاً دقيقاً بما يمكِّن من تشخيص جوانب القوة والضعف واقتراح أساليب العلاج (صلاح علام ، ٢٠٠٠ ، ص ٣١٦) .

٧- تستخدم في التقويم التكويني الذي يعنى بتقييم الطالب بشكل نظامي حيث تمكن من ملاحظة تقدم الطالب وتحديد فيما إذا كان التعليم فعالاً ، وتساعد على تخطيط المهارة التالية التي ستدرس ، وبما أن التركيز على المهارات بدلاً من المقارنة مع الآخرين ، فمعرفة ما يدرس وكيف يقاس يصبح أوضح (Witt et al. , 1994 , p. 30) .

أسس الإرجاع إلى المحك :

يحدد فؤاد أبو حطب وآخرون (١٩٩٧ ، ص ص ٤١٦- ٤١٧) أربع خطوات يتم من خلالها وصف درجات الاختبار وذلك كما يلي :

١- وضع مجموعة من الأهداف تمثل النواتج التعليمية المرغوب فيها عند نهاية التدريس .

٢- اعداد مفردات الاختبار واختبارها لقياس الأهداف قياساً دقيقاً .

٣- التحديد المسبق لمستويات الأداء المقبولة .

٤- تطبيق الاختبار وتقويم أداء المختبرين في ضوء عدد الأهداف التي يستطيعون تحقيقها بنجاح .

هذا ويشير فريدنبرج (Friedenberg , 1995 , pp. 86-87) إلى أن عملية الترجيع إلى المحك تبدأ بتعيين المحك ، تطبيق الاختبار ، ثم تنسب الدرجات إلى المحك .

تفسير الدرجات معيارية المرجع ومحكية المرجع :

درجة الاختبار في حد ذاتها ليس لها معنى أو مدلول وحتى تكون مفيدة يجب أن يعطى لها تفسير (فؤاد أبو حطب وآخرون ، ١٩٩٧ ، ص ١٨) ، فدرجة أداء الطالب تكتسب معنى فقط عندما تفسر مع بيانات أخرى كدرجات أداء الطلاب الآخرين في نفس الاختبار أو عندما تقارن بمعايير أداء محددة (Peers , 1996 , p. 27) .

وعندما نفرق بين القياسين معياري ومحكي المرجع فإننا نميز بين طريقتين مختلفتين في تفسير الدرجات ، وبالرغم من أن لكل نوع منهما خصائصه المحددة إلا أن السؤال يبقى هل الاختبار معياري أو محكي المرجع حتى يتم تفسير درجاته أي أن نوع القياس يعتمد على طريقة تفسير الدرجات (Wiersma & Jurs , 1985 , p. 12) .

ولكن ألا نستطيع عمل تفسيرات معيارية المرجع من اختبارات محكية المرجع وبالمثل هل نستطيع عمل تفسيرات محكية المرجع من اختبارات معيارية المرجع ؟

تقدِّم نادية عبد السلام (١٩٨٦ ، ص ص ٥٥٧ –٥٥٨) إجابة صريحة لهذا السؤال حيث تشيرإلى أنه يمكن عمل تفسيرات معيارية المرجع من اختبارات محكية المرجع والعكس بالعكس ، إلا أن كلا من التفسيرين سيكون ضعيفاً لأن تفسيرات معيارية المرجع لن تتضح من اختبارات محكية المرجع لم تصمم لقياس الفروق الفردية ، وبالمثل لن تتضح تفسيرات محكية المرجع من اختبارات معيارية المرجع لم تبنَ على أهداف محددة بوضوح. وهكذا لن تقدم هذه التفسيرات الغاية المرجوة منها كما لو تمت من اختبارات بنيت خصيصاً لتسهيل النوع المرغوب فيه من التفسيرات .

وهذا ما تحاول الدراسة الحالية التنويه عنه وهو أن مشكلة التقويم المتبع الآن أنه يستخدم اختبارات معيارية المرجع في كافة المعايير ثم يفسر نتائج هذه الاختبارات بالعودة إلى مستوى مطلق كما يتم في الاختبارات محكية المرجع .

ويشير رشدي منصور (١٩٨٧ ، ص ٢٣) إلى أنه لا يجب الاقتصار في تفسير الدرجات على المعيار السيكومتري والاديومتري فحسب ، بل يجب أن يتعداها إلى زاوية رؤية أشمل كمقارنة الفرد بنفسه بين وقت وآخر أو مقارنة أدائه بما يتمنى أن يكون عليه ويؤكِّد أنه كلما تعددت زوايا الرؤية زاد معنى الدرجة ثراء .

الفروق بين الاختبارات معيارية المرجع والاختبارات محكية المرجع :

١- يرجع أداء الفرد في الاختبارات معيارية المرجع إلى أفضل أداء في المجموعة ، في حين يرجع أداؤه في الاختبارات محكية المرجع إلى عدد الأهداف التي يحققها (أحمد السباعي ونجاح محمد النعيمي ، ٢٠٠١ ، ص ٩٩ ؛ جابر عبد الحميد جابر ، ١٩٩٦ ، ص ٢٣٣ ؛ نادية عبد السلام ، ١٩٨١ ، ص ١٨١) .

٢- تبنى الاختبارات معيارية المرجع لتسهل المقارنة بين الأفراد في المجال الذي يقيسه الاختبار ، في حين تبنى الاختبارات محكية المرجع لتقويم أداء الفرد في ضوء كفايات معينة (إبراهيم مبارك الدوسري ، ٢٠٠٠ ، ص ١٣٥) .

وهكذا يهدف القياس معياري المرجع إلى مقارنة أداء الفرد بأداء أقرانه ، بينما يهدف القياس محكي المرجع إلى معرفة ما حققه الفرد .

٣- تفسر درجات الاختبار معياري المرجع بالنسبة لدرجات أفراد آخرين ، في حين تفسر درجات الاختبار محكي المرجع في ضوء محك خارجي مسبق التحديد .

فالأساس في تفسير الدرجة في الاختبار معياري المرجع هو المجموعة المعيارية التي ينتمي إليها الفرد ، في حين يتطلب تفسير الدرجة في الاختبار محكي المرجع تحديد درجة قطع تدل على الكفاءة المناسبة (فؤاد أبو حطب وآخرون ، ١٩٩٧ ، ص ٤٢١) .

٤- تصمم الدرجات معيارية المرجع لتشير إلى المدى النسبي من المعارف والمهارات التي حققها الطالب ، في حين تصمم الاختبارات محكية المرجع لتعرّف المدى المطلق من المعارف والمهارات التي حقّقها الطالب (Friedenberg , 1995 , p. 89) .

وهذا ما تؤكده نادية عبد السلام (١٩٨٦ ، ص ٥٥٨) مشيرة إلى أنه " بينما يجاهد القياس جماعي المرجع لتحديد المستوى النسبي ، فإن القياس محكي المرجع يجاهد لتحديد المستوى المطلق . ومحاولة تحديد مستوى المختبرين بالنسبة إلى موقع كل منهم اتجاه الآخر ، يختلف تماماً عن محاولة تحديد ما إذا يستطيعون فعله " .

٥- يميِّز القياس معياري المرجع بدرجة عالية بين الأفراد الذين يملكون قدراً مختلفاً من الصفة المدروسة حيث تبنى مفردات الاختبار متوسطة الصعوبة وتتوزع الدرجات بقدر معقول من التباين أي تكون منتشرة أكثر من أن تكون متجمعة ، أما في القياس محكي المرجع يميل توزيع الدرجات للتجمع عند نهاية الطرف الأعلى للمقياس مع قدر قليل جداً من التباين (نادية عبد السلام ، ١٩٨١ ، ص ١٨٥) .

وبهذا تحدّد ركيزة الاختبار معياري المرجع في الوسط في حين تحدّد نقطة ارتكاز الاختبار محكي المرجع عند الطرفين حيث تدل أعلى درجة في المقياس على مستوى التمكن الكامل أو على أعلى مستوى للأداء الصحيح لبعض القدرات ، في حين تدل الدرجة التي تكون عند أسفل المقياس على أدنى مستوى لهذه القدرة والذي قد يصل أحياناً للصفر (أنور الشرقاوي ، ١٩٩٦ ، ص ٢٧) .

٦- يتم التركيز في الاختبار معياري المرجع على التباين لذلك تستبعد المفردات السهلة جداً أو الصعبة جداً وكذلك المفردات غير المميزة وذلك لتحقيق التباين بين المختبرين . أما في حالة الاختبار محكي المرجع فإن الهدف هو التأكد من وصول الطالب إلى محك الأداء وإتقان الأهداف التعليمية وعلى هذا لا تستبعد المفردات طالما أنها تقيس الهدف المراد تحقيقه بشكل مناسب .

ويوضح فؤاد أبو حطب وآخرون (١٩٩٧ ، ص ٤٢١) ذلك مشيرين إلى أن عناصر الاختبار معياري المرجع تكتب بحيث تنتج أقصى قدر من التنوع في الأداء بين التلاميذ في حين تكتب عناصر الاختبار محكي المرجع بحيث تمثل نطاق المرامي كما يراه كاتب الاختبار .

وهكذا تختار العناصر في الاختبارات معيارية المرجع بحيث تكون متوسطة الصعوبة ومرتفعة في معامل تمييزها ، في حين تكون الفقرات في الاختبارات محكية المرجع ذات مستوى صعوبة يتناسب مع المهام التعليمية وتستبعد بعض العناصر إذا ثبت ضعف علاقتها بالمحتوى المراد قياسه (إبراهيم مبارك الدوسري ، ٢٠٠٠ ، ص ١٣٦ ؛ صلاح الدين أبو ناهية ، ١٩٩٤ ، ص ٥١) .

بالإضافة إلى ما سبق يميل الطلاب لأن يجدوا مفردات الاختبار محكي المرجع سهلة وتجاب بشكل صحيح غالباً حيث يتم (٨٠%) منهم وحدة التعليم المتوقع أن تجاب كل مفردة منها بشكل صحيح ، بينما يتوقع في الاختبار معياري المرجع أن تتم نسبة (٥٠%) من الطلاب ذلك (Kubiszyn & Borich , 2000 , p. 38) .

٧- يعتمد تحديد صدق وثبات الاختبار معياري المرجع على الارتباط والتباين في الدرجات وذلك على خلاف طريقة تحديدها في الاختبار محكي المرجع الذي يحدد صدقه بمدى ارتباطه بالهدف المراد تحقيقه ويقاس ثباته عن طريق اتساق قرارات التصنيف .

٨- تلائم الاختبارات معيارية المرجع التقويم التجميعي لقياس نتائج جهود التعلم ، أما الاختبارات محكية المرجع فهي أكثر فائدة في التقويم التكويني وذلك لتوجيه عملية التعلم (Ebel , 1987 , p. 242) .

ويؤكّد فؤاد أبو حطب وآخرون (١٩٩٧ ، ص ٤٢٠) ذلك مشيرين إلى أن " الترجيع إلى المعيار يعطي نتائج تعيينية أي أنها تخبرنا أين يقف الفرد أو مجموعة الأفراد ، بينما الترجيع إلى المحك يعطينا نتائج تشكيلية أي أنها تخبرنا في أي المجالات يمكن أن يوجه التدريس بحيث يتيسر تحصيل الكفاءة " .

٩- تلائم الاختبارات معيارية المرجع أهدافاً تعليمية محلية ، في حين تصمم الاختبارات محكية المرجع لتقويم أهداف التعلم في الدولة (Ebel , 1987 , p. 242).

ويشير إبراهيم مبارك الدوسري (٢٠٠٠ ، ص ١٣٦) إلى أنه في الاختبارات معيارية المرجع يتم تعميم درجات الطالب بالنسبة للمجال الذي يقيسه الاختبار بشكل محدود ، في حين يتم التعميم في الاختبارات محكية المرجع من الدرجة التي اكتسبها الطالب إلى المجال الذي يقيسه الاختبار .

وتلخص الباحثة أهم الفروق بين اتجاهي القياس كما يلي :

- تقوم الاختبارات معيارية المرجع على مقارنة أداء الطالب بأداء الجماعة المعيارية التي ينتمي إليها وتهتم بتحديد الوضع النسبي له بالنسبة لأقرانه وعلى هذا فهي تقدّم تفسيرات نسبية ، كما تعتمد على خاصية التباين التي تتحقق عن طريق اختيار مفردات متوسطة الصعوبة ومرتفعة في معامل تمييزها ، وهي تلائم التقويم التجميعي لقياس نتائج جهود التعلم ويمكن تعميم نتائجها بشكل محدود .

- تقوم الاختبارات محكية المرجع على مقارنة أداء الطالب بمحك أداء مسبق التحديد لا يتأثر بطريقة أداء المختبرين أو تكوين الجماعة وبالتالي فهي تقدّم تفسيرات مطلقة ، كما أنها لا تعتمد على خاصية التباين ويتم اختيار المفردة على أساس مدى قدرتها على قياس الهدف المرغوب في تحقيقه ، وهي تلائم التقويم التكويني ويمكن تعميم نتائجها على المجال الذي يقيسه الاختبار .

هذا ويمكن اختيار أحد هذين النوعين من القياس ضمن الاعتبارات التالية :

١- الهدف من الاختبار : إذا كان الهدف من الاختبار هو الانتقاء يجب استخدام الاختبار معياري المرجع أما إذا كان الهدف من الاختبار هو التأكد من إحراز الطالب لأهداف التعلم المرغوب في تحقيقها ووصوله إلى مستوى الإتقان السابق تحديده عندئذ يجب استخدام الاختبار محكي المرجع .

هنا يجب التأكيد على ضرورة معرفة نوع المعلومات التي نحتاجها قبل اختيار نوع الاختبار وتطبيقه ، فإذا فشلنا في ذلك فربما يكون لدينا بيانات ولا نستطيع استخدامها لعمل قرارات هامة . فكثير من المعلمين يقررون أنهم لا يعلمون إلا القليل عن الطلاب بعد تطبيق الاختبار عما قبله وذلك بسبب فشلهم في تعريف الهدف من الاختبار قبل تطبيقه وفي هذه الحالة ربما نكون متسرعين في اتهام أو لوم الاختبار (Kubiszyn & Borich , 2000 , pp. 33-3+4) .

ويؤيد هذا الرأي صلاح الدين أبو ناهية (١٩٩٤ ، ص ص ٤٩-٥٠) الذي يشدّد على المعلم ضرورة تحديد ماهية البيانات التي يحتاجها قبل تصميم الاختبار ليتمكن من الاستفادة منها في اتخاذ قرارات تربوية هامة .

٢- اعتبارات التعلم : إذا كانت أهداف التعلم قليلة العدد ومميز كل منها عن الآخر بشكل كافٍ وكانت هامة على المستوى الفردي عندئذ يجب استخدام الاختبار محكي المرجع ، أما إذا كان جوهر التعلم غير منتهٍ من الجزئيات وكثير العدد عندئذ يستخدم الاختبار معياري المرجع (Ebel , 1987 , pp. 242-243) .

ويشير كبيسزين وبوريش (Kubiszn & Borich , 2000 , pp. 37-38) إلى أنه يجب أن تكون الاختبارات محكية المرجع محددة للغاية إذا كانت ستزود بمعلومات عن مهارات الفرد ، على العكس تميل الاختبارات معيارية المرجع لأن تكون عامة فهي تقيس مهارات عامة ومحددة في آن واحد لكنها تفشل في قياسها بمعنى الكلمة .

