أثر استخدام بيانات التغذية الراجعة وخبرة المحكم في تقدير درجة القطع وممارسات المحكمين وفق طريقة أنجوف المعدلة

د./ إكرام حمزة السيد صهوان[1]

ملخص الدراسة:

هدفت الدراسة إلى تقصي أثر استخدام بيانات التغذية الراجعة وخبرة المحكم في تقدير درجة القطع وممارسات المحكمين وفق طريقة أنجوف المعدلة. ولتحقيق هدف الدراسة قامت الباحثة ببناء اختبار تحصيلي في مادة العلوم للصف الأول الإعدادي وتدريجه وفق نموذج راش، وبناء استبيان حول الممارسات العملية التي يقوم بها المحكمون خلال عملية تقدير درجة القطع، والعوامل المؤثرة في قراراتهم عند استخدام طريقة (نعم/لا) لأنجوف. وشملت عينة الدراسة خمسة عشر محكما لدرجات القطع من المعلمين وأعضاء هيئة التدريس.

تم عقد ورشة لتقدير درجات القطع لتصنيف الطلاب على اختبار العلوم إلى أربعة مستويات (متعثر-نام – مرض-متقدم) من خلال جولتي تحكيم يزود المحكمون بينهما بثلاثة أنواع للتغذية الراجعة متمثلة في: البيانات الجماعية، وبيانات الأثر، والبيانات الواقعية/ التجريبية.

وتمثلت أهم النتائج في:

– انخفاض قيم وسيط درجات القطع لمستويات الأداء " نام -مرض -متقدم " النهائية لأعضاء هيئة التدريس عن مثيلاتها بالنسبة للمعلمين. مع اتساق تقديرات المحكمين من المعلمين وأعضاء هيئة التدريس لكافة مستويات الأداء في كل من جولتي التحكيم.

– عدم وجود فروق دالة إحصائيا في تقديرات درجة القطع لمستويات الأداء (نام – مرض – متقدم) في جولتي التحكيم الأولى والثانية ترجع لخبرة المحكم.

– انخفاض تقديرات درجة القطع في الجولة الثانية عنها في الجولة الأولى بشكل دال إحصائيا لكافة مستويات الأداء (نام-مرض-متقدم) بالنسبة لعينة المعلمين، والعينة ككل وكذلك بالنسبة لمستوى (نام -متقدم) بالنسبة لأعضاء هيئة التدريس؛ في حين لم ترق الفروق في هذه التقديرات لمستوى "مرض" بالنسبة لأعضاء هيئة التدريس إلى حد الدلالة الإحصائية.

– جاءت مراجعة المواصفات الخاصة والعامة لمستويات الأداء أكثر الممارسات التي اعتمد عليها المحكمون في اتخاذ قراراتهم النهائية لتقدير المفردات في الجولتين الأولى والثانية. وأثرت

---

[1] مدرس علم النفس بالمركز القومي للامتحانات والتقويم التربوي

مجموعة من العوامل بدرجة كبيرة في قرارات المحكمين من أعضاء هيئة التدريس والمعلمين في الجولة الثانية، وتمثلت هذه العوامل في: مفهوم الطالب ذى الحد الأدنى من الأداء، والمناقشة بين الجولات والخبرة المهنية، وقيمة معامل الصعوبة والتوزيع التكراري لدرجات الطلاب، والتحليل البياني للمفردات. بينما جاءت درجة القطع بالنسبة للمحكمين الآخرين أقل العوامل تأثيرا في قرارات المعلمين وأعضاء هيئة التدريس.

أثر استخدام بيانات التغذية الراجعة وخبرة المحكم في تقدير درجة القطع وممارسات المحكمين وفق طريقة أنجوف المعدلة

د./ إكرام حمزة السيد صهوان[٢]

مقدمة:

تعد الاختبارات من أهم أدوات القياس التي يستند إليها في اتخاذ قرارات مهمة في حياة الفرد والمجتمع؛ حيث عرفتها البشرية منذ القدم عندما كان الصينيون القدماء يفرضون اختبارا ما على كل من يتقدم لمنصب مهم في الدولة، ثم يحددون نسبة أو مستوى معينا لنجاحه حتى يتم قبوله في هذا المنصب.

وقد شغلت مشكلة تصنيف المتعلمين حسب مستوى تمكنهم من نواتج تعلم معينة علماء القياس على مر الزمن؛ فقبل سبعينات القرن الماضي تمثل الهدف من تصنيف المتعلمين في انتقاء القلة منهم لمتابعة تعليمهم في المرحلة الثانوية والجامعية، وإقصاء الباقين منهم عن متابعة الدراسة؛ مما أدى إلى ظهور الاختبارات معيارية المرجع Norm Referenced Tests. وجاءت هذه الاختبارات أكثر مناسبة لأغراض التقويم المبدئي، والتقويم النهائي.

وقد شاب هذا النوع من الاختبارات عددا من أوجه القصور خاصة في مجال التقويم التشخيصي، وتقويم فاعلية البرامج التعليمية؛ مما ساعد على ظهور فلسفات تربوية جديدة تعتمد بصفة أساسية على مبدأ التعلم من أجل الإتقان وبرامج التعليم الفردي. والتأكيد على أن نسبة كبيرة من المتعلمين يمكنهم أن يتقنوا المحتوى المراد قياسه إذا ما أتيح لهم الوقت الكافى وقدمت لهم المساعدة الملائمة (Kane,2006).

وتزايد نتيجة لذلك الاهتمام بمدخل القياس محكي المرجع الذي يقوم على تصنيف المتعلمين إلى مجموعات وفق درجة إتقانهم للمعارف والمهارات المرتبطة بمقرر دراسي أو برنامج تدريبي معين؛ حيث تقوم فلسفة الاختبارات محكية المرجع على قياس أداء الفرد بالنسبة إلى محك أو مستوى أداء محدد مسبقا، دون الحاجة إلى مقارنة أداء الفرد بأداء أقرانه.

وعلى ذلك فإنه في إطار القياس محكي المرجع يكون تحديد المواصفات القياسية للأداء Setting Performance Standards والتوصل إلى درجات قطع Cut Score مناسبة خطوة

---

[٢] مدرس علم النفس بالمركز القومي للامتحانات والتقويم التربوي

مهمة في سياق استخدام درجات الاختبارات لاتخاذ القرارات حول الممتحنين (فوزية هادي، وصلاح مراد، ٢٠١٣).

كما يعتبر تحديد المواصفات القياسية للأداء هدفا أساسيا لتوثيق النواتج التعليمية في كثير من الأنظمة التربوية/ التعليمية العالمية، بالإضافة إلى أن التوصل إلى درجات القطع ليست خطوة مستقلة في منظومة بناء الاختبار وإنما خطوة مهمة خاصة عند تفسير درجات الاختبار، ومن ثم عند تقديم الأدلة على صدق درجاته (Kane,2006) (Plake,2008). ومن ثم فإن عملية **تحديد مستويات الأداء ومحكاته** وتحديد درجات القطع ترتبط ارتباطا وثيقا بالسياسة التربوية للنظام التعليمي (Cizek, 1996).

وتهدف عملية وضع المواصفات القياسية للأداء إلى وضع درجات قطع فاصلة بناء على المواصفات المرغوبة من المؤسسة التربوية بحيث تتيح تصنيف الأفراد إلى مستويات الأداء المختلفة بناء على درجات الاختبار (Hambleton & Pitoniak, 2006). ومن ثم تؤثر نتيجة عملية تحديد المواصفات القياسية للأداء في حياة الأفراد بقدر أهمية الهدف من الاختبار. ففي بعض المواقف التقييمية خاصة في الاختبارات ذات الأهمية القصوى High Stakes Assessments, فإن درجات القطع غير الدقيقة قد يترتب عليها نتائج بالغة التأثير. فقد تؤدي درجات القطع المرتفعة المبالغ فيها إلى حرمان طلاب أكفاء من الالتحاق ببرامج دراسية معينة، وعلى النقيض من ذلك فإن درجات القطع المنخفضة بشكل مبالغ فيه قد تسمح بالتحاق طلاب على غير المستوى المناسب للالتحاق بالبرامج التي لا تتناسب وقدراتهم؛ في حين أن التصنيف الدقيق للأفراد باستخدام درجات قطع مناسبة يساعد في تطوير مستويات التحصيل والأداء والإنتاج، وكذلك الحد من قيمة الفاقد المادي والبشرى وكذلك ترشيح الأكفاء للمهن التي يرتبط فيها القرار بمصلحة المجتمع وأفراده خاصة في مجالات مهمة كالصحة والتعليم، والمهن الهندسية (راشد الدوسرى،٢٠١٢، ١٠٥) (Sireci et al., 2009).

كما قد يمتد أثر عملية تحديد المواصفات القياسية للأداء ليصل إلى كافة متخذي القرار وأطراف العملية التعليمية؛ لذلك فإن هناك حاجة إلى مزيد من الانتباه والجهد في هذا المجال للوصول إلى آليات مناسبة لتحديد درجات القطع بشكل يسمح بعدالة وموضوعية القياس (Pellegrino, Jones & Michell,1999).

وتبدأ عملية **تحديد المواصفات القياسية للأداء** ببناء مواصفات الأداء Performance Descriptors وتنهي بالتوصل إلى درجة/ درجات القطع. ويتطلب الحكم ما إذا كان الطالب قد

حقق المحك المطلوب أولا تحديد المواصفات القياسية الكيفية Qualitative Standards للأداء المطلوب أي تحديد المعارف والمهارات المرجوة، ثم تحديد كم الأداء " المواصفات القياسية الكمية Quantitative Standardالمطلوب تحقيقه، ويعبر عن هذا الكم من الأداء إجرائيا في درجة القطع Cut Scores على الاختبار.

وقد طور مع ظهور مدخل الاختبارات محكية المرجع **عدد من الطرق والنماذج** لتحديد مستوى الأداء المطلوب للمهام المختلفة. حيث استبدلت الطرق النسبية التي تعتمد على مقارنة درجة المتعلم بمتوسط درجات أقرانه في إطار فلسفة القياس جماعي المرجع بتلك الطرق المطلقة التي تحدد درجة قطع تمثل الحد الأدنى[٣] من الأداء المطلوب لعمل ما في إطار فلسفة القياس محكي المرجع. ويرجع الفضل في ظهور الطرق المطلقة هذه إلى نيدلسكاى (Nedelsky, 1954) حيث تستهدف تحديد مستوى الأداء وفقاً لما يجب أن يصل إليه الطالب من مستوى الإنجاز. ومن هذه الطرق طريقة نيدلسكاي، طريقة المجموعات المتضادة، وطريقة أنجوف، ونموذج ذي الحدين وغيرها (صلاح علام،٢٠٠٧).

وتعددت مسميات مستويات الأداء وعددها، فبالنسبة لبعض الأغراض يكتفي بمستويين فقط (متقن/ غير متقن)، ولأغراض أخرى يستخدم أكثر من مستوى للأداء على متصل القياس (متعثر / نام/ مرض/ متقدم) على سبيل المثال.

وتتعدد الطرق المتبعة في تحديد درجات القطع، وتختلف باختلاف السياق الخاص بالعملية الاختبارية وطبيعة المشاركين، كما تتباين في إجراءاتها. وتعتبر طريقة "أنجوف" Angoff Method من أكثر المداخل انتشارا في تحديد مستويات الأداء ودرجات القطع (Margelios & Clauser,2014,15). وتشير نتائج عدد من الدراسات إلى ثبات طريقة أنجوف وسهولة استخدامها في تقدير درجة القطع كونها تتمتع بمؤشرات ثبات وصدق مقبولة (أحمد الشريم ويوسف السوالمة، ٢٠٠٦). كما توسع الباحثون في تطبيق طريقة أنجوف سواء ورقيا أو إلكترونيا من خلال شبكة الانترنت (Katz &Tannenbaum, 2014, 1-17).

وفي مدخل أنجوف التقليدي يطلب من المحكمين الخبراء في محتوى الاختبار أن يستدخلوا مفهوم الطالب ذى الحد الأدنى من الأداء Minimally Proficient Examinee (MPE) أي الذي

---

[٣] تعبير الحد الأدنى لا يعني ضآلته وغنما تشير إلى أن هذا الحد يمثل الحد اللازم والكاف من المعارف والمهارات التي عندها يمكننا القول بان الطالب قد حقق هذا المستوى من الأداء ويعني ذلك أن الطالب امتلك من المعارف والمهارات ما تضعه على عتبة هذا المستوى.

يمتلك الحد اللازم من المعارف والمهارات الملائم لمستوى الاجتياز. ثم يطلب منهم أن يراجعوا مفردات الاختبار، ويضع كل محكم بالنسبة لكل مفردة تقديرا لنسبة الطلاب ذوي الحد الأدنى من الأداء المحتمل أن يجيبوا على هذه المفردة صوابا. ويتم تجميع هذه النسب لكل مفردات الاختبار لكل محكم، ويمثل متوسط القيمة الناتجة تقديرا للدرجة المتوقع أن يحصل عليها الطالب ذى الحد الأدنى من الأداء أو درجة القطع (Jaeger, 1995).

وقد أجري عدد من التعديلات على طريقة "أنجوف" منها التعديل المعروف "بطريقة توكر" حيث تكون المهمة المطلوبة من المحكم أن يقرر ما إذا كان الطالب ذى الحد الأدنى من الأداء يمكنه أن يجيب على المفردة صوابا أم لا. وذلك تيسرا على المحكم فبدلا من أن يقدر المحكم احتمالية الإجابة الصحيحة لكل مفردة بالنسبة للطالب ذى الحد الدنى من الأداء، فإن تقديره يكون إما "نعم" في حالة ما إذا كان الطالب ذو الحد الأدنى من الأداء المطلوب يمكنه الإجابة صوابا على المفردة، أو يكون تقديره "لا" في حالة عدم قدرة الطالب ذى الحد الأدنى من الأداء الإجابة على المفردة صوابا، ومن ثم عرفت بطريقة نعم/ لا لأنجوف.

مشكلة الدراسة:

مع تعدد الطرق المستخدمة في تحديد مستويات الأداء _ ومن بينها طريقة أنجوف_ فإنها تتضمن مشاركة مجموعة من المحكمين الخبراء (Cizek,2012; Impara & Plake 1997). ومن ثم تتوقف مصداقية الطريقة المتبعة في تحديد درجة القطع على قدرة هؤلاء المحكمين على التمييز بين الأفراد تبعا لمستوى تمكنهم من الكفاية موضع القياس.

ومن ثم يكون الاتساق بين المحكمين عاملا أساسيا في عملية تحديد مستويات الأداء حيث إنه إذا ما تم الاتساق بين المحكمين فإنه يتم تصنيف الأفراد بنفس الطريقة.

وفقا لعدد من الباحثين فإن معظم طرق تحديد مستويات الأداء _ومن بينها طريقة أنجوف_ تتضمن تحدياً معرفياً للمقدرين/ المحكمين؛ حيث تتطلب جميعها أن يصدر المحكمون تقديراتهم حول احتمالية الإجابة الصحيحة على كل مفردة من مفردات الاختبار من قبل طلاب بعينهم (ذوي الحد الأدنى من الأداء/ الطلاب الحديون) خاصة في غياب البيانات عن الأداء الفعلي للأفراد على مفردات الاختبار (Clauser,Mee,Baldwin, Margolis& Dillon,2009).

لذا فقد اقترح بمرور الوقت عدة تعديلات على طريقة أنجوف (Cizek & Bumch, 2007)؛ بحيث يتم تزويد المحكمين ببعض البيانات عن أداء الطلاب الفعلي على الاختبار،

وكذلك حول بعض المؤشرات الإحصائية لمفردات الاختبار. حيث تجرى التقديرات على جولتين أو أكثر بحيث تجرى الجولة الأولى وفق-طريقة أنجوف المعتادة دون تزويد المحكمين بالبيانات الفعلية حول مفردات الاختبار. وبعد انتهاء تقديراتهم للجولة الأولى يتم مناقشة هذه التقديرات معهم وتزويدهم ببعض البيانات المشتقة من تطبيق الاختبار فعليا على عينة مناسبة من الطلاب.

وعرفت البيانات المقدمة للمحكمين بين جولات التحكيم ببيانات التغذية الراجعة Feedbackوتباينت الدراسات في طبيعة وكم بيانات التغذية الراجعة المقدمة للمحكمين حيث تضمنت واحدة أو أكثر من مجموعات البيانات الممثلة في **بيانات جماعية** تتضمن تقديم درجات القطع العامة بالنسبة لكل المشاركين التي توصلوا إليها في جولة التحكيم، والانحراف المعياري، والمتوسط، والوسيط والمنوال لدرجات القطع لمجموعات المحكمين. ومناقشة التقديرات المبدئية الفردية للمحكمين. و**بيانات الأثر** وتتمثل في التوزيع التراكمي لدرجات الطلاب على مستويات الأداء وفق درجات القطع التي حددها المحكمون في جولة التحكيم حيث يتم **تصنيف الطلاب** على مستويات الأداء حيث توضح توزيع الدرجات الخام للممتحنين أو النسبة المئوية للمختبرين في كل فئة وفي كل مستوى من مستويات الأداء عندما تطبق درجة القطع المحددة من قبل المحكمين في الجولة الأولى. و**بيانات تجريبية واقعية** تتعلق بمؤشرات المفردات المشتقة من التجريب الميداني وتتضمن تزويد المشاركين بصعوبات المفردات والتوزيع التكراري لبدائل كل مفردة (Cizek&Bunch,2007).

ومن ثم جاء اهتمام بعض الباحثين بما يمارسه المحكمون من عمليات عقلية عندما يصدرون أحكامهم حول مفردات الاختبار قبل تزويدهم ببيانات التغذية الراجعة أي خلال تقديرات الجولة الأولى، وبعد تزويدهم ببيانات التغذية الراجعة أي خلال تقديراتهم بالجولات التالية؛ ومن ثم الوقوف على ما قد يطرأ على درجات القطع من تغيير نتيجة لبيانات التغذية الراجعة خاصة بالنسبة لطريقة أنجوف المعدلة، وكذلك الوقوف على ما قد يطرأ على ممارسات المحكمين وما يقومون به من عمليات عقلية نتيجة لتزويدهم ببيانات التغذية الراجعة، وتقصي هذه العمليات من خلال الإجابة عن السؤال " **فيم يفكر المحكمون عندما يضعون تقديراتهم حول كل مفردة من مفردات الاختبار؟**"

وتباينت نتائج الدراسات حول تأثير تزويد المحكمين بالبيانات حول أداء الممتحنين ففي حين توصلت نتائج بعض الدراسات إلى أن تزويد المحكمين ببيانات عن أداء الممتحنين ينتج عنه خفض درجات القطع (Hurtz & Auerbach,2003). أسفرت نتائج بعض الدراسات عن زيادة

درجات القطع عند تزويد المحكمين ببيانات التغذية الراجعة، كما أشارت نتائج دراسات أخرى إلى أن التغير في قيمة درجة القطع التي يقدرها المحكمون تختلف سواء بالزيادة أو الانخفاض باختلاف موضوعات المحتوى في الاختبار الواحد (Clauser et,al2009) (Busch& Jaeger,1990) (Cross,Impra,Frary& Jaeger,1984) (Plake, Impara&Pontenza,1994) (Truxillo,Donahue,&Sulzer,1996). ووفقا لكان (Kane,1994) فإن تزويد المحكمين ببيانات واقعية عن خصائص الطلاب ومستواهم الدراسي، ومناقشة ذلك بشكل جماعي معهم، يؤثر في دقة واتساق تقديراتهم لمستوى أداء الطلاب وتحديد درجة القطع ومستوى الإتقان المتوقع.

في حين أشار (Hambleton,2001,102) إلى أن تأثير بيانات التغذية الراجعة قد يكون نفسيا أكثر منه سيكومتريا. وأوضح أن الأثر الرئيس لبيانات الأداء غالبا ما يكون على التباين بين المحكمين أكثر من أثرها على درجة القطع العامة. وراجع (Brandon,2004) ست دراسات وتوصل إلى أن تزويد المحكمين ببيانات التغذية الراجعة قد أدى إلى تغييرات جوهرية في درجة القطع في أربع دراسات منها.

وتباينت نتائج الدراسات من حيث تأثير التغذية الراجعة على ما يمارسه المحكمون من عمليات معرفية أثناء تحديد درجة القطع، وتوصلت بعض الدراسات إلى أن بيانات التغذية الراجعة أثرت بدرجة كبيرة على قرارتهم في الجولة الثانية مثل دراسة (Margolis & Clauser,2014)، في حين أشارت نتائج بعض الدراسات إلى أن قرارتهم لم تتأثر ببيانات التغذية الراجعة، وإنما ساعدتهم التغذية الراجعة فقط في فهم مستويات الأداء وفهم مفهوم الطالب ذى الحد الأدنى من الأداء مثل دراسة (Ferdous and Buckendaht, 2013).