والآن هل يمكن استخدام أحد الاختبارين بمفرده أو كبديل عن الآخر ، أم أنه يجب استخدامهما معاً ؟؟

تجمع العديد من الآراء على ضرورة استخدام نوعي الاختبار معاً وذلك كما جاء في كتابات بابام Popham (١٩٧٦) الذي نشر موضوعاً ألحَّ فيه على جمع البيانات معيارية المرجع إلى البيانات محكية المرجع (Clarizio , Craig , & Mehrens , 1987 , p. 239) .

ويتفق مع هذا الرأي ما ورد عن فوزي طه إبراهيم ورجب أحمد الكلزة (٢٠٠٠ ، ص ١٧٨) في أن " الرد على ما يذاع من وقت إلى آخر حول هبوط مستوى التعليم هو استخدام المعيارين الاديومتري مع السيكومتري " .

فالاهتمام بالاختبارات محكية المرجع لا يعني إهمال الاختبارات معيارية المرجع نظراً لأهميتها في قياس مدى تحقيق الأهداف العامة ، وبهذا يمكن الاعتماد على كلا النوعين طبقاً للأهداف المنشودة والغرض الذي يستخدم الاختبار من أجله (محمد حسين سالم ، ١٩٩٥ ، ص ٢٧) .

ومن خلال العرض السابق تؤيد الباحثة ضرورة الجمع بين نوعي الاختبارات طالما أن كلاً منهما يزودنا بمعلومات مختلفة عن الآخر بشرط استخدام الأسس المناسبة لكل نوعية منهما عند تطبيقه ، مما يساعد على تحقيق التكامل في التقويم والنهوض بالعملية التعليمية إلى المستوى المطلوب .

بناء الاختبارات محكية المرجع :

اتضح من التعريف السابق بنوعي القياس والأسس التي يقوم عليها كل منهما أن وجه الخلاف الأساسي بينهما هو في الهدف ، فالأول يهتم بالمقارنة بين الطلاب ويعتمد على وجود الفروق بينهم فيما يركز الآخر على وصول الطالب محك النجاح وإتقانه لعدد محدد من الأهداف ، ومن هنا كان لابد من وجود اختلاف أيضاً في طريقة بناء هذه الاختبارات . وتعرض الباحثة فيما يلي بعضاً من الاتجاهات في بناء الاختبارات محكية المرجع :

يقدّم جرونلند (Grounlund , 1973 , pp. 24-32) عدة خطوات لبناء الاختبارات محكية المرجع كالتالي :

١- تحديد المجال الذي سيتم اختباره .

٢- تعيين الأهداف وتعريفها في حدود دقيقة .

٣- عمل مخطط تمهيدي للمحتوى .

٤- إعداد جدول المواصفات .

٥- وضع معايير أداء .

٦- كتابة مفردات الاختبار وجمعها .

كما يوضّح هامبلتون (Hambelton , 1980 , pp. 81-82) طريقة بناء هذه الاختبارات في الخطوات التالية :

١- تختار الأهداف أو يحدد المجال قبل أن تبدأ عملية إعداد الاختبار .

٢- توضح مواصفات الاختبار كلاً من أهداف الاختبار ، أشكال المفردات المقبولة ، عدد المفردات ، وتعليمات الإجابة .

٣- تكتب مفردات الاختبار وتقيّم لتحديد اتصالها بالأهداف التي أعدت لتحقيقها .

٤- يجمع الاختبار وتعد معايير تفسير أداء المختبر .

٥- يحلل صدق وثبات البيانات .

ويشير فؤاد أبو حطب وآخرون (١٩٩٧ ، ص ص ٤١٤-٤١٥) إلى أن بناء الاختبارات محكية المرجع يمر بالمراحل التالية :

١- يعد تخطيط المحتوى Content Outline تذكر فيه المهارات والمعرفة التي يحاول الاختبار قياسها .

٢- تحدد مجموعة الأهداف التي يجب أن يكون الطالب قادراً على أدائها.

٣- يميز المجال الذي يحدده كل هدف وتكتب بنود الاختبار وفق التفاصيل المحددة في ذلك المجال ثم يختار عشوائياً عنصران لكل هدف لبناء الاختبار .

٤- يتم التأكد من كون المهارات والمعرفة المقاسة بالاختبار لازمة لقياس الأهداف السلوكية المحددة في الخطوة (٢) .

٥- تحدد درجة القطع .

هذا وأن الطريقة الأكثر شيوعاً واستخداماً في بناء الاختبارات محكية المرجع هي الطريقة المقدّمة من قبل بابام Popham ، والتي تسمح بتكوين " نطاق شامل من المفردات يتميّز بالتحديد الدقيق بحيث يُمكّن مصمّم الاختبار من بناء مفردات متجانسة تقيس هدفاً سلوكياً معيناً " (إسماعيل الوليلي ، ١٩٩٦ ، ١٩) . وقد تم تفضيل واستخدام هذه الطريقة في العديد من الدراسات التي تمّت في هذا المجال كدراسة (آمال محروس ، ١٩٨٨ ؛ إسماعيل الوليلي ، ١٩٩٦ ؛ سعاد حسانين ، ٢٠٠٠ ؛ صلاح عبد الوهاب ، ٢٠٠٠ ؛ محمود إبراهيم ، ١٩٩٠ ؛ مصطفى كامل ، ١٩٩٩ ؛ نادية بعيبع ، ١٩٩٦) .

ويذكر بابام Popham أن الخطوات الأساسية في بناء الاختبار محكي المرجع هي :

١- تحديد المجال السلوكي الذي يقيسه الاختبار .

٢- تحديد الأهداف العامة للمجال السلوكي وتحليلها إلى أهداف سلوكية .

٣- إعداد مواصفات الاختبار .

٤- بناء مفردات الاختبار .

٥- تحديد درجة القطع .

٦- تقدير صدق وثبات الاختبار .

وسيتم شرح هذه الخطوات بالتفصيل كونها المستخدمة في بناء أدوات هذه الدراسة .

١- تحديد المجال السلوكي الذي يقيسه الاختبار :

الخطوة الأولى في تصميم الاختبار محكي المرجع هي تحديد مجال الأداء لوحدة تعليمية حجمها مقبول ، فوحدة دراسية تغطي أسبوعاً أو اثنين ربما تكون أكثر ما هو مرغوب به (Gronlund , 1973 , p. 24) .

ومن الخصائص المرغوب بها في تعريف المجال أن يكون وصفه قصيراً ليساعد على استخدامه لا من قبل مصمِّم الاختبار فحسب بل من قبل المربين ذوي الأعباء الكثيرة ، وأن تحدد فئة العناصر السلوكية موضع الاختبار تحديداً كافياً بحيث يتفق الحكام اتفاقاً كبيراً فيما يتصل بملاءمة العنصر الاختباري لقياس السلوك الموصوف في المجال (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ص ٢٤٠-٢٤١) .

ويحدّد جابر عبد الحميد جابر (١٩٩٦ ، ص ص ٢٤٣-٣٤٧) بعض الاعتبارات التي تساعد على تحديد المجال المراد قياسه وذلك كما يلي :

١- الزمن الذي يستغرقه التعليم : وهو مقدار الزمن المستغرق ليتمكّن المتعلم من إظهار السلوك الذي يحدده المجال .

٢- أولويات محددة : أي أن نضع حداً يتصل بعدد المجالات التي تستخدم في الجهد التقويمي ثم نضع أولويات بحيث نستوعب أهم العناصر السلوكية التي نحاول تحقيقها في العدد الحالي من المجالات .

٣- تجانس العنصر : أي أن نضع حداً أعلى لوصف المجال يساعدنا على التوصل لنمط متجانس من العناصر .

٤- حجم المجال : أي أن نحدد حجم المجال بدرجة تساعدنا على إنتاج العناصر التي إن أجيب عنها إجابة صحيحة فستساعدنا على التنبؤ بالإجابات السليمة لعناصر أخرى تعبر عن المجال وتعكسه .

٥- الاختيار من بين بدائل المجال المتنافسة : يتم اختيار مجال بدلاً من غيره كونه ممثلاً لفئة من العناصر السلوكية والمهارات أو المعرفة التي لها دلالتها أو مغزاها .

٦- القابلية لانتقال الأثر بين بدائل المجال : المجال الذي له أعظم قابلية للتعميم هو الذي ينبغي اختياره .

٧- القابلية لانتقال الأثر خارج المجال : الدرجة التي ينتقل بها أثر المجال إلى خارجه إذا تم إتقانه .

هذا وإن أهمية التعيين الواضح لمهام المجال الذي ستسحب منه مفردات الاختبار لا يمكن تقديرها فبدون بيان واضح لمجال المفردات التي يمكن استخدامها لقياس الإتقان لا يمكن الحكم فيما إذا كانت المفردات المستخدمة ممثلة بالفعل للمجال أو مقتصرة على مجموعة جزئية فحسب .

وهكذا فإن تصورات مختلفة لنفس المجال ستزوّد بمفردات مختلفة وقرارات مختلفة عن الطلاب . فالوصف الواضح للمجال سيسمح لكل شخص أن يحكم على صدق محتوى الاختبار وسيمكن مستخدمي الاختبار من تحديد فيما إذا كانت المفردات ترتبط بمهام المجال بشكل مقبول أو تمثله بشكل ضيق جداً (Millman , 1980 , p. 33) .

وهذا ما يؤكّده هامبلتون وآخرون (Hambelton , Swaminathan , Algina , & Coulson , 1978 , p. 32) في أنه إذا لم يكن المجال مضبوطاً بشكل واضح فمن غير الممكن اختيار عينة ممثلة لمفردات الاختبار منه ، وبما أننا راغبون بتفسير أداء المختبر على عينة من مفردات الاختبار التي تقيس هدفاً محدداً كتقدير لمستوى إتقان ذلك المختبر في مجال أكبر من المفردات التي تقيس ذلك الهدف ، فمن الضروري امتلاك مجال مفردات معين بوضوح لاختيار عينة ممثلة من المفردات منه . وإذا لم يعرف المجال السلوكي بوضوح فسنحصل على ما يسميه بابام (١٩٧٤) " الاختبار المرجع إلى السحابة " .

٢- تحديد الأهداف العامة للمجال السلوكي وتحليلها إلى أهداف سلوكية :

تحدد هذه الخطوة الأهداف التعليمية العامة للمجال الذي تم تحديده في الخطوة الأولى ويمكن أن تكون الأهداف العامة هذه الأهداف العامة لتدريس المادة موضع الاختبار أو أهداف المقرّر في المرحلة الدراسية ، ثم تشتق منها مجموعة الأهداف السلوكية Behavioral Objectives التي تغطي المجال .

هذا ويعرّف الهدف السلوكي على أنه " صياغة توضّح رغبة في تغيير متوقَّع في سلوك التلميذ وهذه الصيغة تعبّر عن مزايا يمكن ملاحظتها وقياسها " ، بمعنى أنه " سلوك سوف يقوم به التلميذ في نهاية الدرس ليدل على حدوث التعلم " (جمال عطية ، ١٩٩٢ ، ص ص ٢١٦-٢١٧) .

وهكذا فإن " الصورة العامة للأهداف تترجم إلى أهداف سلوكية أو اجرائية أي عبارات تحتوي على سلوك أو إجراء قابل للملاحظة والقياس ، وهذه العبارات تصف السلوك النهائي المتوقَّع أن يؤديه التلميذ بعد تعلمه لجزء من المادة " (وليم عبيد ومحمد أمين المفتي وسمير ايليا القمص ، ٢٠٠٠ ، ص ٤٣) .

٣- مواصفات الاختبار :

تحدد هذه الخطوة المجال السلوكي الذي يقيسه الاختبار بصورة مفصلة حيث توضح أنواع المفردات التي ستستخدم في بناء الاختبار وطريقة الإجابة عليها مما يساعد على بناء مفردات متجانسة تقيس هدفاً سلوكياً محدداً .

توضع هذه المواصفات لتوجه معد المفردة بناء على الاعتقاد بأن " الوضوح الأكثر أفضل من الوضوح الأقل " والتي ترجمت فيما بعد إلى " تفاصيل أكثر أفضل من تفاصيل أقل " (Popham , 1994 , p. 16)

وهكذا فإن ما يميز القياس محكي المرجع هو الوضوح الذي يوصف به المجال السلوكي المقاس ، حيث تحاول مواصفات الاختبار أن تعين بإحكام شديد أنواع المفردات التي ستستخدم في بناء الاختبار وذلك من خلال الوصف المفصل في مواصفات مفردة الاختبار التي تحكم توليد مفردات الاختبار وفي مواصفات الاختبار التي تحكم التركيب الكلي للاختبار (Popham , 1992 , pp. 16-17) .

وإذا لم يصف الاختبار محكي المرجع ما يقيسه بشكل واضح فإنه لن يقدم فائدة عن القياسات معيارية المرجع . فإضافة جدول محكم الوصف (مواصفات الاختبار) يساعد على بناء مفردات منسجمة ويزود بتفسير لايخطئ عن معنى أداء المختبر ويسمح بوصف

أدائه بناء على نسبة المفردات التي يستطيع الإجابة عليها بشكل صحيح . وبما أن مفردات الاختبار تشكل عينة من المفردات الممكنة التي يمكن تعميمها لتصل إلى السلوك في السؤال فإننا نستطيع أن نقدم ثقة كبيرة في صحة تفسيراتنا (Popham , 1980 , pp. 16-17) .

بالإضافة إلى ما سبق فإن وجود مواصفات اختبار واضحة تساعد على تقدير معايير إتقان أو درجات اجتياز . ومع أن وضع درجات الاجتياز يتضمن عنصر الحكم إلا أنه يمكن أن تتم هذه المهمة بشكل أقل كيفية فيما يبدو عندما تحدد المهام بوضوح (Millman , 1980 , p. 33) .

هذا وتشتمل مواصفات الاختبار على خمسة عناصر وهي :

٣-١- الوصف العام :

المكون المبدئي لمجموعة مواصفات الاختبار محكي المرجع هو جملة أو اثنتين تقدم كوصف عام لما يقيسه الاختبار وذلك بهدف التزويد بنظرة محكمة لمجموعة السلوكيات الواجب وصفها أكثر في المواصفات فيما بعد ويرجع هذا المكون عادة إلى " الهدف " (Popham , 1980 , p. 22) .

٣-٢- عينة المفردة :

المكون الثاني هوعينة المفردة وهو جزء متمم للوصف العام بالإضافة إلى تعليمات توجه للطالب وقد تستخدم في الاختبار نفسه . وبما أنّ الاختبار يتألف من عدد من المفردات القصيرة نسبياً ولذلك فإن اختيار واحد من هذه المفردات لأغراض توضيحية مسألة بسيطة في العادة ، أما إذا كان الاختبار أكثر تعقيداً والمفردات أكثر طولاً فيصبح من الصعب التزويد بعينة المفردة أحياناً .

وهناك سببان للتزويد بعينة المفردة :

أولاً : ربما يجد الأشخاص الذين يستخدمون المواصفات خصوصاً المشغولين منهم حاجتهم لوصف الاختبار بشكل واف من خلال عبارة الوصف العام والمفردة التوضيحية فقط ، مثل هؤلاء الأشخاص إذا أجبروا على قراءة المجموعة الداخلية للمواصفات ربما يتجنبونها بشكل كامل لكنهم ربما يغروا بالتوصية وان لم تكن كاملة من خلال الوصف العام وعينة المفردة .

ثانياً : أنها تزود بعبارات عن شكل المفردة لهؤلاء الذين سيعدون المفردات التي تشكل الاختبار (Popham , 1980 , p. 23) .