وكما تباينت نتائج الدراسات في تأثير بيانات التغذية الراجعة على تقديرات المحكمين تباينت أيضا في **طبيعة البيانات** المقدمة للمحكمين؛ حيث اقتصرت بعض الدراسات على تقديم صعوبة المفردات، ومنها دراسة أحمد الشريم، ويوسف السوالمة(٢٠٠٦)، ومنها ما امتدت لتزويد المحكمين ببيانات الأثر حول النسب المئوية للطلاب في مستويات الأداء المختلفة نتيجة لتطبيق درجة القطع بالإضافة إلى صعوبات المفردات مثل دراسة (Ferdous and Buckendaht, 2013) ، وأضافت دراسة (Margolis & Clauser,2014) إلى هذه العوامل التوزيع التكراري لبدائل كل مفردة. كما أوضحت دراسة (Engelhard & Stone,1998) أن ثبات الاستقرار لتقديرات المحكمين لدرجة القطع للاختبار أعلى في حال معرفتهم لمعاملات صعوبة الفقرات، وأن كلا من كفاية المحكمين وتجانس صياغة الفقرات يؤثر على درجة القطع الكلية للاختبار.

ويرتبط بتحديد مستويات الأداء ودرجات القطع على الاختبار أيضا متغير خبرة المحكمين وآلية انتقائهم للمشاركة في تقدير درجات القطع. حيث عادة ما يتم اختيار المحكمين بناء على خبرتهم في المادة الأكاديمية محل الاختبار من المعلمين أو المشرفين التربويين والموجهين أو لخبرتهم في التربية بوجه عام من أعضاء هيئة التدريس في مجال القياس والتقويم (Cizek&Bunch,2007). ومع اختلاف خلفية المحكمين يكون من الضروري التحقق من درجة الاتساق في درجات القطع والمحكات التي تعكسها هذه الدرجات.

وينادي كثير من الباحثين في الآونة الأخيرة بضرورة تنوع مجموعات المحكمين وعدم الاقتصار على المعلمين فقط في عملية التحكيم (Barman,2008) (راشد الدوسري، ٢٠١٢). وقد تباينت نتائج الدراسات حول تأثير خبرة المحكم في تقدير درجة القطع فقد أشارت نتائج دراسة (صلاح علام،١٩٩١)، ودراسة (خالد التميمي، ١٩٩٩) إلى أن اختلاف خبرة المحكم لا يؤثر تأثيرا جوهريا في تقديرات درجة القطع ، في حين أشارت نتائج دراسة (Halpin; Sigmon & Halpin,1983) (Margolis & Clauser,2014) إلى أن تقديرات المعلمين لدرجات القطع تختلف اختلاف جوهريا عن تقديرات أعضاء هيئة التدريس. كما خلصت دراسة شانج وزملائه (Chang et.al., 1996, 161 – 173) إلى أن المحكمين يميلون لأن يضعوا تقديرا أعلى للفقرات التي يجيبون عليها إجابة صحيحة من الفقرات التي يجيبون عليها إجابة خطأ. كما أشارت إلى وجود تأثير لمعلومات ومعارف المحكمين في تحديد درجات القطع المتعلق باختبارات الكفاءات وفق طريقة أنجوف.

ويزداد الاهتمام بخبرة المحكم في حالة تزويد المحكمين ببعض بيانات التغذية الراجعة؛ حيث يبرز السؤال حول هل تؤثر بيانات التغذية الراجعة على المحكمين من ذوي الخبرات المختلفة بنفس الطريقة؟

وفي ضوء تباين نتائج الدراسات حول مدى جدوى استخدام بيانات التغذية الراجعة، وتباين نتائجها حول أي بيانات التغذية الراجعة أكثر تأثيرا، وطبيعة هذا التأثير، وندرة الدراسات العربية التي تناولت تأثير بيانات الأثر على تقديرات درجة القطع حيث لا توجد سوى دراسة (أحمد الشريم ويوسف السوالمة،٢٠٠٦) واقتصرت فقط على تزويد المحكمين بمعامل صعوبة المفردات.

ومع التطور في مجال بناء الاختبارات وظهور نظرية الاستجابة للمفردة Item Response Theoryبما تسمح به من تدريج لمفردات الاختبار على ميزان تدرج واحد بصفر

مشترك واحد. كما تساعد في التغلب على مشكلات القياس التقليدي مثل تأثر خصائص مفردات الاختبار بخصائص عينة التقنين، وكذلك تأثر قدرة الفرد بطبيعة المفردات التي أدى الاختبار عليها. حيث تحقق موضوعية القياس، كما تتمثل في تحرر قدرة الفرد من خصائص المفردات وتحرر صعوبات المفردات من قدرات الأفراد.

ومع التطور التكنولوجي وظهور البرامج الإحصائية التي تيسر التعبير عن البيانات الإحصائية الكمية بيانيا. فقد عرف التحليل البياني للمفردات الاختبارية Graphical of Item Analysis (GIA) وهو عبارة عن وسيلة بصرية للتعبير عن المفردات الاختبارية بيانيا توضح العلاقة بين الدرجة الكلية على الاختبار ونسب الاستجابة للبديل الصحيح والبدائل غير الصحيحة للمفردات الاختبارية من نوع الاختيار من متعدد. وتوفر معلومات أساسية سهلة التفسير عن خصائص المفردة (الصعوبة، التمييز، التخمين، فعالية المشتتات). بحيث يسهل التعرف مباشرة من خلال التحليل البياني للمفردة على قدرة المفردة على التمييز بين الأفراد، وكذلك مدى التمييز، وكذلك فعالية المشتتات/ البدائل، ومستوى صعوبة المفردة، ومن ثم يمكن بسهولة لغير المتخصصين بعمق في مجال الإحصاء الحكم على جودة المفردة بدقة واتخاذ القرار المناسب بشأنها (Batenburg & Laros ,2000).

ولما كان تقدير المفردات خلال عملية تحديد درجة القطع يتطلب تزويد المحكمين ببيانات حول المفردات مثل معامل الصعوبة، ونسب اختيارات البدائل لكل مفردة وغيرها وحيث إن المحكمين يأتون من خلفيات إحصائية متباينة، كما أن التعامل مع القيم يكون أحيانا أكثر صعوبة من التعامل مع الأشكال البيانية –خاصة لدى غير المتخصصين– كانت فكرة الباحثة في استخدام التحليل البياني للمفردات في تطوير بيانات التغذية الراجعة المقدمة للمحكمين خلال مرحلة مراجعة التقديرات لتيسير التعامل مع بقية المؤشرات الإحصائية المقدمة للمحكمين من ناحية، وكذلك الاستفادة مما يوفره التحليل البياني للمفردة من إمكان التعرف بسهولة على أكثر المناطق على متصل قدرة الأفراد تكون فيه المفردة أكثر قدرة على التمييز بين الأفراد. بالإضافة إلى أن صعوبة مفردات الاختيار من متعدد تكمن بصفة أساسية في بدائل المفردة فإن تحليل البدائل الذي يوفره التحليل البياني للمفردات قد يسهم في تيسير مهمة المحكمين وتساعدهم في الحكم الدقيق على المفردة بشكل متكامل من حيث قدرة الطالب ذى الحد الأدنى من الأداء يمكنه الإجابة صوابا عن المفردة.

وبالإضافة إلى ما تقدم فقد اهتم عدد كبير من الدراسات السابقة بمستويين للأداء هما النجاح/ الرسوب؛ في حين أن الواقع العملي، وما يتطلب من تحسين لمستويات أداء الطلاب وتعلمهم من

مستوى لآخر يتطلب أن يضع المعلم نصب عينيه مستويات أداء المتعلمين المختلفة سعيا للوصول بهم كل إلى ما تسمح به إمكاناته وقدراته، كما أن فلسفة القياس محكي المرجع، وكذا كافة أنظمة التقويم الحديثة تستهدف التشخيص الدقيق لمستوى المتعلم حتى يمكن تقديم البرامج العلاجية المناسبة له كل وفق احتياجاته.

فقد جاءت فكرة الدراسة الحالية في دراسة أثر بيانات التغذية الراجعة وخبرة المحكم على تقديرات درجة القطع وممارسات المحكمين وفق طريقة أنجوف المعدلة، من خلال بناء اختبار التحصيل الدراسي في مادة العلوم للصف الأول الإعدادي[٤] وتدريجه باستخدام نظرية الاستجابة للمفردة، وتطوير بيانات التغذية الراجعة لتشمل بيانات جماعية (تتمثل في درجات القطع للمحكمين الآخرين-المتوسط – الوسيط-المنوال-الانحراف المعياري-القيم العظمى-القيم الصغرى) لدرجات القطع. كما تشمل بيانات الأثر (النسبة المئوية للطلاب في كل مستوى من مستويات الأداء وفق درجة القطع المقدرة من المحكمين في الجولة الأولى). كما تشمل أيضا البيانات الواقعية (وتشمل معامل الصعوبة، النسب المئوية لاختيار كل بديل، بالإضافة إلى التحليل البياني للمفردات)، وتطبيق درجات القطع الناتجة لتصنيف الطلاب على اختبار العلوم إلى أربعة مستويات أداء (متعثر-نام-مرض-متقدم).

أسئلة الدراسة:

١- ما درجات القطع المحددة من قبل المحكمين عينة الدراسة لكل مستوى من مستويات الأداء (متقدم-مرض-نام) عبر جولتي التحكيم؟

٢- هل يختلف وسيط درجات القطع لمستويات الأداء الثلاثة باختلاف خبرة المحكم في الجولة الأولى قبل تقديم بيانات التغذية الراجعة؟

٣- هل يختلف وسيط درجات القطع لمستويات الأداء الثلاثة باختلاف خبرة المحكم في الجولة الثانية بعد تقديم بيانات التغذية الراجعة؟

٤- هل تختلف تقديرات درجة القطع للمحكمين عينة الدراسة (المعلمون، وأعضاء هيئة التدريس، والعينة ككل) في الجولة الثانية عنها في الجولة الأولى بالنسبة لكل مستوى من مستويات الأداء؟

---

[٤] أجريت الدراسة في مادة العلوم نظرا لأنه تخصص الباحثة في الدرجة الجامعية الأولى، كما شاركت الباحثة كثيرا في مجال عملها بالمركز القومي للامتحانات والتقويم التربوي في بناء اختبارات التحصيل الدراسي في مادة العلوم.