٣-٣- خواص المثير :

توصف في هذه الخطوة المثيرات التي تقدمها مفردات الاختبار مما يساعد على تكوين مفردات تفي بشروط المواصفات المطلوبة . وهنا يجب تحديد كل العوامل المؤثرة في مجموعة مفردات الاختبار وبالتالي علينا أن نفكر ما هي هذه العوامل وكيف يمكن أن توصف بحيث تكون أكثر دقةً وإحكاماً فعدم الدقة في هذه الخطوة سينتج مواصفات غير دقيقة أو مضللة على نحو أسوأ (Popham , 1980 , p. 23) .

ومن الضروري تقديم تفصيلات كثيرة للتقليل من الغموض في جانب المثير وتختلف هذه التفصيلات من موضوع إلى آخر ومن مجال إلى آخر . وعموماً فإن عناصر المثير يجب أن :

١- تحدِّد المحتوى الذي تستند إليه عناصر الاختبار وتقوم عليه .

٢- تصف التعليمات التي يتلقاها الأفراد ليجيبوا على هذه العناصر (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ٢٤٨) .

٣-٤- خواص الاستجابة :

تركز هذه الخطوة على استجابة المختبر للعناصر التي ولدت وفقاً لمقطع المثير وهناك نوعان فقط لاستجابة المختبر فإما أن يختار من بين خيارات الاستجابة المقدمة في الاختبار كأسئلة الصح والخطأ وأسئلة الاختيار من متعدد أو أن يبني استجابة كأسئلة المقالة وأسئلة الإجابة القصيرة ، وإذا تضمن الاختبار اختيار استجابة عندئذ يجب التزويد بقواعد كافية لا لتحديد طبيعة الإجابة الصحيحة فقط ولكن أيضاً لتحديد طبيعة خيارات الإجابة الخاطئة . ففي أسئلة الاختيار من متعدد يجب أن نحدد أولاً موجهات لوضع الإجابة الصحيحة ثم نزود بخيارات متعددة للإجابة الخاطئة التي قد تشكل مشتتات المفردة ، ولا يكفي القول أن مثل هذه المشتتات ستكون " غير صحيحة " بل يجب أن توضح الطبيعة الخاطئة لهذه الإجابات بعناية (Popham , 1980 , pp. 23-24) .

٣-٥- ملحق المواصفات :

تضاف أحياناً هذه الخطوة إلى العناصر السابقة بحيث تشتمل على تفصيلات اضافية تؤدي إلى مزيد من التوضيح لمحتوى المجال السلوكي أو محتوى مفردات الاختبار (إسماعيل الوليلي ، ١٩٩٦ ، ص ٢٣) .

٤- بناء مفردات الاختبار :

يعتمد بناء المفردات على مواصفات الاختبار ويفترض في هذه المفردات أن تقيس المجال السلوكي بدرجة كبيرة من الدقة وأن تكون مشتقة بشكل متجانس من المجال السلوكي موضع القياس الذي تحدده مواصفات الاختبار ، وبالتالي فإن عملية بناء المفردات تكون سهلة في حالة وضوح المواصفات وخلوها من الغموض (إسماعيل الوليلي ، ١٩٩٦ ، ص ص ٢٣-٢٤) .

هذا ويتم اختيار المفردات بطريقة عشوائية بالاعتماد على تطابقها مع الأهداف وحساسيتها للمعالجات التعليمية (Berk , 1988 , pp. 367-368) ، وهناك اتجاهان مختلفان لبناء مفردات الاختبارات محكية المرجع يستخدم أحدهما شكل المفردة لينتج مجموعة من المفردات تقيس كلها نفس الهدف ، في حين ينتج الاتجاه الآخر المفردات بأي وسيلة متاحة وتعدل أو تحذف المفردات التي لم يتم أداؤها كما هو مرغوب فيه على أساس عملي . ويتحدد اختيار الفرد لأحد هذين الاتجاهين تبعاً لطبيعة الأهداف السلوكية (نادية عبد السلام ، ١٩٩٦ ، ص ١٦٣) .

وقد أشار جرونلند (Grounlund , 1973 , pp. 34-37) إلى أن كتابة المفردات فن يتطلب ممارسة كبيرة وقدم مجموعة من القواعد يجب تحقيقها عند كتابة المفردات :

١- يستلزم كتابة مفردات الاختبار وصفاً دقيقاً للسلوك في مخرجات التعلم : فعملية اعداد مفردات ترتبط مباشرة بمخرجات التعلم الواجب قياسها مسألة حكم منطقي وتتألف من صلة كل مهمة في الاختبار بمهمة أداء في مخرج التعلم .

٢- كتابة المفردة بحيث تكون المهمة الواجب أداؤها واضحة ومحدَّدة : وذلك بصياغة المشكلة بلغة سهلة ومباشرة واتباع القواعد في وضع النقط والفواصل بشكل صحيح ، فلا يجب أن يفشل الطالب في الإجابة على مفردة بسيطة لأنه لم يفهم نوع الأداء المتوقع أن يظهره .

٣- كتابة المفردة بحيث تكون خالية من مادة غير وظيفية : يجب أن تحتوي المفردة على مادة وثيقة الصلة بالمشكلة الحالية فقط ، فإضافة مادة غريبة لهدف يطيل المفردة على نحو غير ملائم وفي بعض الحالات ربما يشوِّش الطالب أو يزوِّد بمفاتيح الحل .

٤- كتابة المفردة بحيث لا تكون متصلة بعوامل تمنع الطالب من الاستجابة الصحيحة : يجب أن تصمَّم المفردة بحيث تستدعي نوعاً مخصصاً من الأداء وأن تقلل تأثير العوامل غير المرتبطة بالهدف الرئيسي للمفردة بقدر الإمكان .

٥- كتابة المفردة بحيث تكون غير متصلة بعوامل لا تعوّد الطالب على الإجابة الصحيحة : وذلك بالتركيز على إعداد مفردات تستدعي سلوكاً محدداً موصوفاً في مخرجات تعلم محدّدة وإهمال العوامل غير المتصلة بالإجابة الصحيحة كالتناقض اللغوي الذي يسيطر على بعض أو كل الإجابات الخاطئة الممكنة والمرافقة اللفظية التي تجعل الحل الصحيح واضحاً .

٦- كتابة المفردة بحيث لا تزوّد بتعليمات لإجابة مفردات أخرى في الاختبار .

٧- كتابة المفردة بحيث تكون عند مستوى مناسب من الصعوبة : حيث يتحدّد مستوى صعوبة المفردة حسب الطبيعة الدقيقة لمخرجات التعلم المقاسة .

٨- كتابة المفردة بحيث تكون الإجابة الصحيحة واحدة يتفق عليها الخبراء .

٩- كتابة المفردة بشكل إيجابي قدر الإمكان : وهناك ثلاثة أسباب لوضع المفردة في الشكل الإيجابي :

- من وجهة نظر التعلم من المرغوب فيه عادة التركيز على الحقائق والمفاهيم والمبادئ التي يجب أن يتعلّمها الطالب أكثر من التركيز على الاستثناءات .

- كون الطالب يعرف ما هو " ليس الحالة " لا يزوّد بالثقة التامة أنه عرف ما هي الحالة .

- غالباً ما تهمل الكلمة " لا " في مفردات الاختبار ويستجيب الطالب وكأن العبارة صحيحة .

١٠- كتابة مفردات كافية لتعيين مخرجات التعلم المقاسة بشكل ملائم : يستخدم جدول المواصفات كموجّه لكتابة المفردة حيث يتم الإشارة إلى عدد المفردات اللازمة لقياس كل هدف وكل مجال من المحتوى .

بعد الانتهاء من بناء المفردات يتم جمعها لتشكّل الاختبار وذلك وفق الخطوات التالية :

١- يعاد النظر في مفردات الاختبار للتأكد من صلتها بمخرجات التعلم وخلوها من عيوب فنية .

٢- تفحص مجموعة المفردات ككل لتحديد ملاءمتها مع جدول المواصفات .

٣- تجمع كل المفردات التي تقيس نفس الهدف التعليمي معاً .

٤- ترتب المفردات من السهل إلى الصعب داخل الاختبار وداخل كل مقطع منه .

٥- تكتب توجيهات واضحة للاختبار ككل ولكل نوع مفردة منفصل (Gronlund , 1973 , p. 31) .

٥- تحديد درجة القطع :

يتم تحديد درجة القطع باستخدام إحدى الطرق المتاحة لهذا الغرض والتي سيتم الإشارة إليها لاحقاً من هذا الفصل .

٦- تقدير صدق وثبات الاختبار :

يتم تقدير صدق وثبات الاختبار محكي المرجع كما سيتم شرحه .

صدق الاختبارات محكية المرجع :

الصدق من الصفات الأساسية الواجب توافرها في أداة قبل تطبيقها وهذا لا يعد مشكلة في القياسات الفيزيائية لأن قياس طول لا يختلط بسهولة بقياس وزن ، لكن الأمر مختلف في القياسات التربوية ، فربما يريد المعلم أن يقيس مستوى معارف عليا لكنه يختار أداة تقيس شيء مختلف تماماً ، والأكثر سوءاً قد تعكس بعض درجات القياس تحيز الدرجة أكثر من أداء الطالب .

كما يجب التأكد من أن أدوات التقييم ملاءمة للطلاب الذين يأخذونها ، مثلاً إذا لم يعط الطالب فرصة لتعلم المادة أو إذا احتوى اختبار رياضيات على مسائل مكتوبة باللغة الانكليزية لكن الطالب لم يتعلم الانكليزية بعد فلن يكون الاختبار قياساً صادقاً لقدرته على حل المسائل الرياضية وهكذا فإن الصدق هو المصطلح المستخدم من قبل اختصاصيي القياس للإشارة إلى مدى قياس الاختبار لما استخدم لقياسه حقيقة ، أي كم هو وثيق الصلة بموضوع القياس (Ward & Murray – Ward , 1999 , pp. 76-77) .

وعلى هذا يُتفق على أن صدق الاختبار يعني أن يقيس الاختبار ما صمّم لقياسه (Alken , 1997 , p. 93 ; Peers , 1996 , p. 28) .

أي أنه يركّز على ما يقيسه الاختبار وكيف يفعل ذلك وبهذا يخبرنا ما يمكن أن نستنتجه من درجات الاختبار ، ولا يمكن أن يسجَّل في حدود عامة كأن يقال أن للاختبار صدق " مرتفع " أو "منخفض " بشكل مجرد بل يجب أن يعيَّن تبعاً للاستخدام الخاص الذي أعدّ له الاختبار (Anastasi , 1990 , p. 139) .

وهكذا " يرتبط الصدق دائماً بالاستخدام الخاص بالمقياس ولا يمكن أن تكون نتائج المقياس صادقة لجميع الأغراض " (رجاء أبو علام ، ١٩٩٩ ، ص ٤١٣) .

أي أن الصدق يتعلق بهدف استخدام الاختبار وليس فقط بغرض وتكوين الاختبار ، ويعتبر الفرق الأساسي بين القياسين معياري المرجع ومحكي المرجع هو في تأسيس صدق الاختبار (نادية عبد السلام ، ١٩٩٦ ، ص ٧١) .

أما عن طرق تحديد صدق الاختبار محكي المرجع فهي :

١- صدق المحتوى

٢- الصدق الوظيفي

٣- صدق تحديد المجال السلوكي

وسيتم شرح النوع الأول بالتفصيل كونه المستخدم في هذه الدراسة .

١- صدق المحتوى :

أكثر الأنواع الثلاثة استخداماً وقد وافق كثير من مطوري الاختبارات أنه لتأييد تفسيرات درجات الاختبار محكي المرجع من الضروري تقويم " صدق محتواها " بمعنى أن تتم أحكام تركِّز على مدى الاتصال بين محتوى الاختبار والأهداف الواجب قياسها عن طريق الاختبار ، وبما أن التحليل في حدود محتوى الاختبار لذلك يستخدم عادة التعبير " صدق المحتوى " (Hambelton , 1980 , p. 82) .

وعلى هذا يفضل واضعو الاختبارات محكية المرجع صدق المحتوى نظراً للعلاقة بين المفردات والموضوعات المحددة (نادية عبد السلام ، ١٩٩٦ ، ص ٧٣) .

ويلاحظ أنه خلال بناء الاختبارات محكية المرجع يجري التركيز بشكل رئيسي على صدق المحتوى فالمجال الذي تعيَن منه المفردات بشكل ملائم للأهداف بالإضافة إلى تخطيط الاختبار تخدم كموجهات لاعداد اختبارات صادقة المحتوى ، وهكذا تعرّف قائمة الأهداف التعليمية ومخطط المحتوى مجال السلوك الواجب قياسه ، ويوضّح جدول المواصفات طبيعة عينة الاختبار ، إذا أعدت مفردات الاختبار بعد ذلك بعناية وفقاً لمواصفات الاختبار ستزود استجابات الطلاب بمؤشرات أداء صادقة . أي أن صدق المحتوى هو مسألة حكم ، علينا أن نحكم على مناسبة وكفاية عينة مفردات الاختبار لقياس مخرجات التعلم المتوقعة (Gronlund , 1973 , p. 47) .

وهناك طريقتان يمكن استخدامهما لتقويم صدق محتوى مفردات الاختبار محكي المرجع وهما :

١- الحكم على مفردات الاختبار من قبل اختصاصيي المحتوى ، حيث يركز على مدى التناسق بين مفردات الاختبار والمجالات التي صممت لقياسها . هذا وتتركز الأسئلة التي توجه للمحكمين عن صدق محتوى مفردات الاختبار حول نقطتين هامتين :

- هل كتبت مواصفات المجال بوضوح ؟

- هل هناك اتفاق فيما بين مواصفات المحتوى التي تعين مجموعة مفردات المجال بشكل كاف ؟

وقد اقترح روفينلي وهامبلتون تقنيات لجمع وتحليل أحكام اختصاصيي المحتوى حول صدق محتوى مفردات الاختبار بغية تصنيف المفردات بالنسبة إلى مجموعة الأهداف التي يقيسها الاختبار في إحدى الحالات التالية :

+١ : تقابل الاعتقاد التام أن المفردة تقيس هدف ما .

٠ : عدم التأكد فيما إذا كانت المفردة تقيس هدف ما .

-١ : تقابل الاعتقاد التام أن المفردة لا تقيس هدف ما (Hambelton et al. , 1978 , p. 34) .

ويتضح هنا أن وجود مجال ذي مواصفات محددة يسهل تحديد مطابقة فقرات الاختبار للمجال وأنه " كلما كان المجال الدراسي أو الهدف التعليمي محدداً بشكل دقيق كلما زاد الاتفاق بين المحكمين على مدى تطابق الفقرات في ذلك المجال أو الهدف . وإذا شكّك المحكمون في صدق الاختبار فان ذلك يتطلب تعديل الفقرات أو تعديل الأهداف أو تعديل كليهما معاً " ولا يشترط أن يكون عدد المحكمين كبيراً ، ولكن يجب أن يكونوا على معرفة جيدة بمحتوى المجال الدراسي وبمستوى الطلاب المستهدفين ويمكن الاستعانة بالمعلمين في هذه المهمة (صلاح الدين أبو ناهية ، ١٩٩٤ ، ص ص ٣٥٠-٣٥١) .

٢- تركّز هذه الطريقة على مقارنة مستويات الصعوبة المتوقعة والملاحظة لكل مفردة حيث يقدَّر مستوى الصعوبة المتوقع للمفردة ثم يقارن مع مستوى الصعوبة الملاحظ بعد تطبيق الاختبار فإذا كان المستويان متقاربين اعتبرت المفردة مناسبة وإلا يعاد فحص المفردة ، وبنفس الطريقة إذا وجدت مجموعة من المفردات تقيس جانباً من المعارف مرتب بشكل هرمي يتوقع من الطلاب الذين استجابوا بطريقة صحيحة على المفردات ذات المستويات العليا أن يستجيبوا أيضاً بطريقة صحيحة على المفردات ذات المستويات الأدنى في نفس

التدرج الهرمي ، وعلى هذا فإن هذه الطريقة تعتمد على التنبؤ ويمكن أن تتم باستخدام إحدى الطرق التالية :

أ- نكوّن مجموعتين من المختبرين ونبدأ مع المجموعة ذات المعرفة الأقل في مجال محتوى الاختبار حيث يعطى لها الاختبار ثم يعاد تطبيقه عليها ثانية بعد تقديم معالجة تعليمية مناسبة ، فإذا وجدنا زيادة في مستوى سهولة المفردة قبل الاختبار وبعده اعتبرت المفردة صادقة .

ب- نختار مجموعتان إحداهما تجريبية والثانية ضابطة ، تقدم المعالجة التعليمية للمجموعة التجريبية فقط ثم تختبر المجموعتان ، فيدل ارتفاع نسبة الاستجابات الصحيحة في المجموعة التجريبية على صدق المفردة وذلك بفرض فاعلية المعالجة التعليمية .

ج- يقدَّم الاختبار لمجموعة من المختبرين ثم يتم تصنيفهم إلى مجموعتين متفوقين وغير متفوقين بناء على درجاتهم في الاختبار ، ويعتبر تمييز المفردة بين المجموعتين دالاً على صدقها (نادية عبد السلام ، ١٩٩٦ ، ص ص ٧٤-٧٦) .

ويجدر بالذكر أن صدق المحتوى هو إحدى خواص الاختبار أي أنه لن يتغير باختلاف مجموعة المختبرين إلا أن صدق تفسيرات درجة الاختبار يمكن أن يتغير من مجموعة لأخرى ، فمثلاً إذا طبق اختبار محكي المرجع بالخطأ تحت ظروف السرعة العالية فإن صدق التفسيرات المبنية على درجات الاختبار ستكون أقل مما لو طبق الاختبار في حدود زمن مناسب أكثر . أي أن صدق محتوى الاختبار لا يتغير لكن صدق أي تفسير يمكن أن يتغير من حالة اختبار لأخرى (Hambelton et al. , 1978 , pp. 38-39) .

وبهذا فإن الصدق " يشير إلى الدرجة التي يمكن بها تفسير نتائج المقياس ولذلك نحن نتكلم عن صدق التفسيرات التي نخرج بها من النتائج " (رجاء أبو علام ، ١٩٩٩ ، ص ٤١٣) .

٢- الصدق الوظيفي :

يقصد بهذا النوع من الصدق أن يؤدي الاختبار الوظيفة التي أعدَّ لها ، وهناك فرق بين كون نتائج الاختبار وصفية فقط لمجال السلوك الذي يفترض أن تقيسه وبين كونه ملائماً لما وضع من أجله ، والدقة التي يحقق بها الاختبار محكي المرجع الغرض الذي أعدَّ له يمكن أن يوصف بالصدق الوظيفي (نادية عبد السلام ، ١٩٩٦ ، ص ١٤٥) .

٣- صدق اختبار المجال السلوكي :

" يتعلق هذا الصدق بالدقة التي يختار بها الموجه مجالاً محدداً يستخدم كمؤشر يوضح مستوى المتعلم بالنسبة إلى مجال عام مثل هدفي إدراكي أو وجداني " (نادية عبد السلام ، ١٩٩٦ ، ص ١٤٧) .

ويقابل هذا النوع من الصدق في الاختبارات معيارية المرجع صدق التكوين .

وهناك نوع خاص من صدق بناء درجات الاختبار محكي المرجع يستحق تركيزاً خاصاً ألا وهو صدق القرار الذي يعتبر الدليل على دقة قرارات تصنيف الإتقان وعدم الإتقان . وعندما لا يكون صدق القرار متاحاً (ممكناً) فلا داعي أيضاً لحساب الثبات لأن أي مؤشر ثبات مرتفع مبني على اختبار " غير صادق " قد يعني أن الاختبار يستطيع تصنيف الطلاب باتساق في مجموعات خاطئة ، وهكذا فإن عمل قرار اتساق بدون دقة عمل القرار له قيمة مشكوك فيها في التقويم محكي المرجع (Berk , 1988 , p. 369) .

ثبات الاختبارات محكية المرجع :

يعد الثبات من الشروط الأساسية المطلوب توافرها بالإضافة إلى الصدق في أي أداة قياس ، فعندما يقوم باحث ما ببناء أداة خاصة بعمله عليه أن يتأكد أن هذه الأداة تتمتع بدرجة مقبولة من الثبات حتى يمكنه الثقة بالنتائج التي ستوفرها هذه الأداة . كما أن اعتماد الباحث على أداة جاهزة سبق تقنينها لا يعطي المبرر الكافي لاستخدامها دون التأكد من ثباتها لأن الثبات وكما سيتم توضيحه لاحقاً ليس من خصائص الاختبار وإنما هو خاصية من خواص درجات الاختبار .

هذا ويعرَّف ثبات الاختبار محكي المرجع على أنه اتساق الاختبار في صنع تقديرات مستوى إتقان المختبرين في مجال المحتوى (Borg & Gall , 1983 , p. 290) ، وهو ما " يقصد به اتساق قرارات تصنيف الأداء خلال القياسات المتكررة " (آمال محروس ، ١٩٨٨ ، ص ٥٣) .

وقد اقترح هامبلتون ونوفيك أن يعرَّف ثبات قرارات الإتقان في حدود اتساق القرارات الناتجة عن تطبيق أشكال متكافئة من الاختبار (Hambelton et al. , 1978 , p.21) .

ويذكر شحته عبد المولى (١٩٩٩ ، ص ١٠) أن ثبات الاختبار محكي المرجع هو " اتساق قرار تصنيف مجموعة محددة من الأفراد اعتماداً على درجاتهم في الاختبار بناء على درجة القطع أو مستوى القدرة المحدد ، وهذا المستوى هو أدنى مستوى للأداء المقبول

للإتقان أو عدمه على أساس أن هذا المستوى يتخذ أساساً لتصنيف الطلاب إلى متقنين وغير متقنين " .

وهكذا يشير ثبات الاختبار محكي المرجع لاتساق موقع الطالب في فئة الناجحين / الفاشلين ، فإذا تمَ وضع كل أو معظم الطلاب في نفس فئة النجاح / الفشل خلال القياسين سيوصف الاختبار بأنه ثابت ، أما إذا اجتازت فئة من الطلاب الاختبار الأول وفشلت في الاختبار الثاني وفئة من الذين فشلوا في الاختبار الأول اجتازوا الاختبار الثاني سيكون الاختبار غير متسقاً أو غير ثابتاً وببساطة لن يزود باتساق موقع الطالب (Chase , 1999 , p. 72) .

يتضح مما سبق أن ثبات الاختبارات محكية المرجع يعني اتساق قرارات تصنيف الطلاب في إحدى فئات الإتقان وذلك من خلال القياسات المتكررة لنفس الاختبار أو من خلال تطبيق صور متكافئة له .

ويجدر بالذكر أن الثبات ليس من خصائص الاختبار وإنما هو خاصية من خواص درجات الاختبار يستعمل لصنع القرارات ، ويمكن أن يختلف تماماً باختلاف أهداف القياس ويختلف أيضاً في شمولية التعميم (Brennan , 1998 , p. 7) .

القيم المقبولة لمعامل الثبات :

تتراوح قيمة معامل الثبات بين الصفر والواحد والسؤال المطروح دائماً ما هي القيمة المقبولة لمعامل الثبات للثقة بنتائج الاختبار ، الحقيقة أنه لا توجد إجابة قاطعة لجميع الحالات إلا أنه يمكن الاعتماد على أهمية القرار الذي سيتخذ ، فالقرارات التربوية نادراً ما تكون مفردة بل تميل لأن تكون تتابعية تبدأ بأحكام واجراءات بسيطة وتمر بسلسلة أحكام أكثر تعقيداً ، في الخطوات الأولى من صنع القرار قد تكون قيم معاملات الثبات المنخفضة مقبولة إلى حد ما لأن نتائج الاختبار ستستخدم كموجِّه لجمع معلومات اضافية . ولكن المهم ألا تعامل الدرجات كما لو كانت مضبوطة / دقيقة تماماً عندما يكون معامل الثبات منخفضاً بل يجب عمل أحكام مؤقتة ومحاولة تثبيت البيانات لاحقاً وأن يكون هناك استعداد لعكس القرار عندما يتضح أنه خاطئ .

وهكذا يمكن الاعتماد على أهمية القرار فيما إذا كان بالامكان تأكيده أو عكسه في وقت لاحق وعلى النتائج المترتبة عنه ، فلقرارات هامة غير قابلة للتغيير يجب أن يكون معامل الثبات مرتفعاً ، في حين أنه لقرارات أقل شأناً ويمكن تغييرها لاحقاً يمكن قبول قياسات أقل ثباتاً ، بمعنى أن القيم المقبولة لمعامل الثبات تعتمد بشكل كبير على مقدار الثقة

التي نحتاجها في القرار الذي سيتم عمله ، فالثقة الأكبر تحتاج لمعامل ثبات أكبر (Gronlund , 1981 , p. 114) .

وتخلص الباحثة إلى أنه لا توجد قيمة محددة تناسب كافة الأغراض وعموماً كلما ارتفعت قيمة معامل الثبات كان أفضل لأن اعتماد اختبارات معاملات ثباتها منخفضة سيترك مجالاً للشك في نتائجها وسيعتبر ضياعاً للوقت والجهد .

هذا ويشير قاسم علي الصراف (٢٠٠٢ ، ص ١٩٨) إلى أنه يمكن زيادة قيم معاملات ثبات الاختبارات محكية المرجع عن طريق زيادة عدد المفردات التي تقيس الأهداف التي يغطيها الاختبار ومن خلال التأكد من أن الأهداف محدّدة بدقة ووضوح .

ضرورة استخدام طرق جديدة لحساب معامل ثبات الاختبارات محكية المرجع :

اعتمدت النظرية التقليدية للقياس على التباين واستخدمت العديد من الطرق لحساب معامل ثبات الاختبارات معيارية المرجع من أهمها :

- معامل الاستقرار
- معامل التكافؤ
- معامل الاتساق الداخلي

وقد استخدمت جميعها معاملات الارتباط معتمدة على أن معامل الثبات هو " معامل الارتباط بين الاختبار ونفسه " (عادل محمد محمود العدل ، ١٩٨٦ ، ص ٤٩) ، أي أنها اعتمدت على وجود التباين في درجات المختبرين ، وكان ذلك ملائماً لتلك الاختبارات التي تهدف إلى تعرُّف الفروق بين الطلاب . إلا أن الوضع أصبح مختلفاً مع الاختبارات محكية المرجع التي لم تصمَّم لنفس الهدف وإنما اهتمت بالدرجة الأولى بوصول المتعلم إلى حد الإتقان المطلوب واتخذت الأساليب الكفيلة بذلك ، فبدأ توزيع الدرجات يتحول من المنحنى الاعتدالي إلى منحنى ملتو نحو اليمين ، وبهذا قلَّ انتشار الدرجات وتجمعت معظمها عند قمة المقياس أي نقص التباين فيما بينهما ، وبالأحرى أصبح كما يقول رشدي فام منصور يبدأ بلا تعلم (لا تباين) وينتهي بتعلم متقن (لا تباين أيضاً) (١٩٨٧ ، ص ٢٢) .

وبالتالي لم تعد الطرق المعتمدة في حسابها على معامل الارتباط وسيلة مناسبة لقياس الثبات وهذا ما أشارت إليه جميع الكتابات في هذا المجال .

حيث يشير جرونلند (Grounlund , 1981 , p. 111) إلى أنه عند استخدام الاختبارات معيارية المرجع فإننا نريد أن يكون أداء الفرد متسقاً من :

١- مفردة لأخرى حيث أن جميع المفردات تقيس نفس مخرجات التعلم (اتساق داخلي) .

٢- من وقت لآخر حيث يتوقع أن يكون لمخرجات التعلم درجة معقولة من الاستقرار (استقرار) .

٣- من شكل اختبار لآخر حيث تميل الأشكال لقياس نفس العينة من مهام التعلم (تكافؤ) .

إلا أن الاختبارات محكية المرجع لم تصمَّم لتؤكد التباين بين الأفراد وبالتالي فإن معاملات الثبات التقليدية المعتمدة على الارتباط من المحتمل أن تزوِّد بنتائج مضللة عندما تستخدم مع الاختبارات محكية المرجع .

فالاستقرار هام للاختبار محكي المرجع ولكن ليس من الضروري أن يكون معامل الارتباط وسيلة لتحقيقه لأنه يعتمد على التباين ، فقد يكون الاختبار متسقاً بدرجة عالية ومع ذلك لا تعكس هذه الطريقة المعتمدة على التباين هذا الاتساق (نادية عبد السلام ، ١٩٨١ ، ص ١٨٩) .

كما أن تقديرات الاتساق الداخلي التي تعكس درجة الترابط بين العناصر وتبين تجانس الاستجابة تعتبر عملية زائدة عن الحاجة في المقاييس محكية المرجع لأن درجة تجانس الاستجابة تتحدد عند تحسين تعريفات المجال وإعداد عناصر الاختبار (جابر عبد الحميد جابر ، ١٩٩٦ ، ص ٢٦٠) .

وهكذا فإن المعنى الإحصائي المتضمن في الاختبارات محكية المرجع يؤدي إلى خفض تباين الدرجات بين الأفراد ، فإذا تابع كل شخص التدريب حتى اتقن المهارة المطلوبة سينخفض التباين إلى الصفر وكلما انخفض تباين العينة انخفض معامل الثبات . وبالتالي من غير الملائم تقويم ثبات هذه الاختبارات بتطبيق الطرق السابقة ، ففي هذه الظروف حتى الاستقرار والاتساق الداخلي المرتفع للاختبار يمكن أن يعطي معامل ثبات قريباً من الصفر (Anastasi , 1990 , pp. 137-138) .

ويتفق مع ما سبق رجاء أبو علام (١٩٨٧ ، ص ٢٩٠) حيث يؤكد على أن استخدام الطرق المناسبة لحساب معامل ثبات الاختبارات معيارية المرجع قد يؤدي إلى نتائج مضللة إذا استخدمت مع الاختبارات محكية المرجع التي تتصف بانخفاض انتشار درجاتها .

وخلاصة ما سبق أن الطرق المستخدمة لتحديد معامل ثبات الاختبارات معيارية المرجع غير ملاءمة لتحديد معامل ثبات الاختبارات محكية المرجع ، لأن معامل الارتباط غير مناسب لمقارنة مجموعة الدرجات قليلة التباين (Hambelton , 1994 , p. 25 ; Alken , 1997 , p. 93) .

وعلى الرغم من ذلك لا تزال الطرق الارتباطية هي الأساليب الشائعة في حساب معاملات ثبات الاختبارات المدرسية بغض النظر عن نوعها ، وأكثر من هذا أن بعض الدراسات العلمية الحديثة والمهتمة بالنوعية الجديدة من الاختبارات لا تزال تستعين بالطرق التقليدية في حساب معامل ثبات درجات اختباراتها ومن أمثلة هذه الدراسات :

– دراسة محمد فتح الله سيد (١٩٩٥) التي تمَّ فيها بناء اختبار محكي المرجع في مادة العلوم وحسب معامل ثباته باستخدام الطرق التالية (معامل الاتساق الداخلي ، معامل الاستقرار والتكافؤ ، ليفنجستون ، معامل كابا) أي أنه جمع بين النوعين من الطرق .