٥- ما نسبة الطلاب في كل مستوى من مستويات الأداء عند استخدام درجة القطع المحددة لكل مستوى في جولتي التحكيم الأولى والثانية؟

٦- ما مدى وضوح المواصفات القياسية للأداء والمفاهيم المرتبطة بها قبل إجراء الجولة الأولى من التقديرات من وجهة نظر المحكمين عينة الدراسة؟

٧- ما الممارسات التي يتبعها المحكمون عندما يصدرون أحكامهم على المفردات في الجولة الأولى قبل تقديم بيانات التغذية الراجعة من وجهة نظرهم؟

٨-ما الممارسات التي يتبعها المحكمون عندما يصدرون أحكامهم على المفردات في الجولة الثانية بعد تقديم التغذية الراجعة من وجهة نظرهم؟

هدف الدراسة:

هدفت الدراسة الحالية إلى دراسة أثر تزويد االمحكمين ببعض بيانات التغذية الراجعة(جماعية، بيانات الأثر، البيانات الواقعية للمفردات) وخبرة المحكم (المعلمين وأعضاء هيئة التدريس) على تقديراتهم لدرجات القطع وفق طريقة أنجوف المعدلة لمستويات الأداء المتعددة. و تحديد أكثر العوامل تأثيرا في الممارسات العملية للمقدرين أثناء إصدار تقييماتهم على مفردات الاختبار بالنسبة لمستويات الأداء المختلفة (متقدم – مرض-نام-متعثر) ، بالإضافة إلى تحديد أكثر المؤشرات الإحصائية للمفردات التي يمكن أن تسهم في فهم المحكمين لمتطلبات الأداء عبر مفردات الاختبارات.

أهمية الدراسة:

- يمكن الاستفادة من نتائج الدراسة الحالية في تقديم التغذية الراجعة اللازمة لتطوير عملية تحديد مستويات الأداء ودرجات القطع، وكذلك تطوير أداء المحكمين الأمر الذي من شأنه أن يحسن في الممارسات التقويمية المختلفة. ويسهم في إصلاح المنظومة التعليمية وتحقيق العدالة في اتخاذ القرارات.
- قد تسفر الدراسة عن تطوير طريقة أنجوف الأمر الذي من شأنه أن يساعد في اتخاذ القرارات الصائبة فيما يتعلق بتصنيف الطلاب وتقويمهم وتقديم الخدمة التفسيرية المناسبة لأداء كل منهم.
- التوصل إلى درجة القطع بناء على معلومات وبيانات ميدانية تجريبية واقعية بالإضافة إلى خبرة المحكم قد يساعد في الوصول إلى درجة أكثر تعبيرا عن مستوى الطالب.
- الإضافة إلى الأدبيات في مجال تحديد مستويات الأداء ودرجات القطع من خلال الدراسة المنظمة لممارسة تقديم بيانات الأداء للمحكمين في وضع المواصفات القياسية القائمة على

المحتوى (المواصفات الكمية) في عدد من الاختبارات ذات الأهمية القصوى.

- قد تساعد نتائج الدراسة الحالية في تحسين جودة الممارسات التربوية المرتبطة بتصنيف الطلاب وتوجيههم إلى المسارات التعليمية المختلفة.
- قد تساعد الطريقة الحالية في تطوير آلية التحكيم لفحص تقديرات المحكمين وتقديم وسيلة يستفيد منها الباحثون في التحقق من صدق أدواتهم. ومن ثم تحسين جودة حكم الخبراء في الاختبارات من خلال تلافي مشكلات التحكيم.
- مع تنامي أهمية العلم والتعلم وظهور حركة الإصلاح التربوي في العالم وأهمية المساءلة التربوية من قبل النظام التعليمي لكل من المدرسة والمعلم ومسئوليتهما عن مستوى أداء الطالب وجودة تعلمه، أصبح النظام التعليمي بشكل عام والمدرسة بشكل خاص يتعرضان لضغوط شديدة لاستخدام نتائج تقويم أداء الطالب لتحديد الطلاب الذين لا يمتلكون الحد الأدنى من المهارات والمعارف العلمية الضرورية لمتابعة تحصيلهم من صف إلى آخر، أو من مرحلة إلى أخرى، أو بعد تخرجهم وانخراطهم في سوق العمل أو متابعة تحصيلهم في مؤسسات التعليم العالي ويتطلب ذلك بالطبع التحديد المسبق لمستويات الأداء المرغوبة ودرجات القطع المرتبطة بها، وذلك من خلال التوصل إلى درجات القطع المناسبة بالطرق العلمية التي تتميز بدرجة عالية من الصدق والثبات.

مصطلحات الدراسة:

- درجة القطع Cut Score: تعد مفهوما أساسيا في ضوء فلسفة القياس محكي المرجع الذي يقارن أداء الطالب بمواصفات قياسية محددة. وتعرف درجة القطع بأنها نقطة أو أكثر على مقياس الدرجات، تفصل بين مستويات الأداء المختلفة. لذلك يتطلب تحديدها تحديد المواصفات القياسية للأداء بحيث تتضمن مواصفات للمعارف والمهارات المطلوبة لتفسير مستويات أداء/ تحصيل الطلاب. ومن ثم تتطلب تحديد نوعين من المواصفات هما المواصفات الكيفية: مواصفات المحتوى Content standards وتحدد ما ينبغي أن يعرفه الطلاب ويستطيعون فعله. والمواصفات الكمية: مستويات الأداء Performance Standards وتحدد مقدار ما ينبغي أن يعرفه الطلاب (مستوى الأداء الذي يتوقع عنده أن يحقق الطلاب مواصفات المحتوى).
- وإجرائيا في الدراسة الحالية تم تحديد أربعة مستويات للأداء هي: "متعثر- نام- مرض- متقدم" تفصل بينها ثلاث درجات للقطع.

طريقة أنجوف المعدلة Modified Angoff Method: إحدى طرق تقدير درجة القطع، وفيها

يُطلب من كل محكم أن يتصور مجموعة من الأفراد الذين حققوا الحد الأدنى للكفاية التي يقيسها الاختبار، ثم يقدر نسبة عدد الأفراد الذين يُحتمل أن يجيبوا إجابة صحيحة على كل مفردة من مفرداته، ويمثل متوسط هذه النسب الحد الأدنى لمستوى الاجتياز في الاختبار.

وقد أجريت بعض التعديلات على طريقة أنجوف منها تبسيط مهمة المحكمين إلى أقل حد ممكن، وذلك بأن يطلب من كل منهم أن يبدى رأيه "بنعم" أو "لا" حول كل مفردة فيما يتعلق بما إذا كان الفرد ذو الحد الأدنى من الكفاية ينبغي عليه أن يجيب إجابة صحيحة على المفردة حتى يعد متمكناً ونظراً لوجود بديلين أمام المحكم فإن القيم الاحتمالية الناتجة تكون إما صفراً أو واحداً صحيحاً لكل مفردة.

وقد أجريت عدة تعديلات على طريقة أنجوف منها تزويد المحكمين ببعض البيانات التجريبية عن الأداء الفعلي للأفراد والمفردات. وإجرائيا تستخدم طريقة (نعم/ لا) لأنجوف المعدلة بتزويد المحكمين بالبيانات التجريبية لتحديد مستويات الأداء المتعددة المعروفة بطريقة توكر.

بيانات التغذية الراجعة Feedback Data: يقصد بها إجرائيا في الدراسة الحالية مجموعة البيانات التي يزود بها المحكمون بعد كل جولة من جولات التطبيق وتشمل:

- **بيانات الأثر Impact Data**: وتتمثل في النسب المئوية للطلاب الذين يقعون في كل مستوى من مستويات الأداء عند تطبيق درجات القطع التي يحددها المحكمون في جولتي التطبيق على بيانات الطلاب الفعلية الناتجة عن استجاباتهم على الاختبار.
- **البيانات الواقعية وتشمل:**

- **مستوى الصعوبة الكلاسيكي P-Value** الفعلي لكل مفردة كما يعبر عنه بعدد الأفراد الذين أجابوا صوابا عن المفردة.
- **التكرارات والنسب المئوية لاستجابات الأفراد الفعلية على كل بديل لكل مفردة من مفردات الاختبار** كما يعبر عنه بعدد الأفراد الذين يختارون كل بديل من بدائل السؤال وكذلك عدد الأفراد الذين تركوا السؤال دون إجابة والنسبة المئوية لكل بديل والإجابة المتروكة.
- **التحليل البياني لكل مفردة** Graphical of Item Analysis (GIA): وسيلة بصرية توضح العلاقة بين الدرجة الكلية على الاختبار ونسب الاستجابة للبديل الصحيح والبدائل غير الصحيحة للمفردات الاختبارية من نوع الاختيار من متعدد. وتوفر معلومات أساسية سهلة التفسير عن خصائص المفردة (الصعوبة، التمييز، التخمين). كما أن المفردات قليلة الجودة يمكن التعرف عليها بسهولة.

- **البيانات الجماعية:** وتتمثل في البيانات المتعلقة بدرجات القطع التي يحددها المحكمون بكل جولة وتشمل(درجة القطع لكل محكم، متوسط، وسيط، منوال، الانحراف المعياري، الدرجة العظمى، الدرجة الصغرى لدرجات القطع لكل المحكمين).

خبرة المحكم: ويقصد بها إجرائيا في الدراسة الحالية مجال عمل المحكم؛ حيث تضمنت الدراسة فئتين من المحكمين هما أعضاء هيئة التدريس بالجامعة، والمعلمون بوزارة التربية والتعليم.

الممارسات العملية للمحكمين: ما يمارسه المحكمون من عمليات تفكير حتى يصلون إلى الحكم على تقدير كل مفردة وفق إرشادات طريقة أنجوف، وما قد يطرأ على هذه العمليات من تغيير نتيجة لتقديم بيانات الأثر.

الإطار النظري:

١ الأنظمة المرجعية لتفسير الدرجات المستمدة من الاختبارات:

تتعدد الأنظمة المرجعية المستخدمة لتفسير الدرجات المستمدة من أدوات القياس النفسي والتربوي؛ ووفق أدبيات القياس فإن هناك أربعة أنظمة مرجعية رئيسة مؤثرة في تطوير أساليب التقويم وأدواته تتمثل في النظام جماعي/ معياري المرجع Norm- Referenced System ، النظام الذي ينسب أداء الفرد لنفسه Self- Referenced System ، النظام الذي ينسب أداء الفرد إلى محك Criterion-Referenced System ، والنظام الموضوعي القائم على نظرية الاستجابة للمفردة Objective System. وتتناول الباحثة فيما يلي النظامين المحكي المرجع والموضوعي؛ حيث تأتي الدراسة الحالية تطبيقا لهذين المدخلين.