– دراسة إسماعيل الوليلي (١٩٩٦) التي تمَّ فيها بناء اختبار محكي المرجع ثم حسب معامل ثباته باستخدام معامل كابا ومعامل التكافؤ .

– دراسة مصطفى محمد كامل (١٩٩٩) التي تمَّ فيها أيضاً بناء اختبار محكي المرجع وحسب معامل ثباته باستخدام معامل كيودر ريتشاردسون (٢١) .

ومن هنا تؤيد الباحثة ضرورة استخدام الأساليب والتقنيات المناسبة لكل نوعية من الاختبارات .

الطرق المتبعة في حساب معامل ثبات الاختبارات محكية المرجع :

عرف القياس التربوي العديد من طرق حساب معامل ثبات الاختبارات محكية المرجع والتي جمعت في ثلاث فئات :

– دالة عتبة الفاقد Threshold loss function

– دالة فاقد مربع الخطأ Squared - error loss function

– تقدير مجال الدرجة Domain score estimation

وتمر عملية اختيار مؤشر الثبات المناسب بمرحلتين :

١– اختيار فئة الثبات الملاءمة بالاعتماد على اعتبارات تمهيدية محددة متصلة بالافتراضات ، التفسيرات ، واستخدام المؤشرات . هذه الاعتبارات تتضمن توضيح المصطلح ، افتراض شكل الاختبار ، وضع درجة القطع ، تفسير درجة الاختبار وصنع القرار .

٢- اختيار مؤشر محدد داخل الفئة المختارة (Berk , 1980 , p. 323) .

وعموماً يمكن تقسيم هذه الطرق إلى :

أ – طرق تتطلب تطبيق الاختبار مرتين أو تطبيق صور متكافئة للاختبار . وهذه الطرق هي التي سيتم التركيز عليها في هذه الدراسة .

ب- طرق تتطلب تطبيق الاختبار مرة واحدة : ومن هذه الطرق طريقة ليفنجستون ، طريقة هينا ، وطريقة صبكوفياك .

طرق حساب معامل ثبات الاختبارات محكية المرجع التي تتطلب تطبيق الاختبار مرتين أو تطبيق صور متكافئة من الاختبار :

١- طريقة كارفر (١٩٧٠) :

رفض كارفر (١٩٧٠) الطريقة الارتباطية المستخدمة في حساب معامل الثبات محاولاً أن يبرهن أن الثبات يعتمد على التطابق ولكن التطابق لا يعتمد على التباين . وقد اعتمدت اجراءات كارفر على تطابق التوزيعات في حين أن المفهوم العام للثبات في القياس التربوي يعتمد على تطابق درجات الأفراد (Hambelton et al. , 1978 , p. 20) .

وقد اقترح كارفر طريقتين لتحديد الاتساق وهما :

– الطريقة الأولى : تتطلب تطبيق اختبارين متكافئين على نفس المجموعة ثم تقارن النسب المئوية للطلاب الذين حددوا كمتقنين في كل اختبار ، فإذا كانت النسبتان متساويتين أومتساويتين تقريباً اعتبر الاختباران ثابتين ، ويلاحظ أنه ربما يكون الاختباران غير ثابتين تماماً ويعطيان نسباً متساوية للمتقنين . كمثال .. يمكن أن تحدد نصف المجموعة كمتقنين في الاختبار الأول في حين قد يحدد النصف الآخر من المجموعة كمتقنين في الاختبار الثاني في هذه الحالة النسبة (٥٠%) ولكن مجموعة المتقنين مختلفة تماماً.

– الطريقة الثانية : تتطلب تطبيق نفس الاختبار على مجموعتين قابلتين للمقارنة ، لكنها ستعاني من نفس القصور ، ولهذا فإن هذه الطريقة ليست حسَّاسة لاتساق تصنيفات الأفراد (Subkoviack , 1980 , p. 131) .

ولحساب معامل كارفر تنظّم نتائج الاختبار كما في الجدول التالي :

جدول (١) توزيع الطلاب في حالات الإتقان باستخدام طريقة كارفر

التطبيق الأول / التطبيق الثاني	متقن	غير متقن
متقن	a	b
غير متقن	d	c

$$\text{معامل كارفر} = \frac{a+c}{N}$$

حيث أن :

$$N = a + b + c + d$$

ويلاحظ أن هذا المعامل سهل الحساب لكنه يركِّز فقط على نسبة الأفراد المصنفين كمتقنين وغير متقنين في مرتي التطبيق دون الاهتمام باتساق قرارات التصنيف هذه .

٢- طريقة نسبة الاتفاق لهامبلتون ونوفيك (١٩٧٣) :

يقدم المعيار المقترح من قبل كارفر دليل ضعيف على ثبات الاختبار محكي المرجع حيث أن شروطه ضرورية ولكنها غير كافية لتقدير الثبات ، لذا اقترح هامبلتون ونوفيك (١٩٧٣) أن يعرَّف ثبات اتساق قرارات الإتقان في حدود اتساق القرارات الناتجة عن تطبيق نفس الاختبار مرتين أو تطبيق أشكال متكافئة من الاختبار .

فبفرض أن المختبرين قد صنفوا في حالات الإتقان (N) فإن موشر الثبات المقترح من قبل هامبلتون ونوفيك والذي يطلق عليه نسبة الاتفاق (P_o) هو :

$$P_o = \sum_{K=1}^{N} P_{KK}$$

حيث أن :

P_{kk} : هي نسبة المختبرين المصنفين في حالة الإتقان (k) في كلا التطبيقين ، وبالتالي فهو يمثل نسبة إتقان القرارات الملاحظة (Hambelton et al. , 1978 , pp. 20-21) .

ويمكن تعريف (P_o) على أنه نسبة الدرجات الواقعة تحت أو فوق مستوى المحك في كلا التطبيقين أو في الصورتين المتكافئتين (Alken , 1997 , p. 93) .

هذا ويقدّم المؤشر (P_o) نسبة الاتساق الكلي للتصنيف التي تحدث لأي سبب في كلا الاختبارين ، وهناك عاملان يسهمان في هذا الاتساق وهما العدد النسبي للمتقنين وغير

المتقنين في المجموعة المختبرة (تركيب مجموعة الإتقان / عدم الإتقان) ودقة أو إحكام الاختبار (Subkoviack , 1980 , p. 152) .

ويشير صلاح علام (٢٠٠١ ، ص ٢٩٩) إلى أنه " على الرغم من وضوح هذه الصيغة وسهولة تطبيقها إلا أنها لا تأخذ بعين الاعتبار اتساق التصنيف الذي ربما يرجع إلى عامل الصدفة مما يؤدي إلى نتائج غير دقيقة " .

ويلاحظ أن هذا المؤشر حساس لاختيار درجة القطع ، طول الاختبار ، وتباين الدرجة . وتعتبر درجة القطع المتغير الأكثر تأثيراً على قيمة نسبة الاتفاق من المتغيرين الآخرين ، حيث ترتفع قيمته مترافقة مع درجات قطع عند أطراف توزيع الدرجات وتنتج أقل قيمه مع درجات قطع قرب المتوسط . كما ترتفع قيمته عندما يزداد عدد مفردات الاختبار ويزيد تباين الدرجة ومن الممكن الوصول إلى القيمة (٠,٧٥) أو أعلى مع اختبارات فرعية تحتوي على (١٠) مفردات على الأقل ولها تباين درجات قليل نسبياً . هذه الخصائص هامة عند بناء واستخدام اختبارات محكية المرجع من إعداد المعلم ، وتشير محاسن (P_o) لماذا كونه المفضّل لاختبارات معدة لاتخاذ قرارات على مستوى الفصل (Berk , 1980 , pp. 327-332) .

وقد أظهرت دراسة بيرك (Berk , 1980 , pp. 333-334) التي قارنت طريقة نسبة الاتفاق (P_o) ، معامل كابا (K) ، مارشال وهيرتل ، صبكوفياك ، هينا (١٩٧٦) ، وهينا (١٩٧٧) أن (P_o) هو التقدير الوحيد غير المتحيز وأنه أسهل طريقة للفهم والحساب والتفسير ولكن من مساوئ هذه الطريقة أن لها أعلى خطأ معياري يتراوح بين (٦ – ٨%) لعينات بحجم الفصل الدراسي . وعلى المعلمين تفسير قيم المؤشر بحذر ، حيث ينقص الخطأ عندما ترتفع درجة القطع في اختبارات فرعية قصيرة ويصل إلى (٦%) عند اتخاذ درجات قطع (٨٠%) فما فوق .

وإذا كان من الصعب على المعلمين بناء اختبارين متكافئين لمجموعة أهداف يوصى بحساب المؤشر من إعادة تطبيق صورة واحدة للاختبار . لكن الأخذ بعين الاعتبار التدريس الذي سيؤثر على التطبيق الثاني لأن المعارف التي اكتسبها الطالب بعد التطبيق الأول ستنتج تضخماً زائفاً في درجات الاختبار الثاني فقد تتغير حالة بعض الطلاب غير المتقنين في التطبيق الأول إلى متقنين في التطبيق الثاني ، وفي هذه الحالة ستقل قيمة (P_o) زيفاً وبالتالي يجب أخذ هذا التحيز المحافظ بعين الاعتبار في تفسير المؤشر (P_o) .

كما توصلت دراسة آمال محروس (١٩٨٨ ، ص ٢٠٣) التي قارنت الطرق : نسبة الاتفاق ، ليفنجستون ، هارس ، وصبكوفياك إلى تفوق طريقة نسبة الاتفاق على الطرق الثلاث الأخرى التي تعتمد على تطبيق الاختبار مرة واحدة . وعلى الرغم من تفوقها لا يشيع استخدامها لأنها تعطي نتائج مرتبطة بمدى تطابق صورتي الاختبار ، فإذا كانت الصورتان أكثر تطابقاً أعطت ثباتاً أعلى وإذا كانت الصورتان أقل تطابقاً أعطت ثباتاً منخفضاً .

وفي دراسة كالون (Kalohn , 1992) لتعيين مميزات مؤشرات الثبات : نسبة الاتفاق ، كابا ، كابا المصححة ، وفاي وجد أن نسبة الاتفاق وكابا المصححة قدَّمتا تقديرات غير متحيزة في قيم البارامترات في كافة الظروف (شكل توزيع الدرجات ، حجم العينة ، معاملات الثبات التقليدية ، ودرجة القطع) .

٣- طريقة معامل كابا لسوامنيثان وهامبلتون والجينا (١٩٧٤) :

اقترح سوامنيثان وهامبلتون والجينا أن معامل نسبة الاتفاق المقترح من قبل هامبلتون ونوفيك (١٩٧٣) لا يأخذ بعين الاعتبار نسبة الاتفاق المتوقع أن تحدث بالصدفة وحدها ، وبهذا لا يبدو ملائماً بكل معنى الكلمة (تماماً) لقياس ثبات الاختبار محكي المرجع الذي يستخدم لتعيين المختبرين في حالات إتقان الأهداف التي يغطيها الاختبار .

ووفقاً للهدف من هذه الاختبارات يبدو معقولاً اعتبار اتساق القرارات حول حالات الإتقان الناتجة عن إعادة تطبيق الاختبار كقياس للثبات ، بمعنى أن عملية اتساق القرار هذه هي انعكاس للمحتوى واستخدام الاختبار لصنع القرارات ولهذا يعرَّف ثبات الاختبار محكي المرجع على أنه قياس الاتفاق بين القرارات الناتجة عن إعادة تطبيق الاختبار وكبديل عن معامل الاتفاق اقترح استخدام معامل كابا (K) الذي وضعه كوهين (Kohen , 1960) والذي يأخذ بعين الاعتبار نسبة الاتفاق الناتجة بالصدفة وبهذا يبدو أكثر ملاءمة (Swamianathan et al. , 1974 , pp. 263-264) .

أي أن هذا المعامل يعبّر عن درجة اتساق تصنيفات الطلاب في مرتي تطبيق الاختبار بمعنى أن يصنَّف الطالب المتقن في التطبيق الأول على أنه متقن في التطبيق الثاني وأن يصنَّف الطالب غير المتقن في التطبيق الأول على أنه غير متقن في التطبيق الثاني .

فبعد أن يتم تطبيق الاختبار مرتين يصنّف الطلاب إلى حالات إتقان عند درجة قطع محددة وترتب النتائج في جدول كالتالي :

جدول(٢) توزيع الطلاب في حالات الإتقان باستخدام طريقة معامل كابا

التطبيق الأول / التطبيق الثاني	متقن	غيرمتقن	المجموع
متقن	م١١	م٢١	م٠١
غيرمتقن	م١٢	م٢٢	م٠٢
المجموع	م١٠	م٢٠	م٠٠

حيث أن :

م١١ : عدد الطلاب الذين أتقنوا الاختبار في مرتي التطبيق .

م٢٢ : عدد الطلاب الذين لم يتقنوا الاختبار في مرتي التطبيق .

م١٢، م٢١ : عدد الطلاب الذين أتقنوا الاختبار في التطبيق الأول ولم يتقنوا الاختبار في التطبيق الثاني وبالعكس .

م٠١ : عدد الطلاب الذين أتقنوا في الاختبار الأول .

م٠٢: عدد الطلاب الذين لم يتقنوا في الاختبار الأول .

م١٠ : عدد الطلاب الذين أتقنوا في الاختبار الثاني .

م٢٠ : عدد الطلاب الذين لم يتقنوا في الاختبار الثاني .

م٠٠ : عدد الطلاب الكامل .

ثم يحسب معامل كابا (K) من العلاقة :

$$K = \frac{P_o - P_c}{1 - P_c}$$

حيث أن :

$$Pc = \sum_{i=1}^{N} P_{0i} \cdot P_{i0}$$

$$Po = \sum_{i=1}^{N} P_{ii}$$

P_{i0} : نسبة الطلاب المصنفين في حالة الإتقان (i) في التطبيق الأول للاختبار .

P_{0i} : نسبة الطلاب المصنفين في حالة الإتقان (i) في التطبيق الثاني للاختبار .

P_{ii} : نسبة الطلاب المصنفين في حالة الإتقان (i) في مرتي التطبيق .

P_o : نسبة الاتفاق الملاحظة .

P_c : نسبة الاتفاق المتوقعة .

K : نسبة الاتفاق الموجودة وهي مبنية على النسب الملاحظة والمتوقعة على القطر الرئيسي لمصفوفة النسب المشتركة ولا تتأثر بالتعارضات الموجودة خارج القطر الداخلي (Swamianathan et al. , 1974 , p. 264) .

وقد برهن هينا (Huynh , 1976 , p. 253) على أن القيمة الدنيا لمعامل كابا تتحقق عندما لا تسهم معلومات الاختبار في عملية صنع القرار ، ومن ناحية أخرى يصل الحد الأعلى له عندما تنتج بيانات متكافئة نفس التصنيفات تماماً .

ومن أهم خصائص هذا المؤشر ما يلي :

١- تصحيح نسبة الاتفاق من عامل الصدفة مقيد بالتكرارات الهامشية لجدول البيانات ٢×٢ فهذا المؤشر يمكن أن يكون في أعلى قيمة (+١) فقط عندما تتساوى القيم الهامشية في كلتا الصورتين أو القياسين .

٢- تزيد قيم (K) مع زيادة طول الاختبار وهذا ينطبق على معظم معاملات الثبات الأخرى ، حيث تتكون عينة المفردة أو الاختبار الفرعي المصمَّم لتقييم إتقان هدف تعليمي عموماً من (٣- ١٠) مفردات لاختبارات طوِّرت لتستخدم على مستوى الفصل الدراسي ، ونادراً أكثر من (٢٠) مفردة لاختبارات طوِّرت لتستخدم على مستوى الولاية أو المقاطعة ، ولهذا يتوقع أن تزود اختبارات محدودة الطول بمعلومات أقل ثباتاً وأن تقدم مؤشرات أقل قيماً .

٣- تزداد قيم (K) مع زيادة تباين الدرجة . فكلما قلَّ تباين الدرجات سيؤدي إلى حصر في مجال (K) في الاختبارات الفرعية القصيرة .

هذا وإن أعلى قيم (K) تعود لدرجات قطع قرب المتوسط ، في حين تترافق أقل القيم له مع درجات قطع عند الأطراف (Berk , 1980 , pp. 332-333) .

وقد أظهرت دراسة كالون (Kalohn , 1992) لتعيين مميزات مؤشرات الثبات محكية المرجع : نسبة الاتفاق ، كابا ، كابا المصححة ، وفاي مع دراسة تأثيرات شكل توزيع الدرجات وحجم العينة ومعاملات الثبات التقليدية وموضع درجة القطع أن لموضع درجة القطع وشكل التوزيع التأثير الأكبر في التحيز لتقديرات كابا وفاي ، فقد زوِّدت درجات القطع الموجودة قرب منتصف التوزيع بتقديرات غير متحيزة لكلا المعاملين كابا وفاي ، وكلما اقتربت درجة القطع من نهاية التوزيع وكانت كثافة الدرجات قليلة زاد التحيز . كما أظهرت الزيادة في حجم العينة نقص تحيز معنوي خصوصاً عندما وجدت درجة القطع عند نهاية التوزيع ، وربما انخفض تأثير حجم العينة عندما وجدت درجة القطع عند منتصف التوزيع .

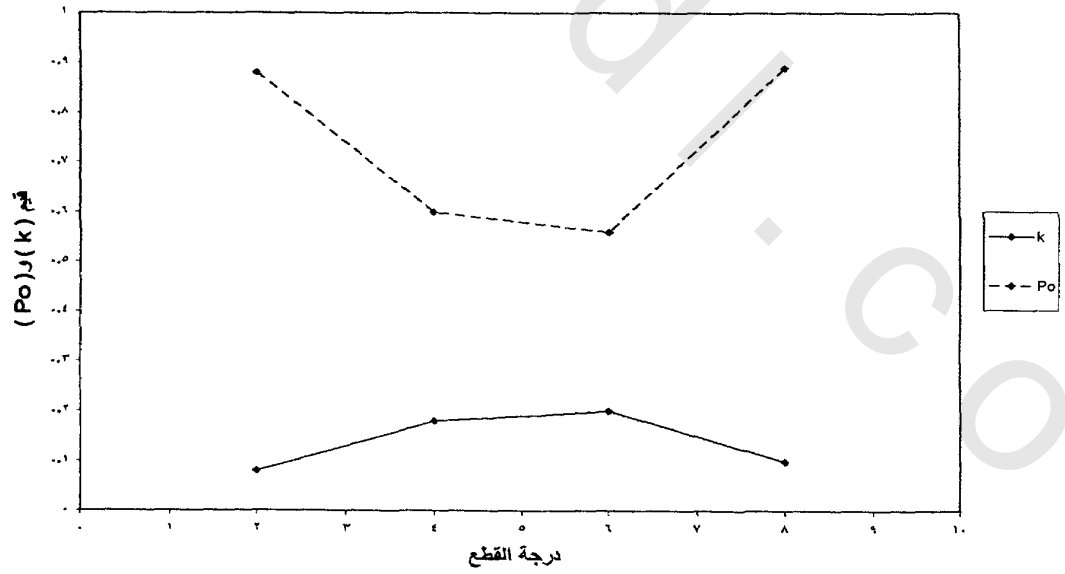
كما أظهرت دراسة سعاد حسانين (٢٠٠٠) التي قارنت معاملات الثبات : هارس ، معامل كابا ، وليفنجستون ، واستخدمت أنواع المفردات الاختبار من متعدد ، أسئلة الصواب والخطأ ، أسئلة المقال . أنه لم يختلف معامل ثبات الاختبار محكي المرجع باختلاف نوع المفردة عند حساب الثبات بطريقة معامل كابا .

المقارنة بين (P_o) و (K) :

١- المؤشران (P_o) ، (K) حسّاسان لأنواع مختلفة من الاتساق ، والسبب الرئيسي في اختيار أحدهما بدل الآخر يأتي من الرغبة في قياس نوع الاتساق الخاص بذلك المؤشر ، أي قياس الاتساق الكلي أو اتساق الاختبار فقط (Subkoviak , 1980 , pp. 152-153) .

٢- (P_o) و (K) غير متكافئين إحصائياً ، حيث أن (P_o) لا يأخذ بعين الاعتبار وصول الطالب فئة النجاح / الفشل بالصدفة ، في حين يعتبر عامل الصدفة في حساب (K) ولهذا السبب يكون أحياناً أقل من (P_o) (Chase , 1999 , p. 72) .

٣- حساسية (P_o) لدرجة القطع هي تماماً عكس (K) ، حيث أن أعلى القيم لـ (P_o) تترافق مع درجات قطع عند أطراف التوزيع وأقل القيم له تترافق مع درجات قطع قرب درجة المتوسط . وبالعكس تنتج أعلى القيم لـ (K) مع درجات قطع قرب درجة المتوسط وأقل القيم تترافق مع درجات قطع عند أطراف التوزيع . ويمكن توضيح العلاقة بين المؤشرين (P_o) ، (K) ودرجة القطع بالشكل البياني التالي :



شكل (١) العلاقة بين المؤشرين (P_o) و (K) ودرجة القطع

(Subkoviack , 1980 , pp. 152-153) .

٤- يناسب المؤشر (P_o) اختبارات محكية المرجع عند اختيار درجة قطع مطلقة ولاختبارات قد تحتوي اختبارات فرعية قصيرة و/ أو تزود بتباين درجة منخفض . بينما يناسب المؤشر (K) الحالة التي تكون درجة القطع فيها نسبية وذلك بالنسبة لنتائج اجتياز أو فشل نسبة مخصصة من الطلاب ، لكن المشاكل المتصاحبة مع (K) تجعلها أقل فائدة من (P_o) ويجب توخي الحذر في استخدامها وتفسيرها . وبسبب اعتماد المؤشرين على درجة القطع ، طول الاختبار ، تباين الدرجة ، وتكرار الخلايا في جدول احتمالي ٢×٢ يوصى مصمِّم الاختبار بتسجيل هذه المعلومات كافة مع مؤشر الاتفاق لتسهيل تفسيرات المؤشر (Berk , 1980 , p. 333) .

٥- يلاحظ أن الخطأ المعياري لـ (P_o) أقل أحياناً من الخطأ المعياري لـ (K) ، ولحجم عينة ثابتة مثلاً (٣٠) طالباً يمكن أن يقدر (P_o) بدقة أكثر من (K) عموماً (Subkoviak , 1980 , p. 154) .

العوامل المؤثرة في حساب معامل ثبات الاختبار محكي المرجع :

تختلف العوامل المؤثرة في حساب معامل ثبات الاختبارات محكية المرجع عن العوامل المؤثرة في حساب معامل ثبات الاختبارات معيارية المرجع نظراً لاختلافهما في الهدف والتصميم كما تم بحثه سابقاً . حيث أن العوامل المؤثرة في حساب معاملات ثبات الاختبارات محكية المرجع والتي وردت في تراث القياس التربوي وتناولتها الدراسات السابقة في هذا المجال هي : طول الاختبار ، درجة القطع ، حجم العينة (عدد المختبرين) ، قدرة الطالب ، خواص محتوى الاختبار ، شكل توزيع الدرجات ، طريقة اختيار المفردة ، طريقة تقدير الدرجات (التقليدية ، أسلوب الدرجات الذاتية ، وزن الثقة (احتمال مقترح) ، وزن الثقة (لوغارتيمية) ، توزيع الثقة الكروية) ، التغذية المرتدة (كلية ، جزئية ، مؤجلة) ، نوع المفردة (الاختيار من متعدد ، الصواب والخطأ ، المقال) ، ترتيب المفردات .

وسيتم في هذه الدراسة التركيز على أكثر هذه العوامل أهمية في حساب معامل ثبات الاختبار محكي المرجع ألا وهي :

١- طول الاختبار Test length

٢- درجة القطع Cut-off score

٣- حجم العينة Sample size

١- طول الاختبار :

أحد العوامل التي تؤثر في حساب معامل ثبات الاختبار محكي المرجع هو طول الاختبار والذي يحدد عادة بعدد الأهداف التي يقيسها الاختبار وعدد المفردات التي تقيس كل هدف ، فكلما زاد طول الاختبار ارتفعت قيمة معامل الثبات ، وقد أكدت الدراسات السابقة في هذا المجال ذلك ومن هذه الدراسات ما يلي :

بيّنت دراسة هاجين (Hagen , 1983) التي قامت على مقارنة اجراءات تقدير طريقة هينا وطريقة صبكوفياك لمؤشرات اتساق التصنيف (P_o) و (K) وعالجت أطوال الاختبار (٢٥ ، ٥٠ ، ٧٥) مفردة ، أن طول الاختبار أثّر في كافة التقديرات حيث ظهرت زيادة في أهمية التقديرات مع زيادة طول الاختبار .

كما أشارت دراسة وانغ (Wang , 1983) التي قامت على مقارنة طرق حساب معامل الثبات معامل كابا ، هينا ، صبكوفياك ، ومارشال وهيرتل وذلك للتحقق من تأثيرات عدد المفردات ، عدد المختبرين ، ودرجات القطع على دقة التقديرات المستمدة من الطرق الأربع وذلك باستخدام عدد المفردات (٥ ، ١٠ ، ٢٠) مفردة . أن متوسط التقديرات يرتفع مع زيادة طول الاختبار وتغير الدرجة الحقيقية في معظم الظروف المدروسة .

وبما أن الاختبارات محكية المرجع تستخدم لتصنيف المختبرين في حالات إتقان محددة فإن ارتفاع قيمة معامل الثبات يعني زيادة اتساق قرارات التصنيف وبالتالي زيادة الثقة في نتائج الاختبار .

أي أن مشكلة تحديد طول الاختبار ترتبط بعدد أخطاء التصنيف المسموح التجاوز عنها ، وإحدى طرق تقليل احتمالات خطأ سوء التصنيف هي جعل الاختبار طويلاً جداً وربما يكون هذا غير مناسب دائماً (Hambelton et al. , 1978 , p. 24) .

ويشير جرونلند (Gronlund , 1981 , pp. 112-113) إلى أنه عند استخدام الاختبارات محكية المرجع في التعليم الصفي فإننا نستطيع زيادة احتمالية صدق وثبات النتائج باستخدام عينة كبيرة بشكلٍ كافٍ من مفردات الاختبار لكل هدف تعليمي أو مجال مخصص من مهام التعلم الواجب قياسها . فإذا كانت المخرجات المقصودة كثيرة التخصيص وعالية التركيب (مثل جمع عددين بمرتبة واحدة) فعدد قليل نسبياً من المفردات (وليكن خمس) ربما يكون كافياً لاعتماد قرار فيما يتعلق بالإتقان .

ولمعظم قرارات الإتقان / عدم الإتقان قد تزود عشر مفردات لكل مجال منفصل من المهام بحد أدنى معقول وإذا كانت القرارات التعليمية مبنية على أقل من عشر مفردات فسيتم عمل قرارات مؤقتة فقط مع محاولة الاستعانة ببيانات أخرى متاحة .

بالإضافة إلى ما سبق توجد عدد من العوامل الهامة يجب أخذها بعين الاعتبار عند تحديد عدد المفردات الواجب بناؤها لقياس كل هدف من أهداف الاختبار ، ويمكن توضيح هذه العوامل من خلال العرض التالي :

- عرّف بيرك ورونالد (Berk & Ronald , 1979) أربعة عوامل أساسية تساعد على تحديد عدد المفردات الواجب بناؤها أو تعيينها لمجموعة أهداف مختبرة ، وهذه العوامل هي:

١- أهمية ونوع القرارات التي ستصنع من النتائج .

٢- أهمية الأهداف السلوكية والتعليمية .

٣- عدد الأهداف .

٤- بعض القيود العملية مثل وقت كتابة المفردة ، الوقت المتاح لتطبيق الاختبار ، خصائص الطالب وذلك في سياق ربط نتائج البحث بدرجات القطع ومعاملات الثبات .

وقد تم التوصية على أنه يجب أن تعدّ ما بين (٥ – ١٠) مفردات لقياس كل هدف لمعظم القرارات على مستوى الفصل الدراسي ، وما بين (١٠ – ٢٠) مفردة يجب أن تستخدم لقرارات على مستوى المدرسة ، المقاطعة .

- أشار بيرك (Berk , 1988 , p. 366) إلى أن تحديد عدد المفردات التي يجب كتابتها لكل هدف في اختبار محكي المرجع يتم من خلال منظورين ، منظور عملي والآخر إحصائي .

يأخذ المنظور العملي عدة عوامل بعين الاعتبار وهي نوع وأهمية صنع القرار ، أهمية وتشديد تعيين الأهداف وعدد الأهداف . في حين يركّز المنظور الإحصائي على العلاقة بين عدد المفردات وصدق وثبات مؤشرات القرار .

- أشار نورمان (Norman , 1984) إلى أن تحديد العلاقة بين طول الاختبار واتساق قرارات تصنيف الاختبارات محكية المرجع يعتمد على مجموعة من العوامل الاضافية كتجانس الاختبار ، توزيع مجال الدرجات ، ومستويات صعوبة المفردة . ومن الصعب ايجاد قاعدة عامة فيما يتعلق بأطوال الاختبار المعقولة بدون أخذ هذه العوامل بعين الاعتبار.

وقد أشارت نتائج هذه الدراسة إلى أن الأهداف المبينة على (٢٠) مفـردة لـم تـزوّد بتحسينات عملية ومعنوية في اتساق القرار أكثر من الأهداف المبينة على (١٠) مفـردات ، كما أظهرت فئة الأهداف المبنية على مفردة من نوع الاختيار من متعدد (ذات أربعة بنود) تحسينات معتدلة ومعنوية في اتساق القرار أكثر من الأهداف المبينة على مفـردة محـددة ، ولم تظهر الأهداف المبينة على (١٠) مفردات تحسينات اتساق قرار معنوية أكثر مـن فئـة الأهداف المبينة على مفردة من نوع الاختيار من متعدد ذات أربعة بنود.

ومن ناحية أخرى يرى بيرلينر إلى أنه تصلح مفردة واحدة مرتبطة بموضوع ما أن تكون اختباراً نقياً أو كاملاً (نادية عبد السلام ، ١٩٨١ ، ص ١٨٨) ، ويشير أحمد جاسم السباعي ونجاح محمد النعيمي (٢٠٠١ ، ص ١٠٠) إلى أنه " لابد من أن يخصص بند على الأقل لكل هدف من الأهداف وذلك تحقيقاً لمصداقية الاختبار وموضوعيته " وهذا يعني أنه يمكن قياس الهدف بمفردة واحدة . كما يشير بابام Popham (١٩٧٨) إلى أنه توجد اختبارات محكية المرجع تتكون من مفردة واحدة أو مفردتين لكل هدف سلوكي مقاس وليس هناك شك في أن مفردة واحدة للاختيار من متعدد يمكن أن تعطي تقديراً ثابتاً لوضع الممتحن فيما يتصل بمستوى أدائه الحقيقي (محمود إبراهيم ، ١٩٩٠ ، ص ٦١) .
وتتفق مع الرأي السابق نتائج دراسة نورمان (Norman , 1984) التي تم الإشارة إليها في أن الأهداف المبينة على (١٠) مفردات لم تظهر تحسينات اتساق قرار معنوية أكثر من فئـة الأهداف المبينة على مفردة من نوع الاختيار من متعدد ذات أربعة بنود .