١-١ النظام الذي ينسب أداء الفرد إلى محك (القياس محكي المرجع) Criterion-Referenced System

ظهرت الاختبارات محكية المرجع نظرا لما وجه إلى فلسفة القياس جماعي المرجع، وما يبنى عليها من اختبارات مقننة من أوجه نقد؛ مما حذا بالباحثين بالبحث عن نظام ينسب أداء الفرد إلى محك أداء متوقع. ويهتم بتفسير الدرجات المستمدة من الاختبارات والمقاييس في ضوء هذا المحك. ويصاغ هذا الأداء عادة على صورة **كفايات محددة أو نواتج متوقعة أو أهداف سلوكية مرتبة** بحيث **تصف مختلف مستويات الأداء.** ولا تستند مرجعية تفسير الدرجة في الاختبار أو المقياس مرجعي المحك إلى أداء الأقران أو معيار التقنين، وإنما إلى الأداء المتوقع أو المرجو تحققه والذي يحدد تحديداً دقيقاً من خلال توصيف مستويات الأداء كيفيا وكميا توصيفا دقيقا.

ويفترض في إطار هذا النظام وجود متصل لاكتساب المعارف والمهارات يُمثل أحد طرفيه

عدم الكفاءة، ويمثل الطرف الآخر الكفاءة التامة، ويقابل أداء الفرد في الاختبار مرجعي المحك إحدى نقاط هذا المتصل. حيث تحدد كل نقطة مستوى أداء أو مستوى معيناً من مستويات الكفاءة كما يقيسه الاختبار مرجعي المحك، ويقارن مستوى الأداء الفعلي للفرد في الاختبار بمستوى أداء مقبول محدد مسبقاً، ويمثل مدى هذا المتصل ما يستطيع الفرد أداؤه وما لا يستطيع، وبذلك توفر الاختبارات محكية المرجع معلومات محددة عن درجة الكفاءة التي حققها الفرد مستقلة عن أداء أقرانه.

ومن ثم يتطلب استخدام هذا المنحى في بناء الاختبارات البدء أولا بافتراض مُتصل للسمة المراد قياسها، وتحديد السلوك الذي يمكن أن يؤديه الفرد باستخدام الاختبار ثم تحديد مستوى الأداء المرغوب على الاختبار. وتُعد الدرجة التي يحصل عليها الفرد في الاختبار بمثابة قيمة تقديرية لمستوى أدائه أو كفاءته على هذا المتصل الافتراضي للسمة المقاسة، ثم يقارن مستوى الأداء الفعلي للفرد بمستوى الأداء المرغوب المحدد مسبقا (صلاح علام، ٢٠٠٧) .(Vukmirovic,2009)

ووفقا لدرجة تحديد النطاق السلوكي الذي يقيسه الاختبار تقسم الاختبارات محكية المرجع إلى ثلاثة أنواع تتمثل في: الاختبارات مرجعية الهدف Objective Referenced Tests، والاختبارات مرجعية النطاق Domain Referenced Tests، واختبارات التمكن أو الإتقان Mastery Tests.

١-١-١- استخدامات الاختبارات محكية المرجع:

ما إن استقرت فلسفة القياس محكي المرجع حتى توسعت تطبيقاتها لتشمل **تقويم الفرد أو مجموعة أفراد لتحديد مكانة أدائهم في مجال تعليمي أو تدريبي معين**، حيث تستخدم في تسكين الطلاب على متصل التعلم، وتشخيص تحصيل الطلاب، ومراقبة تقدمهم، وتشخيص تحصيل الصف المدرسي ومراقبة تقدم الصف المدرسي، وعقود الأداء. وامتدت لتشمل **تقويم المناهج والبرامج التعليمية والتدريبية لتقدير فاعليتها ومردودها** (صلاح علام، ٢٠٠٧).

١-٢ نشأة طرق تحديد مستويات الأداء ودرجات القطع:

حتى عهد قريب لم يكن هناك اهتمام جوهري بطرق وضع مستويات الأداء ودرجات القطع؛ حيث كان الاهتمام منصبا على النسبة المئوية الصحيحة المقبولة للدرجات، وكانت تتراوح بين ٦٠%، ٧٥% وأطلق عليها مصطلح الحد الأدنى للنجاح، لكن مع ظهور حركة الإصلاح التربوي وتنامى مفهوم المحاسبية التربوية من قبل النظم التعليمية المختلفة لكل من المدرسة والمعلم

ومسئولياتهما عن مستويات الطلاب وجودة تعلمهم. أصبحت النظم التعليمية عامة والمدارس خاصة تخضعان لضغوط كبيرة لتوظيف نتائج تقويم أداء الطلاب لتحديد الطلاب ذوي الحد الأدنى اللازم من المهارات والمعارف، وذلك حتى يتسنى متابعتهم من صف لآخر ومن مرحلة لأخرى، وكذلك تقديم الخدمة التربوية المناسبة لأولئك الذين لم يحققوا الحد الأدنى من المهارات والمعارف. ويتطلب ذلك بالضرورة التحديد المسبق لمستويات الأداء المستهدفة، ومن ثم درجات القطع المرتبطة بها.

١-٣ درجة القطع Cut Score

تعد درجة القطع مفهوما أساسيا في ضوء فلسفة القياس محكي المرجع الذي يقارن أداء الطالب بمواصفات قياسية محددة. وتعرف درجة القطع بأنها نقطة أو أكثر على مقياس الدرجات، تفصل بين مستويات الأداء المختلفة (Vukmirovic,2009). كما تعرف بأنها نقطة على متصل درجات الاختبار تستخدم لتقسيم الطلاب إلى مجموعتين (المتمكنين وغير المتمكنين) بمستويات كفاية مختلفة بالنسبة للأهداف التي يقيسها الاختبار، أي الدرجة التي يمكن أن تدل على الحد الأدنى للأداء المقبول لمهارة ما، والتي ينبغي أن يمتلكها الطالب كحد أدنى ليكون ناجحا أو متفوقا في هذه المهارة (صلاح علام، ١٩٩١)؛ (Shepard,1984). أو الدرجة التي ينبغي أن يحصل عليها الفرد في النطاق الشامل لمفردات الاختبار لكي يعد متقناً لمحتوى أو مهارة معينة. وأحياناً تشير درجة القطع إلى الحد الأدنى للأداء المقبول لكي يتمكن الفرد من أداء مهام تالية. كما تناولها "تيراتيرا" (Tiratira, 2009, P41) بأنها نقطة على متصل يعد عندها الحد الأدنى المقبول فضلاً عن التصنيفات الإيجابية والسلبية ، وعرفها (Klein et al., 2009, P163) بأنها نقطة على مقياس الدرجات الملاحظة تميز بين الأشخاص المتقنين وغير المتقنين. وعرفها (Biddle, 1993) على أنها درجة تستخدم للفصل بين الأشخاص الذين نجحوا في الاختبار والذين لم ينجحوا.

١-٤ متطلبات تحديد درجة القطع

يتطلب تحديد درجة القطع تحديد المواصفات القياسية للأداء **Performance Standards** بحيث تتضمن مواصفات للمعارف والمهارات المطلوبة لتفسير مستويات أداء/ تحصيل الطلاب. ومن ثم تتطلب تحديد نوعين من المواصفات هما: **المواصفات الكيفية:** مواصفات المحتوى Content standardsوتحدد ما ينبغي أن يعرفه الطلاب ويستطيعون فعله. **والمواصفات الكمية:** مستويات الأداء Performance / Standards وتحدد مقدار ما ينبغي أن يعرفه الطلاب (مستوى الأداء الذي يتوقع عنده أن يحقق الطلاب مواصفات المحتوى).

لذلك فإنه إجرائيا من الضروري تحديد مستويات الأداء (عددها – مسمياتها)، ووضع توصيف لهذه

المستويات، ثم يلي ذلك تحديد درجات القطع.

١-٥ المواصفات القياسية للأداء Performance Standards:

ترجم مصطلح Standards في الأدبيات العربية إلى مسميات متعددة منها "المستويات المعيارية"، "المعايير"، "المواصفات القياسية". وتتبنى الدراسة الحالية ترجمته تحت مسمى "المواصفات القياسية" وذلك تجنبا للخلط بين مفهوم المواصفات القياسية وبين مفهوم المعايير جماعية المرجع Norms.

وتعرف المواصفات القياسية للأداء بأنها مواصفات المعرفة والمهارات اللازمة لسمة ما لمستويات التحصيل الأكاديمي. مواصفات مستوى الأداء تلك عبارة عن المعرفة والمهارات الأساسية التي يجب على الطالب إتقانها عند كل مستوى أداء (إذا كان هناك أكثر من درجة قطع مطلوبة). والتي يقوم بوضعها عادة خبراء المحتوى العلمي للمقرر وصناع القرار التربوي (Perie,2008).

وتنقسم المواصفات القياسية للأداء إلى نوعين هما المواصفات القياسية الكيفية والمواصفات القياسية الكمية:

- **المواصفات القياسية للمحتوى (المواصفات القياسية الكيفية)** Content Standards وهي في طبيعتها مواصفات نوعية/ كيفية ترتبط بالإجابة على السؤال "ما" **"What "** وتعرف محتوى المعارف والمهارات المرغوبة وتتجسد إجرائيا في جدول مواصفات الاختبار؛ حيث تراعي ما يعرفه المتعلمون ويكونون قادرين على القيام به وترتبط بشكل مباشر بالمنهج والمحتوى.

- **المواصفات القياسية للتحصيل (المواصفات القياسية الكمية)** **Achievement Standards** هي مواصفات كمية في طبيعتها تتجسد في الإجابة على السؤال "ما كم"؟ "How Much"حيث تعرف / تصف مستوى الأداء Performance Level المتوقع أن يظهر فيه الطلاب مواصفات المحتوى المحددة. وتقاس إجرائيا من خلال متوسطات درجات القطع.

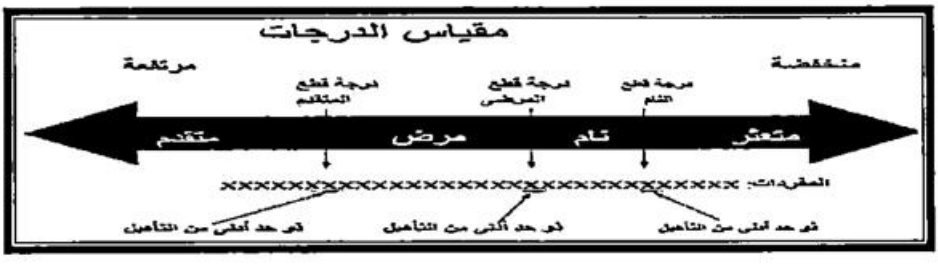
١-٦ تحديد / وضع المواصفات القياسية للأداء Standard Setting

تعرف عملية تحديد المواصفات القياسية بأنها اشتقاق مستويات الأداء في التقويمات التربوية أو المهنية والتي يتم بناء عليها اتخاذ القرارات أو تصنيفات الأفراد (Cizek,1996). وذلك من خلال تعريف مستويات التحصيل على صورة درجات الاختبار حيث تعتبر عملية إصدار أحكام كيفية وكمية. ويتجسد الشكل الإجرائي لمستويات التعلم Learning Standards في درجات الاختبار. وتوظف عملية تحديد درجات القطع لتصنيف نواتج تعلم الطلاب على هيئة مستويات الأداء.

١-٧ إجراءات تحديد المواصفات القياسية للأداء

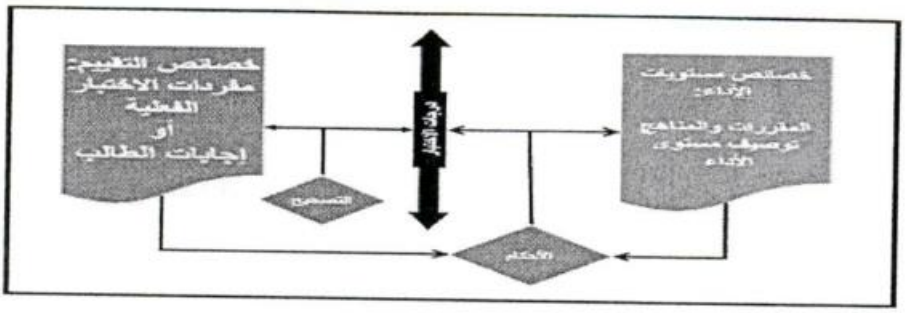
- **الإجراءات العامة** وفيها يتم بناء أو تطوير مواصفات مستويات الأداء العامة والخاصة، وتحديد المشاركين، والمواد والإجراءات اللوجستية.
- **التأطير لمستويات الأداء** وفيها يتم تحديد عدد مستويات الأداء، واتخاذ القرارات حول مسميات مستويات الأداء Performance Levels Labels وبناء مواصفات مستويات الأداء Performance Levels Descriptors(PLDs). ويختلف عدد مستويات الأداء باختلاف الهدف من الاختبار ففي حالات منح الرخصة للمعلمين فإنه يمكن تحديد درجة قطع واحدة للتعرف على المعلمين الأكفاء / المؤهلين، في حين قد تحتاج الاختبارات التحصيلية إلى عدد من مستويات الأداء لوضع الطلاب في فئات مثل: مقبول، كفء، متقدم. ومن الضروري أن تبنى هذه المواصفات القياسية على توصيف واضح لمستويات الأداء من خلال توصيف مستويات الأداء باستخدام عبارات وصفية كاملة يستخدمها المحكم لتحديد ما إذا كان الأفراد قادرون على تحقيقها أم لا، ومن ثم يتحدد في ضوئها درجة القطع. وبصفة أساسية يتضمن وضع المواصفات القياسية للأداء الأحكام حول مواصفات الأداء المطلوب عند كل مستوى من مستويات الأداء ودرجات الاختبار التي تعكس هذه المواصفات (Heiach,2013,492).
- **تحديد درجات القطع:** وتشمل الاختيار من بين الطرق المختلفة ووضع محكات التقييم وإجراءات قبول درجات القطع.

ويوضح شكل (١) التالي متصلا للتحصيل الدراسي تتدرج عليه درجات الاختبار من الدرجات المنخفضة إلى الدرجات المرتفعة. هذا المتصل مقسم إلى ٤ مستويات (متعثر- نام- مرض- متقدم) وذلك باستخدام ثلاث درجات قطع؛ حيث تحدد كل درجة قطع منها كم المفردات التي يمكن أن يجيب عليها الطالب ذو الحد الأدنى من الأداء إجابة صحيحة لكل مستوى.



شكل (١): أربعة مستويات للأداء ثلاث درجات قطع

ويتم ذلك من خلال اختبار الطلاب باختبار يقيس مهارات ومعارف محددة يتم تصحيحها وتحويلها إلى درجات على الاختبار، ويتم الحكم على هذه الدرجات في ضوء مواصفات مستويات الأداء المحددة والمقررات والمناهج المستخدمة كما هو موضح بشكل (٢) التالي.



شكل (٢): مخطط توضيحي لعملية تحديد درجة القطع

١-٨ طرق تحديد مستويات الأداء ودرجات القطع:

نظراً لأهمية "مستويات الأداء" في اتخاذ القرارات التعليمية المتعلقة بالأفراد، فقد اهتم علماء القياس التربوي باقتراح عدد من الطرق التي يمكن الاسترشاد بها في تحديد هذه المستويات، وأصبحت في الآونة الأخيرة مجالاً خصباً للدراسات الإمبريقية التي هدفت للمقارنة بينها في مواقف اختبارية متنوعة لتحديد أكثرها اتساقاً ودقة (علام، ١٩٨٥، ١٩٩١)، (Jaeger, 1994; Koffler, 1980).

وتتعدد الأسس التي يتم في ضوئها تصنيف طرق تحديد مستويات الأداء فقط صنفها "مسكاوسكاس" إلى نماذج منفصلة (استاتيكية) State Models تقوم على ثنائية التصنيف لقدرة المتعلم، والنماذج المتصلة Continuum Models حيث تعتبر قدرة الطالب/ الفرد متدرجة على متصل؛ ومن ثم حظيت هذه النماذج باستخدام أوسع.

وتفترض بعض هذه الطرق أن التمكن أو الكفاية تعد "حالة"، أي أن الفرد إما أن يكون متمكناً أو غير متمكن. في حين يفترض آخرون أن التمكن أو الكفاية تتوزع توزيعاً متصلاً، أي أن التمكن ليس نقطة على المتصل، وإنما يكون محدوداً بمدى على هذا المتصل.

وتعتمد الطرق المتصلة على اختلاف تصنيفاتها على وجود العنصر البشرى في كافة إجراءات تحديد مستويات الأداء في حين تقوم النماذج المنفصلة على النماذج الرياضية والإحصائية كنموذج "بيز" Baysan Model ونموذج ذي الحدين.

كما تصنف طرق تحديد مستويات الأداء ودرجات القطع أيضا إلى طرق تحكيمية، وطرق تعتمد على التحكيم وتسترشد بالبيانات التجريبية، وطرق تعتمد على التجريب وتسترشد بالتحكيم، ويمكن تفصيل هذه الطرق كما يلي:

١-٨-١ **الطرق التحكيمية/ الطرق المطلقة Judgmental Methods**: تقوم بصفة أساسية على أحكام الخبراء المتخصصين في مجال القياس أو المتخصصين في محتوى الاختبار المراد قياسه، سواء أكانت أحكاماً فردية أم جماعية. وتتعلق هذه الأحكام بمحتوى الاختبار ومفرداته دون الاسترشاد ببيانات تجريبية مستمدة من التطبيق الميداني للاختبار. لذلك يُطلق عليها "الطرق المطلقة Absolute Methods. حيث يسترشد بآرائهم في تحديد مستوى الطالب ذو الحد الأدنى من الكفاية؛ ولتحديد عدد الأسئلة التي يجب أن يجيب عليها هذا الطالب الحدي إجابة صحيحة.

ويتطلب تحديد مستويات الأداء في إطار هذه الطرق بالضرورة أحكاماً قيمية تأخذ بعين الاعتبار الهدف من الاختبار، ومحتواه، وطبيعة السمة أو القدرة أو الكفاية التي يقيسها الاختبار.

ومن الطرق شائعة الاستخدام طريقة ندلسكاى (Nedlesky, 1954، وطريقة أنجوف Angoff, 1961، وطريقة إيبل Ebel, 1962، وطريقة جيجر Jaeger, 1982).

وتتمثل أهم **مميزات** الطرق التحكيمية في سهولة استخدامها وتطبيقها وتفسيرها، كما تتميز بسهولة الحساب نظرا لعدم اعتمادها على أساليب إحصائية معقدة، كما أنها لا تحتاج إلى جهد ووقت كبيرين في تطبيقها.

وتتمثل أهم **محددات** الطرق التحكيمية في اعتمادها كلياً على آراء المحكمين؛ حيث تلعب كفاية هؤلاء المحكمين دوراً بارزاً في الوصول إلى المحك المناسب لتصنيف الطلاب وتحديد مفردات الاختبار التي يمكن أن يجيب عليها الطالب الحدي صوابا، وصعوبة تحديد الطلاب الأقل كفاءة، وعدم اعتمادها على بيانات تجريبية فعلية لكي يسترشد بها المحكمون في تقدير أحكامهم، وكذلك الحصول على درجات قطع مختلفة عند تطبيقها على اختبار موحد وعينة دراسية واحدة، مع ذاتية المحكمين في اختيار الحد الأدنى من الكفاية؛ بسبب اختلاف تصوراتهم حول عملية الإتقان، وتفاوت عملية الاتساق الداخلي في عملية التحكيم، وصعوبة التنبؤ بأداء الطلاب في الاختبار.

وتقدم الأدبيات في مجال القياس بعض المعايير للحد من تأثير محددات الطرق التحكيمية، منها استخدام طرق لتوحيد بيانات الأداء الفعلي. ومراعاة أن تكون عملية تحكيم مفردات الاختبار متزامنة مع تطبيق الاختبار، والاهتمام بتدريب المحكمين، وتنظيم عملية تحكيم بنود الاختبار، وإيجاد الحلول المناسبة لتسوية الاختلافات في أحكام المحكمين. وتصميم استمارات تنظم عملية التحكيم، وتزويد المحكمين ببيانات مسبقة حول أداء الطلاب على الاختبار (صلاح علام، ٢٠٠٧)(Vukmirovic,2006)(Cizek,2012).

١-٨-٢ طرق تعتمد جزئياً على التحكيم وتسترشد ببيانات تجريبية/ الطرق النسبية Judgmental Empirical Methods

أدرك الباحثون وجود أوجه قصور ملحوظة في الطرق التحكيمية، منها اعتمادها على تحكيم مفردات الاختبار دون النظر إلى الأداء الفعلي للمختبرين مما قد يؤدي إلى الحصول على معايير تفتقر إلى درجة عالية من الصدق (صلاح علام،١٩٩٥، ٢٤٢-٢٤١). لذا اقترح علماء القياس عددا من الطرق التجريبية لتحديد مستويات الأداء للحد من أوجه قصور الطرق التحكيمية.