يتضح من العرض السابق أنه يوجد عدد من العوامل المؤثرة في عملية اتخاذ القـرار بشأن طول الاختبار إلا أنه لا توجد قاعدة عامة تنفع لجميع الأغراض وتحدّد عدد المفردات الواجب استخدامها لقياس كل هدف ، فهناك من يرى أن استخدام خمس مفـردات يناسـب قياس المخرجات عالية التركيب في حين يلزم لعمل قرارات الإتقان / عدم الإتقـان عشـر مفردات كحد أدنى ، وهناك من يؤيّد استخدام ما بين (٥ – ١٠) مفـردات لقيـاس أهـداف تستخدم في صنع قرارات على مستوى الفصل الدراسي ، في حين تشير بعـض الدراسـات إلى أن مفردة واحدة من نوع الاختيار من متعدد يمكـن أن تعطـي تقـديراً ثابتـاً لوضـع الممتحن.

وتخلص الباحثة إلى القول بأن هناك آراء ترشد وتوجه في عملية تحديد طول الاختبار ولكن لايوجد رأي قاطع ومتفق عليه بهذا الشأن مما يترك الباب مفتوحاً لعمل مزيد من الدراسات والأبحاث في هذا المجال .

٢- درجة القطع :

تعرّف درجة القطع على أنها الدرجة التي تفصل بين الطلاب الناجحين والراسبين في اختبار ما ، وتحدّد نسبة الطلاب القادرين على تحقيق الحد الأدنى المقبول كشرط للإتقان أو النجاح ، وهي " تعبر عن الأداء على الاختبار الذي يجب أن يحصل عليه المختبر ليظهر أنه قد أحرز كفاءة كافية في المهارات والمعرفة وليكون قادراً على أداء ألوان السلوك المحكية Criterion Behaviors " (فؤاد أبو حطب وآخرون ، ١٩٩٧ ، ص ٤١٥) .

وقد ظهرت في الكتابات المبكرة لجليزر بعض الاشارات بأن الاختبارات محكية المرجع يمكن أن تستخدم لتعيين درجات القطع بين الكفاءة وعدم الكفاءة أو فيما دعيت بعد ذلك نجاح / فشل وإتقان / عدم إتقان . وكان هناك افتراض بوجود " اتصال معرفي يتراوح بين عدم وجود كفاءة على الإطلاق إلى كفاءة تامة " ، وقد أشار جليزر (١٩٦٣) إلى علم درجات القطع حين كتب " نحتاج تخصيص حد أدنى من مستويات الأداء تصف أقل كمية من الكفاءة يتوقع أن يحرزها الطالب عند نهاية المقرر ، أو التي يحتاجها لكي ينتقل إلى الخطوة التالية من المنهج " .

وفي نفس الوقت الذي طوّر فيه جليزر معتقداته عن القياس محكي المرجع نشر ماجر (١٩٦٢) شرحاً عن الأهداف السلوكية وإعدادها حيث كتب " إذا استطعنا تحديد حد أدنى للأداء المقبول لكل هدف سيكون لدينا محك أداء لاختبار برامجنا التعليمية وسيكون لدينا وسائل لتحديد فيما إذا كانت برامجنا قد نجحت في إحراز المقصد التعليمي . ما الذي يجب أن نحاول عمله عندئذ أنشير في صياغة أهدافنا ما سيكون عليه الأداء المقبول بإضافة كلمات تصف محك النجاح " وهكذا أضاف ماجر فكرة محك الأداء .

ثم استخدم بابام وهيسيك (١٩٦٩) في دراسة لهما عن الاختبارات محكية المرجع مصطلح ماجر " محك أداء " حيث أشارا إلى أن الاختبارات محكية المرجع هي تلك الاختبارات التي تستخدم لتحديد حالات الفرد بناء على محك أداء وهكذا توالت الاضافات والتعريف بهذا المفهوم (Glass , 1978 , pp. 240-241) .

وقد تعددت المسميات التي أطلقت على درجة القطع Cutoff score فأحياناً يطلق عليها مستويات الإتقان Mastery level أو درجات النجاح Passing score أو الحد الأدنى للكفاية Minimum competency level أو مستويات المحك Criterion levels وأحياناً يطلق عليها درجات القطع Cutoff score (صلاح علام ، ٢٠٠١ ، ص ٢٥٣) .

هذا وتحدّد درجة القطع في مرحلة اعداد الاختبار وقبل تطبيقه حيث يحدد مستوى الأداء المطلوب أن يحققه الطالب لكي يعتبر متقناً ، ويمكن أن يكون مستوى الأداء هذا هدفاً اجرائياً أو نسبة مئوية من الإجابات الصحيحة . وهكذا تفصل درجات القطع بين مجموعة المختبرين وتقسمها إلى قسمين :

– القسم الأول وهو مجموعة المختبرين الذين وصلوا درجة المحك أو تجاوزوها وهي فئة المتقنين / الناجحين والتي سيسمح لها بالانتقال إلى الوحدة التالية أو المرحلة الأعلى .

– أما القسم الثاني فهو مجموعة المختبرين الذين لم يبلغوا درجة المحك وهي فئة غير المتقنين / الراسبين والتي يتوجب عليها إعادة التعليم حتى تصل إلى مستوى الإتقان المطلوب .

وبما أن عملية تعيين درجة القطع تسبق تطبيق الاختبار وبالتالي فهي لا تتأثر بحجم عينة المختبرين أو مستوى أدائها ولا تتغير من عينة لأخرى . وإذا دعت الضرورة لتغيير درجة القطع فإن ذلك سيؤدي إلى إدخال بعض الأفراد إلى فئة المتقنين أو استبعادهم منها ، ويكون ذلك عموماً مع الأفراد الذين تقترب درجاتهم من درجة المحك سواء بالزيادة أو النقصان ، أما الأفراد الذين تجاوزوها بقدر معقول فإن موقعهم في فئة المتقنين لايتغير وكذلك الأفراد الذين تقع درجاتهم تحت درجة المحك بمدى كبير فإنه لا توجد إمكانية لدخولهم فئة المتقنين مهما انخفضت درجة القطع .

أثر موضع درجة القطع على قيمة معامل الثبات :

جميع معاملات ثبات الاختبارات محكية المرجع حسّاسة لموضع درجة القطع وبالتالي فإن تفسير أي مؤشر يجب أن يتضمن تعيين درجة القطع ولا أهمية لحسابه بدون تبرير اختيارها لأن القيمة المرتفعة لمؤشر ثبات مبني على درجة قطع ليس لها تبرير أو لها تبرير ضعيف سيشير إلى أن الاختبار يستطيع تصنيف الطلاب باتساق في مجموعات خاطئة ، فاتساق صنع القرار بدون دقته له قيمة مشكوك فيها في التقويم محكي المرجع (صلاح علام ، ٢٠٠٠ ، ص ص ٢٨٨-٢٨٩ ؛ Berk , 1980 , p. 325) .

وقد أكدت الدراسات التي تمت في هذا المجال تأثير موضع درجة القطع على قيمة معامل ثبات الاختبارات محكية المرجع ، ففي دراسة وانغ (Wang , 1983) للتحقق من دقة أربع طرق لحساب مؤشر الثبات وهي معامل كابا ، هينا ، صكبوفياك ، ومارشال وهيرتل في ظروف تغيير عدد المفردات ، عدد المختبرين ، ودرجات القطع تبين تأثر

مؤشرات الثبات بموضع درجة القطع وأن تأثير درجة القطع على متوسط التقديرات يتغير تابعاً لتغيرات الدرجة الحقيقية والطريقة .

كما أشارت دراسة هاجين (Hagen , 1983) التي قامت على مقارنة اجراءات تقدير هينا (١٩٧٦) وصبكوفياك (١٩٧٦) لمؤشرات اتساق التصنيف (P_o) و(K) في ظروف تغير الاختبار ، شكل التوزيع ، ودرجات القطع (٧٠%،٨٠%،٩٠%) إلى أن شكل التوزيع يتبع للقرب من درجة القطع ومتوسط التوزيع فقد كان (P_o) في أقل قيمة و(K) في أعلى قيمة عندما اقتربت درجة القطع من المتوسط .

كما أظهرت دراسة كالون (Kalohn , 1992) لتعيين مميزات مؤشرات الثبات والتي تم الإشارة إليها سابقاً والتي اختبرت تأثير شكل التوزيع ، حجم العينة ، معاملات الثبات التقليدية ، وموضع درجة القطع – حيث استخدمت خمس درجات قطع – تأثير قيم البارامترات في جميع الإحصاءات بموضع درجة القطع وكان لموضع درجة القطع وشكل التوزيع التأثير الأكبر في التحيز لتقديرات كابا وفاي فقد زودت درجات القطع الموجودة قرب منتصف التوزيع بتقديرات غير متحيزة لكلا المؤشرين كابا وفاي . وكلما اقتربت درجة القطع من نهاية التوزيع وكانت كثافة الدرجات قليلة زاد التحيز ، كما أظهرت الزيادة في حجم العينة نقص تحيز معنوي خصوصاً عندما كانت درجة القطع عند أطراف التوزيع وربما انخفض تأثير حجم العينة عندما وجدت درجة القطع قرب منتصف التوزيع .

وفي دراسة لي (Lee , 1996) لمعرفة حساسية مؤشر ثبات الاختبار محكي المرجع لموضع درجة القطع والتي تم فيها تغير درجة القطع لعشر مواضع مختلفة اعتباراً من منتصف التوزيع وباتجاه الأطراف في كل نوع توزيع ظهر تأثر مؤشر ثبات كل اختبار بموضع درجة القطع على الرغم من نشوء علاقة عكسية لما تم افتراضه فقد ازداد المعامل كابا كلما اقتربت درجة القطع من منتصف المنحنى .

وهكذا نجد إجماع جميع الدراسات على تأثر مؤشر ثبات الاختبارات محكية المرجع بموضع درجة القطع ، مما يؤكد على ضرورة تعيينها عند حساب مؤشر الثبات .

طرق تحديد درجة القطع :

تعدّدت طرق تحديد درجة القطع فقد صنّفها هامبلتون في ثلاث مجموعات كالتالي :

أ- طرق التحكيم Judgmental Methods : تتطلب بيانات يجب جمعها من المحكمين لاعداد المعايير أو تتطلب صنع أحكام عن سيمة المتغيرات (مثلاً ، التخمين) .

ب- طرق أمبيريقية Empirical Methods : تتطلب جمع بيانات من اختبارات لتساعد في عملية اعداد المعايير .

ج- طرق مركبة Combination Methods : تستخدم بيانات التحكيم وبيانات أمبيريقية في عملية إعداد المعيار (Hambelton , 1980 , pp. 104-106) .

كما صنّفها صلاح علام (٢٠٠١ ، ص ص ٢٥٤-٢٥٦) بالشكل التالي :

أ- طرق تعتمد على التحكيم : تعتمد على أحكام الخبراء سواء أكانت أحكام فردية أو جماعية وتضم عدة طرق أهمها طريقة ندلسكاي - طريقة أنجوف - طريقة ايبل - طريقة جيجر .

ب- طرق تعتمد جزئياً على التحكيم وتسترشد بيانات أمبيريقية : حيث تعتمد على أحكام الخبراء مع وجود بيانات أمبيريقية يسترشد بها هؤلاء الخبراء في إصدار أحكامهم وتضم عدة طرق أهمها طريقة التحكيم المعزّزة بالمعلومات - طريقة انجوف المعدّلة - طريقة توفّق بين الطرق المطلقة والطرق النسبية.

ج- طرق تعتمد على البيانات الأمبيريقية وتسترشد بالتحكيم : تعتمد على بيانات أمبيريقية مستمدة من أداء فعلي للمختبرين وتعرف محكات تصنيف الأفراد بالاعتماد على المحكمين وتضم عدة طرق أهمها طريقة المجموعات المحكية - طريقة المجموعات الحدية - طريقة المجموعات المتناقضة .

وقد بني بعض من الطرق السابقة على الاعتبارات التالية :

١- محتوى المفردة .

٢- تخمين وتعيين المفردة .

٣- بيانات أمبيريقية من مجموعات إتقان وعدم إتقان .

٤- اجراءات قرار نظرية .

٥- قياسات محك خارجية .

٦- نتائج تربوية (Hambelton , 1980 , p. 101) .

والسؤال الذي يطرح نفسه ، كيف سنختار الطريقة المناسبة من بين الطرق الكثيرة المتوفرة لتحديد درجة القطع ؟

لقد وضع هامبلتون (Hambelton , 1980 , p. 103) عدة عوامل يجب أخذها بالحسبان عند اختيار طريقة تحديد المحكات وهذه العوامل هي :

أ- أهمية القرارات .

ب- مقدار الوقت المتاح لوضع المحك .

ج- المصادر المتاحة (أشخاص ومال) للحصول على العمل المطلوب .

د- إمكانيات الأحكام (فبعض الطرق تتطلب معرفة المحتوى والمختبرين الواجب اختبارهم أكثر من غيرها) .

هـ- ملاءمة الطريقة لنوع الاختبار تحت الدراسة .

ويضاف على ذلك التكاليف النسبية لأخطاء سوء التصنيف والتي تقسم إلى نوعين :

- الخطأ من النوع الأول ويتمثل في نجاح طالب لا يستحق اجتياز الاختبار .

- الخطأ من النوع الثاني ويتمثل في رسوب طالب كان من الواجب أن ينجح (Geisinger, 1991 , P. 17) .

أما من ناحية تفضيل إحدى الطرق على غيرها فإن الدراسات التي تمت حتى الآن والتي قارنت بعض من هذه الطرق للتعرُّف على مدى الفروق بينها ما زالت محدودة ولم يحسم الأمر بشكل نهائي لصالح إحدى الطرق ، وبالتالي يوصى باستخدام أكثر من طريقة عند تقدير درجة القطع .

٣- حجم العينة :

حجم العينة هوعدد الأفراد الذين تطبق عليهم أداة البحث .

ويؤثر حجم العينة في قيمة معامل ثبات الاختبارات محكية المرجع فكلما زاد حجم العينة ارتفعت قيمة معامل الثبات وإن كان العكس غير صحيح ، فربما يرجع ارتفاع قيمة معامل الثبات لعوامل أخرى غير زيادة حجم العينة .

وقد أشارت دراسة داتشك (Dutschke , 1988) التي هدفت إلى تعيين مميزات وخصائص مؤشرات ثبات الاختبارات محكية المرجع إلى اقتراب خصائص التوزيع لمؤشرات الثبات من المميزات الطبيعية مع زيادة حجم العينة .

ولكن إلى متى تبقى هذه العلاقة صحيحة بمعنى هل ستستمر قيمة معامل الثبات بالارتفاع كلما زاد حجم العينة ، أم أنه عند حجم معين للعينة سيصل معامل الثبات الحد

الأعلى له طبعاً ضمن الظروف الأخرى كطول الاختبار ودرجة القطع ولن يزيد بعدها مهما زاد حجم العينة ؟

لقد توصلت دراسة وانغ (Wang ,1983) التي حاولت معرفة تأثير بعض العوامل على طرق حساب معامل الثبات : معامل كابا ، هينا ، صبكوفياك ، ومارشال وهيرتل حيث استخدمت عدد المختبرين (١٥ ، ٣٠) ، إلى أنه لا يتغير متوسط تقديرات مؤشر نسبة الاتفاق عندما ينقص عدد المختبرين من (٣٠) إلى (١٥) .