وتعتمد هذه المجموعة من الطرق أيضاً على أحكام الخبراء مع تقديم بيانات تتعلق بالأداء الفعلي للمختبرين لكي يسترشد بها هؤلاء الخبراء في أحكامهم.

فالبيانات المستمدة من تطبيق الاختبار على عينة مناسبة من الأفراد تجعل أحكام الخبراء أكثر واقعية. ولذلك يطلق عليها الطرق النسبية Relative Methods. ومن أهم هذه الطرق **طريقة التحكيم المعززة بالمعلومات** Informed Judgment (Popham, 1983)، وطريقة أنجوف المعدلة (Berk, 1984)، وطريقة توفق بين الطرق المطلقة والطرق النسبية (Beuk, 1984).

١-٨-٣ طرق تعتمد على البيانات التجريبية، وتسترشد بالتحكيم Empirical Judgmental Methods:

تعتمد هذه المجموعة من الطرق اعتماداً أساسياً على البيانات التجريبية المستمدة من تطبيق الاختبار على عينة أو أكثر من الأفراد، وتحليل هذه البيانات باستخدام الأساليب الإحصائية. غير أن التعريف الإجرائي لمحكات تصنيف الأفراد يتحدد استناداً إلى أحكام الخبراء في ضوء الأداء الفعلي لمجموعة متمكنة من النطاق السلوكي الذي يقيسه الاختبار.

حيث تتعلق عملية التحكيم في هذه الطرق بانتقاء الطلاب المتمكنين وغير المتمكنين في

الفصول المدرسية أو في مواقع العمل وليس بتحكيم محتوى الاختبار ومفرداته. وتعد هذه الطرق أكثر توجهاً نحو الجوانب الكمية، حيث تسترشد بالنماذج الرياضية والإحصائية. وتعتمد مصداقيتها على مدى الثقة في قدرة الخبراء المحكمين على الفصل بين الأفراد المختبرين تبعاً لمستوى تمكنهم فيما يقيسه الاختبار. ومن أهم هذه الطرق: طريقة المجموعات المحكة Criterion Groups وطريقة المجموعات الحدية Borderline Groups، وطريقة المجموعات المتناقضة (صلاح علام، ٢٠٠٧)(Vukmirovic,2006)(Cizek,2012).

ونظرا لأن طريقة أنجوف محل اهتمام الدراسة الحالية لذا تتناولها الباحثة بشيء من التفصيل.

٩-١ طريقة أنجوف Angoff's Method

في هذه الطريقة يُطلب من كل محكم أن يتصور مجموعة من الأفراد الذين حققوا الحد الأدنى للكفاية التي يقيسها الاختبار، ثم يقدر نسبة عدد الأفراد الذين يُحتمل أن يجيبوا إجابة صحيحة على كل مفردة من مفرداته، ويمثل متوسط هذه النسب الحد الأدنى لمستوى الاجتياز في الاختبار. وتتميز هذه الطريقة باعتمادها على عدد من الأفراد الذين حققوا الحد الأدنى للكفاية، كما يسهل على مجموعة المحكمين فهمها واستخدامها. غير أنه ربما يصعب على المحكمين تعريف الفرد الذي حقق الحد الأدنى للكفاية المطلوبة، وكذلك تقدير القيم الاحتمالية لبعض المفردات (Shepard,1984).

وقد أجرى أنجوف (١٩٧٦) تعديلات على طريقته بحيث تُيسر تقدير القيم الاحتمالية لمفردات الاختيار من متعدد. لذلك يمكن تصميم استمارة تشتمل على التعليمات، وأرقام مفردات الاختبار، وتوضع أمام كل منها الحرف المناظر للإجابة الصحيحة لمفردة داخل مربع. وكذلك ميزان يشتمل على عشر فترات متساوية تقريباً. ثم يطلب من كل محكم تقدير احتمال أن يعرف الفرد الذي يمتلك الحد الأدنى من الكفاية المطلوبة إجابة المفردة دون أن يلجأ إلى التخمين، وأن يضع دائرة حول الرقم الذي يحتوي هذه القيمة على الميزان المُعطى. أي أن هذه الطريقة تتطلب من كل محكم التنبؤ بمستوى صعوبة كل مفردة في الاختبار لمثل هذا الفرد. وإذا وجد اختلاف في القيم التقديرية بين المحكمين تساوى أو تزيد عن ٠.٢٠ لأي مفردة يمكن أن يُطلب من المحكم الذي أعطى أعلى تقدير والذي أعطى أقل تقدير تفسير وجهة نظره وتعديل التقدير إذا رأى ذلك.

ثم تجمع القيم الاحتمالية التي قدرها كل محكم لكل مفردة من مفردات الاختبار. وتُمثل القيم الناتجة الحد الأدنى لمستوى الأداء في الاختبار من وجهة نظر كل منهم، ومن ثم يتم الحصول على

متوسط هذه القيم لجميع المحكمين ولجميع مفردات الاختبار.

١-١٠ التعديلات التي أجريت على طريقة أنجوف:

وحيث إن أنجوف لم يقدم أي تفاصيل حول طريقة التطبيق لطريقة تحديد درجة القطع أو مستوى الأداء المطلوب عمليا. كما لم يذكر على نحو التفصيل كيفية اختيار المحكمين وطريقة تدريبهم على تطبيق طريقة حساب درجة القطع، أو آلية مناقشة أحكام المحكمين ومراجعة تلك الأحكام أو منحهم مفاتيح الإجابات الصحيحة للفقرات التي يحكمون عليها، فقد أجريت تعديلات عديدة على طريقة أنجوف كتقديم بيانات للمحكمين عن أداء الأفراد على الاختبار وأطلق عليها طريقة أنجوف المعدلة Modified Angoff Method (Cizek,2001).

١-١٠-١ طريقة أنجوف التجريبية المعدلة Modified Angoff Empirical Method

تعتبر هذه الطريقة تعديل لطريقة أنجوف التحكيمية، وتتطلب تزويد المحكمين ببعض البيانات المستمدة من التجريب الميداني لمفردات الاختبار، للاسترشاد بها في عملية التحكيم، مثل طريقة أنجوف التحكيمية فإنه يطلب من كل محكم تقدير احتمال إجابة الأفراد ذوي الحد الأدنى من الكفاية المطلوبة إجابة صحيحة على كل مفردة من مفردات الاختبار، وبعد الحصول على هذه التقديرات يتم تزويد المحكمين ببيانات تجريبية عن مفردات الاختبار مثل متوسط صعوبة كل مفردة، ثم يطلب منهم مراجعة تقديراتهم الاحتمالية في ضوء تلك البيانات التجريبية، ويكون الحد الأدنى لمستوى الاجتياز في الاختبار هو متوسط مجموع تقديراتهم المعدلة .ومن مميزات هذه الطريقة سهولة التطبيق والفهم وسهولة حسابها وتعمل على ربط المحك بالكفايات المقاسة، والاستعانة بمعامل صعوبة المفردة الحقيقي خلال عمليات التحكيم (Berk , 1986) (صلاح علام،٢٠٠٧).

وأضاف جيجر Jaeger بأنه أجريت تعديلات عام ١٩٧٨م على طريقة أنجوف الأساسية، وذلك بهدف تبسيط مهمة المحكم بحيث يطلب من كل محكم تقييم كل مفردة من مفردات الاختبار بالإجابة بنعم أو لا حول كل مفردة فيما يتعلق بما إذا كان الفرد ذو الحد الأدنى من الكفاية ينبغي عليه أن يجيب إجابة صحيحة على المفردة حتى يعد متمكناً وإلا اعتبر غير متمكن. ونظراً لوجود بديلين أمام المحكم هما "نعم"، "لا" فإن القيم الاحتمالية الناتجة تكون إما صفراً أو واحداً صحيحاً لمفردة معينة. ومن ثم يتم تحديد الحد الأدنى للكفاية، وبعد الحصول على هذه التقديرات يستعرض عدد مختلف من المحكمين عدة أنواع من البيانات التجريبية، خلال ثلاثة لقاءات منفصلة، وتشتمل هذه البيانات التجريبية على التوزيع التكراري لدرجات الطلاب في

الاختبار والإحصاءات الوصفية لدرجات القطع المحددة من قبل المحكمين، ثم يطلب من عينات المحكمين المختلفة تعديل تقديراتهم بناء على تلك البيانات التجريبية، وتصبح درجة القطع عبارة عن متوسط تلك التقديرات التي حددتها عينات المحكمين في اللقاءات الثلاثة المنفصلة. ومن مميزات هذه الطريقة أنها تتيح للمحكمين فرصة لتحسين تقديراتهم بناء على ثلاثة أنواع من البيانات المختلفة، ويحدد درجة القطع عينات مختلفة من المحكمين وتميل هذه الطريقة إلى زيادة ثبات درجة القطع، ويؤخذ على هذه الطريقة اقتصار الإجابة على كل سؤال بنعم أو لا، بالإضافة إلى أن هذه الطريقة تفرض على المحكمين أن يلتقوا ثلاثة لقاءات منفصلة، كما أنها تستغرق كثيراً من الوقت. وتعتبر هذه الطريقة أكثر تعقيداً من الطرق السابقة، ويعتبر تفسير المحك النهائي صعب جداً بالنسبة للمحكمين. ولتلافي بعض من هذه العيوب قام كل من سنديرس ومابوس بإجراء تعديلات على هذه الطريقة فأصبحت تحتوى على عينة واحدة من المحكمين، وأصبح تحديد المحك النهائي يتم خلال لقاء واحد يجمع بين المحكمين (Berk, 1986). كما قام شيبرد أيضاً بإجراء تعديلات على طريقة أنجوف الأساسية بحيث أصبحت هذه الطريقة تتطلب أربعة أنواع من المعطيات تتمثل في:

- الأحكام المطلقة القائمة على فحص المفردات مثل طريقة أنجوف .
- بيانات تجريبية على عينة محكمة صنفت المتقنين وغير المتقنين مثل طريقة المجموعات المتضادة.
- تقديرات المحكمين لمستوى النجاح المقبول.
- التوفيق بين تقديرات المحكمينّ المختلفة.

ومن أهم مميزات هذه الطريقة أنها تضم أربعة أنواع من البيانات التجريبية والتحكيمية، كما يراعى فيها المصادر الرئيسة للقيام بعملية تحديد المواصفات القياسية للأداء، كما تحتفظ بمميزات طريقتي أنجوف والمجموعات المتضادة؛ إلا أن لها نفس أوجه قصور طريقتي أنجوف والمجموعات المتضادة (Berk, 1986).