أي أن تأثير حجم العينة على معامل الثبات سينتهي عند حجم معين ، إلا أن الباحثة لم تعثر على دراسة أخرى تفيد في هذا المجال وتؤكد أو تنفي ما توصلت إليه الدراسة السابقة فربما يكون الأمر مختلف مع طرق أخرى لحساب معامل الثبات غير الطرق المذكورة في هذه الدراسة .

ومن ناحية أخرى فقد أظهرت دراسة كالون (Calohn , 1992) لتعيين مميزات مؤشرات الثبات : نسبة الاتفاق ، كابا ، كابا المصححة ، وفاي والتي استخدمت أربعة أحجام للعينات (٣٠ ، ٦٠ ، ١٢٠ ، ٢٤٠) ، أن الزيادة في حجم العينة أدت إلى نقص تحيز معنوي خصوصاً عندما وجدت درجة القطع في أطراف التوزيع . وربما انخفض تأثير حجم العينة عندما وجدت درجة القطع قرب منتصف التوزيع .

تعقيب :

تؤثر العوامل التي تم ذكرها – طول الاختبار ، درجة القطع ، حجم العينة – في حساب معامل ثبات الاختبارات محكية المرجع ، فزيادة طول الاختبار تؤدي إلى ارتفاع قيمة معامل الثبات وبالتالي زيادة اتساق قرارات التصنيف ، إلا أن تحديد العدد المطلوب من المفردات لقياس كل هدف سلوكي يقيسه الاختبار لم يتم تحديده والاتفاق عليه بشكل نهائي ، أي أن هناك آراء ترشد في عملية تحديد طول الاختبار ولكن لا توجد قاعدة عامة في ذلك وهذا ما تم الإشارة إليه عند بحث طول الاختبار .

وبالنسبة لدرجة القطع فقد أكدت كافة الدراسات حساسية معاملات الثبات لموضع درجة القطع ، حيث تزيد قيم معظم معاملات الثبات كلما اقتربت درجة القطع من أطراف التوزيع في حين يكون الأمرمختلفاً بالنسبة لمعامل كابا الذي يكون في أعلى قيمه عندما تقترب درجة القطع من متوسط التوزيع . وهناك بعض الاعتبارات عند اختيار طريقة تحديد درجة القطع إلا أنه لا يوجد تفضيل واضح لإحداها عن الأخرى ، ولا عجب في ذلك فالاختبارات محكية المرجع حديثة العهد مقارنة مع الاختبارات معيارية المرجع وبالتالي فإن ما يتعلق ببعض جوانبها لم تتخذ فيه قرارات حاسمة ولم تنل بعد الحظ الوافر من الدراسة وهذا ما يدعو لعمل مزيد من الأبحاث في هذا المجال .

أما بالنسبة للعامل الأخير ألا وهو حجم العينة فقد أشارت الدراسات أيضاً إلى أن زيادة حجم العينة سيؤدي إلى ارتفاع قيمة معامل الثبات .

وتخلص الباحثة إلى القول أن العوامل السالفة الذكر تؤثر في قيمة معامل ثبات الاختبارات محكية المرجع ولا يمكن القول بأن هناك طريقة لحساب معامل الثبات أفضل من غيرها من كافة الجوانب وفي كل الظروف ، فكل طريقة لها محاسنها وإسهاماتها ولها مساوئها في نفس الوقت .

وبالتالي من الأفضل تحديد أكثر الطرق مناسبة في ضوء ظروف معينة (طول اختبار محدد ، درجة قطع محددة ، حجم عينة محدد) ، وهذا ماستحاول هذه الدراسة بحثه من حيث اختيار أفضل طريقة من الطرق التي تقوم على حساب معامل ثبات الاختبارات محكية المرجع عند إعادة تطبيق الاختبار ضمن شروط محددة .

الخصائص الإحصائية لمفردات الاختبارات محكية المرجع :

١- تحليل المفردات :

يشير جرونلند (Gronlund , 1973 , p. 47) إلى أن الطرق التقليدية لتحليل المفردة والتي صممت للاستخدام مع الاختبارات معيارية المرجع التي تتطلب تباين في درجة الاختبار غير ملاءمة لتحليل مفردات الاختبارات محكية المرجع ، وذلك لأن تباين الدرجات غير مناسب طالما أننا نريد من كل الطلاب أن يحصلوا على درجات تامة في الاختبار عند نهاية التعليم .

وهذا ما تؤكده نادية عبد السلام (١٩٩٦ ، ص ص ١٦٢-١٦٣) في أنه لا يمكن استخدام طرق تحليل المفردة الكلاسيكية لاختيار مفردات الاختبار محكي المرجع لأنها صممت خصيصاً لإنتاج درجات لها أقصى تباين .

ولكن جاءت دراسة سعاد حسانين (٢٠٠٠) لتناقض ما سبق حين درست أثر طريقة اختيار المفردة على ثبات الاختبارات محكية المرجع وذلك باستخدام طرق اختيار المفردة : قوة تمييز المفردة ، معامل صعوبة المفردة ، معامل كوكس – فارجاس ، معامل برينان . حيث أظهرت نتائج هذه الدراسة إمكانية استخدام الطرق التقليدية لتحليل المفردة في تحليل مفردات الاختبار محكي المرجع .

وما دام الغرض من الاختبارات محكية المرجع هو قياس إتقان الطالب للمادة التعليمية يفترض ألا يستطيع الطالب الإجابة عن مفردات الاختبار قبل أن يتلقى التعليم ثم يتمكن من الإجابة عليها إجابة صحيحة بعد تلقي التعليم فيكون ذلك مؤشراً على الفاعلية التعليمية وعلى أن المفردة كانت حساسة لهذا التعليم واستطاعت قياسه .

وعلى هذا فإن المفردة الفعالة هي المفردة التي تبدو صعبة جداً على الطلاب قبل التدريس ولا يستطيع أن يجيب أحد عنها إجابة صحيحة ، ثم تصبح سهلة جداً بعد التدريس ويستطيع أن يجيب عنها معظم الطلاب إجابة صحيحة بعد إتقان المادة التعليمية موضوع الاختبار . أما في حالة عدم تمكن الطالب من الإجابة على المفردة لا قبل التعليم ولا بعده فهذا يعني وجود مشكلة فقد تكون المفردة صعبة جداً أو ضعيفة جداً ولا تقيس أثر التدريس أو أن التدريس نفسه غير فعال (صلاح الدين أبو ناهية ، ١٩٩٤ ، ص ٣٢٩) .

وإذا كانت الفكرة ايجاد مفردات حسَّاسة للتغيرات داخل الفرد ، عندئذ ستستبعد المفردات التي تبدي عدم اختلاف أو اختلاف قليل قبل وبعد التعليم . وأفضل المفردات هي تلك التي لها قيم معامل حساسية قريب من الصفر قبل التعليم وقيم تقترب من الواحد بعد التعليم (Keeves , 1988 , p. 383) .

ومن أهم طرق تحليل المفردات :

أولاً : أساليب تعتمد على القياس القبلي ــ البعدي :

١- معامل كوكس وفارجاس (١٩٦٦) : يطبق الاختبار على نفس مجموعة الطلاب قبل أن تتلقى التعليم وبعد أن تنتهي منه .

٢- معامل برينان (١٩٧٢) : يعتمد في تكوين المجموعات على قرار الإتقان وعدم الإتقان ويصف كيف تميز المفردة بين المتقنين وغير المتقنين .

٣- مقياس رودا بوش (١٩٧٣) : يحدد نسبة عدد الطلاب الذين أجابوا عن المفردة بشكل خاطئ قبل التعليم ولكنهم أجابوا عنها بطريقة صحيحة بعد التعليم .

٤- مقياس بابام (١٩٧٥) : يقارن استجابات الطلاب الصحيحة والخاطئة على المفردة قبل وبعد التعليم .

٥- معامل فاي : يستخدم معامل ارتباط بيرسون لفحص المفردة والأداء على الاختبار

هذا وتتراوح قيم المعاملات بين (-١ ، +١) حيث تشير القيم السالبة إلى وجود خلل في الهدف أو المفردة أو أن طريقة التعليم غير فعالة ، ووجود فروق كبيرة في قيم معامل السهولة قبل وبعد التعليم سيكون بمثابة دليل على صدق المفردة في قياس المهارة التعليمية ، أما إذا كانت الفروق صغيرة أو معدومة فإن ذلك يدل على وجود خلل بالمفردة .

وما يجمع بين هذه الطرق أنها تحتاج تطبيق الاختبار مرتين على نفس مجموعة الطلاب ، وأنه لا يمكن الحكم على فاعلية المفردة حتى يكتمل التعليم ويجرى الاختبار البعدي ، كما تؤثر الفترة الزمنية الفاصلة بين التطبيقين في النتائج (سعاد حسانين ، ٢٠٠٠، ص ص ٧٠-٧٨) .

ثانياً : أساليب تعتمد على اختبار مجموعتين منفصلتين من الطلاب في نفس الوقت :

وذلك باختبار مجموعة لم تتلق التعليم بعد وأخرى تلقت التعليم . وتتراوح قيم المعاملات أيضاً بين (-١ ، +١) . وهنا يجب التأكيد على ضرورة تكافؤ المجموعتين في كافة الخصائص بحيث يكون الفارق بينها فقط هو تلقي التعليم أو عدمه (سعاد حسانين ، ٢٠٠٠ ، ص ص ٧٨-٨٠) .

٢- مؤشر الصعوبة :

تبنى مفردات الاختبارات معيارية المرجع بحيث تكون متوسطة الصعوبة أي أن معامل صعوبتها حوالي (٥٠%) ، وبالتالي فإن المفردات التي يجيب عنها كل الطلاب (السهلة جداً) والمفردات التي لا يجيب أحد منهم عليها (الصعبة جداً) تحذف من الاختبار .

أما في حالة الاختبارات محكية المرجع عندما يكون معامل صعوبة مفردة منخفضاً فإن هذا ينبه المعلم لوجود خلل ما قد يرجع ذلك إلى أن أساليب التدريس غير فعالة أو أن توقعات المعلم عن مستوى طلابه ليست مناسبة وغير واقعية أو أن الفقرة ضعيفة وغير مناسبة . وبالتالي على المعلم أن ينظر في هذه الأسباب ويحدد أياً منها المسؤول عن فشل عدد كبير من الطلاب في الإجابة على المفردة ، وعليه أن يدرك أن معامل صعوبة المفردة مرتبط بأمرين هامين :

١- المستوى التعليمي الذي يوجد عليه الطلاب .

٢- المحك أو درجة القطع التي تم اختيارها (صلاح الدين أبو ناهية ، ١٩٩٤ ، ص ص ٣٢١-٣٢٢) .

ويرتبط مستوى صعوبة المفردة في الاختبار محكي المرجع بمستوى صعوبة الهدف الذي يقيسه الاختبار.

وهذا ما يشير إليه رجاء أبو علام في أن صعوبة البند تتوقف على صعوبة المهمة التي يتناولها مخرج التعلم ، وذلك لأن الاختبار محكي المرجع مصمَّم لتحديد العمل الذي يمكن للمتعلم القيام به فإذا كان العمل سهلاً ستكون بنود الاختبار سهلة ، وإذا كان العمل صعباً ستكون البنود صعبة . بمعنى أنه يجب أن تتطابق البنود مع مهام التعلم من جميع الوجوه على قدر الإمكان بما في ذلك مستوى صعوبة العمل (رجاء أبو علام ، ١٩٨٧ ، ص ص ١٤٦-١٤٨) .

وبما أن مستوى صعوبة المفردات يتحدد حسب الطبيعة الدقيقة لمخرجات التعلم المقاسة ، فعند إعداد الاختبار يكون من الأفضل تحديد مستوى صعوبة كل هدف مرغوب في قياسه وبالتالي نحتاج لبناء مفردات تتراوح من السهولة إلى الصعوبة بحيث نستطيع قياس درجة تقدم الطالب اتجاه الأهداف التي لم يتم تحقيقها ، وهكذا فإن مستوى صعوبة

المفردة ليس ضرورياً لقياس إتقان الحد الأدنى من الأساسيات لكنه ضروري لتحديد الحد الأعلى من مستوى الأداء الذي أحرزه الطالب (Gronlund , 1973 , p. 36).

وعموماً تميل مفردات الاختبارات محكية المرجع لأن تكون مكافئة لبعضها في مستوى الصعوبة ، ويميل معظم الطلاب لأن يجدوا المفردات سهلة وتجاب بشكل صحيح غالباً ، حيث يتم (٨٠%) منهم وحدة التعليم المتوقع أن تجاب كل مفرداتها بشكل صحيح (Cubiszyn & Borich , 2000 , p. 38) .

٣- مؤشر تمييز المفردة :

عند فحص المفردات في الاختبارات معيارية المرجع يتم حذف المفردات غير المميزة وذلك لتحقيق الهدف من هذه الاختبارات وهو مقارنة الطلاب ببعضهم وإظهار الفروق بينهم مما يؤثر على بناء الاختبار .

أما في الاختبارات محكية المرجع فيتم التركيز على ما يستطيع الطالب أن يؤديه ومدى وصوله إلى محك الإتقان أكثر من الاهتمام بمقارنته بأقرانه والتمييز بينهم ، ولذلك فإن معاملات التمييز ليس لها أهمية في هذه النوعية من الاختبارات .

فإذا قاست مفردة سلوكاً أساسياً بدقة وأجاب عنها كل الطلاب إجابة صحيحة أو خاطئة فإن ذلك لا يعتبر عيباً أو نقيصة في المفردة (Crehan , 1974 , p. 255) .

وبالتالي يسعى مصمم الاختبار للتأكد من أن المفردات تمثل انعكاساً للسلوك المحك وليس من المهم أن تكون مميزة أو غير مميزة ، وإنما المهم أن تمثل هذه المفردات مجموعة السلوك الذي يقيسه الاختبار (نادية عبد السلام ، ١٩٨١ ، ١٨٨) ، وهكذا تستبعد المفردات فقط إذا ثبت ضعف علاقتها بالمحتوى أو ثبت في تحليلها وجود مشكلة في صياغتها لا يمكن تصحيحها (إبراهيم مبارك الدوسري ، ٢٠٠٠ ، ص ١٣٦) .

تعقيب :

يتضح من العرض السابق أن تطور فكرة القياس من معياري المرجع إلى محكي المرجع رافقه تطور في التقنيات المستخدمة معه ومن بينها طرق حساب معامل الثبات الذي لا مجال للشك في أهميته عند استخدام أي أداة .

تحاول الدراسة الحالية معرفة التغيرات التي تطرأ على قيمة معامل الثبات المحسوب باستخدام كل طريقة من الطرق موضع الدراسة وذلك عند زيادة طول الاختبار وحجم العينة وتغيير موضع درجة القطع ، كما تحاول المقارنة بين بعض طرق حساب معامل ثبات

الاختبارات محكية المرجع التي تعتمد على تطبيق الاختبار مرتين أو تطبيق صور متكافئة منه ، كون هذه الطرق سهلة وسريعة الحساب ولا تتطلب تحقيق شروط معينة أو مجهود كبير في حسابها وذلك في ضوء أكثر العوامل أثراً على قيم معاملات الثبات .