١-١١ دور المحكمين في تحديد درجات القطع:

عادة ما يتم الاستعانة بعدد من المحكمين من الخبراء في المادة العلمية ومحتوى المناهج الدراسية وخبراء في القياس والتقويم التربوي والمعلمين كمحكمين لتحديد مستويات الأداء ودرجات القطع في مادة دراسية تكون في صميم تخصصهم وخبرتهم.

وتفاديا للمشكلات التي قد تنجم عن تحديد درجات القطع في تقديرات المحكمين ولزيادة ثبات تلك التقديرات وتقليل الخطأ المعياري. وبالتالي رفع مستوى صدق الاستدلالات الناتجة عنها

يؤكد بليك (Plake,2008) على ضرورة التحقق من حسن اختيار المحكم عبر مجموعة عناصر أساسية تلعب دورا مهما في تحديد درجة القطع مثل:

- سنوات الخبرة في التدريس، وخاصة مدى انخراط المحكم في العملية التعليمية بصفة مستمرة مع الطلاب، حيث يجعله ذلك أكثر معرفة ودقة في تقديره لمستويات الأداء ودرجات القطع؛ لأن المحكم ذا الخبرة التعليمية القليلة تتأثر تقديراته (ارتفاعا أو انخفاضا) بالتغذية الراجعة التي يحصل عليها بين جلسات التحكيم. وذلك أكثر من المحكمين الذين لديهم خبرة تعليمية واسعة.
- إشراك محكمين من خلفيات معرفية وخبرات متنوعة، مثل المعلمين المتخصصين التربويين ممن لهم دور في تحصيل الطالب.
- إلمام المحكم بالمحتوى الدراسي المستهدف تحديد درجة قطع ومستوى أداء له، حيث يعد أهم عنصر في اختيار المحكم لعملية التحكيم.
- يحتاج المحكمون إلى جلستي تحكيم على الأقل، سواء مع التغذية الراجعة بين الجلسات أو من دونها. والسؤال المهم الذي يطرح نفسه كثيرا على من يتولى مهمة تدريب المحكمين على تحديد درجات القطع هو: هل يحق للمدربين تعديل درجات القطع ومستويات الأداء التي حددها المحكمون بعد التدريب والمداولات والتغذية الراجعة بين جلسات التحكيم؟ وأجاب عن هذا ليتز ووي (Lissitz & Wei,2008) بطريقة علمية؛ حيث أجازا ذلك ولكن من خلال قنوات علمية وإجراءات رسمية رصينة. طبق الباحثان طريقتين للوصول إلى تعديل مقبول من جميع الأطراف التربوية لدرجات القطع. الطريقة الأولى تتم عبر اطلاع اللجان التربوية المسئولة عن متابعة موضوع درجات القطع في المدارس، واعتمادها على درجات القطع ومستويات الأداء التي توصل لها الباحثان خلال جلسات التحكيم مع المعلمين، وتوضيح الرتبة المئينية لكل منها التي تقابل درجة قطع محددة (على اعتبار أن هناك مستويات متعددة لدرجات القطع). أي أن نسبة الطلاب الناجحين في المدرسة إذا تم اعتماد درجة قطع معينة تقابل مستوى أداء محدد، يمثل المعارف والمهارات المطلوب من الطالب إتقانها.عندئذ يتم الاتفاق بين الباحث واللجنة المذكورة على اعتماد ما اقترحه الباحث. أما الطريقة الثانية فهي تعتمد على مجموعة من الأساليب الإحصائية وعلى سياسة تربوية معينة يريد المختصون التربويون تحقيقها بالنسبة لمستويات الأداء الوطنية.

- ومن المشكلات التي يواجهها الباحثون عادة عند تدريب المحكمين على تحديد درجات القطع ومستوى الأداء، مشكلة عدم قدرة بعض المحكمين على التمييز بين المحتوى الجوهري والمحتوى غير الجوهري المرتبط بالمهارات الأساسية التي يجب على الطالب إتقانها. والمشكلة الأخرى هي تحديد درجات قطع متأثرة بما يعرف إحصائيا بالنزعة إلى الوسط مما يجعل التوصل إلى

استدلالات صادقة من درجات القطع تلك عملية صعبة.

- وجد كارانتونيس وسيرسي: (Karantonis& Sireci ,2006) أن استخدام الوسيط بدلا من المتوسط في تحديد درجات القطع ومستوى الأداء لإتقان الطالب يمكن أن يكون أفضل، لأنه يقلل من تأثير التقديرات المتطرفة التي يحددها بعض المحكمين ، ومدى تأثيرها في رفع أو خفض درجة القطع إذا ما استخدم المتوسط في حسابها؛ حيث إنه إحصائيا أن قيمة المتوسط تتأثر كثيرا بالقيم المتطرفة. وأضافا أيضا ، أنه من الجانب الآخر يجب دعم جلسات تدريب المحكمين بما يعرف بمواصفات مستويات الأداء. للحصول على تقديرات دقيقة لدرجات القطع ومستويات الأداء

١-١٢ بعض القضايا المرتبطة بتقدير درجة القطع: هناك عدد من القضايا المرتبطة بتقدير درجة القطع ما زالت في حاجة إلى البحث من أهمها:

- عدم وضوح مفهوم الطالب ذي الحد الأدنى من الكفاية / الأداء لعدد كبير من المحكمين؛ حيث إن الخبراء المؤهلين لتقدير درجة القطع لاختبار ما غالبا ما يفكرون بالطالب المتوسط أو فوق المتوسط أكثر من تفكيرهم في الطالب الحدي أي الذي يمتلك الحد الأدنى من الكفاية؛ حيث يميل أغلبهم إلى تصنيف معظم الطلاب في هذين المستويين. ومن ثم فإن المحكم الذي لا يمكنه تصنيف فرد ما في المستوى الحدي فإنه يلجأ إلى تصنيفه في مستوى أعلى، مما قد يسفر عنه التوصل إلى درجة قطع مرتفعة. ويمكن التغلب على هذه المشكلة من خلال التدريب الجيد للمحكمين على مفهوم الحد الأدنى من الكفاية قبل عملية التقدير الفعلية والتحقق من فهمهم الواضح لها.

- مدى إمكانية تزويد المحكمين بالمؤشرات الإحصائية عن الفقرات والمفحوصين: حيث إن أحد الافتراضات المهمة لدرجة القطع هو أن هذه الدرجة تمثل الحد الأدنى من الكفاية في موضوع ما بشكل عام دون الاهتمام بخصائص مجموعة معينة من الأفراد، ولكن في حال إطلاع المحكمين على المؤشرات الإحصائية للفقرات أو معلومات عن المفحوصين وأخذ المحكمين لهذه المعلومات بعين الاعتبار من شأنه أن يؤثر في تقديراتهم. أي أن عاملا أو أكثر من العوامل جماعية المرجع قد يؤثر في تقدير درجة القطع. الأمر الذي يستدعي إجراء مزيد من الدراسات التجريبية.

- أي درجات القطع أفضل؟ في الحقيقة لا يوجد اتفاق على أن درجة القطع الأعلى هي الأفضل أو العكس ، فقد تكون درجة القطع المناسبة لقيادة السيارة منخفضة بالنسبة لقيادة الطائرة، كما أن درجة القطع في اللغة الأجنبية قد تكون كافية لكنها تكون منخفضة تماما بالنسبة للغة الأم.

فهناك عدد من الاعتبارات التي تؤخذ بالحسبان عند تحديد مستوى التمكن المطلوب تتعلق بالنتائج المترتبة على نوعية وكفاية الفئة التي يمكن اعتبارها متمكنة (Bowers & Shindoll,1989).

١-١٣ أسس تقويم طرق تحديد درجة القطع: تتعدد معايير تقويم طرق تحديد درجة القطع لتشمل محكات إجرائية وداخلية وخارجية ومنها (صلاح علام، ٢٠٠٧) (Hambleton & Pitoniak, 2006):

- **معيار الكفاية الفنية:** بمعنى أن تكون الطريقة قادرة على تصنيف المختبرين بشكل مناسب، مع حساسية الطريقة لأداء المختبرين وصعوبة المفردات وتمييزها، وكذلك حساسيتها لعمليات التعليم والتدريب. وأن تتضمن الطريقة طرقا إحصائية مناسبة تفسر بشكل صحيح، مع مراعاة أخطاء القياس. وتثبت صدق القرار.

معيار الكفاية العملية: وتشمل سهولة تطبيق الطريقة وسهولة استخدام الطرق الإحصائية المرتبطة بها وقابليتها للتفسير والفهم، مع الموثوقية والإقناع. بالإضافة إلى أهمية القرارات المبنية على درجة القطع، ومقدار الوقت المتاح لتحديد درجة القطع، ومدى توافر المصادر المادية والطاقة البشرية لإنجاز المهمة مع كفاية المحكمين.

١-١٤ بيانات التغذية الراجعة Feedback Data

ووفقا لسيزك وبونش (Cizek&Bunch,2007) يمكن تصنيف التغذية الراجعة إلى ثلاثة أنواع تتضمن:

١-**بيانات جماعية / معيارية المرجع** Normative Data: وتتضمن تقديم درجات القطع العامة بالنسبة لكل المشاركين، درجات القطع المتطرفة، الانحراف المعياري، المتوسط، الوسيط لدرجات القطع. كما يعبر عنها متوسط أو وسيط كافة درجات القطع بعد استبعاد القيم المتطرفة.

٢- **بيانات الأثر** Impact Data: توضح توزيع الدرجات الخام للممتحنين أو النسبة المئوية للمختبرين في كل فئة وفي كل مستوى من مستويات الأداء. وتمثل بيانات الأثر بصفة أساسية مجموعة من البيانات التجريبية الناتجة من تطبيق الاختبار المراد تحديد مستويات الأداء ودرجة القطع للطلاب عليه بعد تطبيقه على عينة مناسبة من الطلاب. ويرجع استخدام بيانات الأثر بداية إلى جهود بعض العلماء في تطوير الطرق التحكيمية ومنها طريقة أنجوف من خلال الربط بين أحكام المحكمين المطلقة والواقع الفعلي للمتعلمين. وبدأت في البداية بتزويد المحكمين ببعض-البيانات عن تصنيف الطلاب في مستويات الأداء المختلفة عند تطبيق درجة القطع على استجابات الطلاب الفعلية.

٣-**البيانات الواقعية:** Reality Information : حاول بعض الباحثين تطوير بيانات التغذية

الراجعة من خلال تزويد المحكمين ببعض المؤشرات الإحصائية للمفردات مثل صعوبة المفردة، والنسب المئوية لكل بديل من بدائل الإجابة في حالة أسئلة الاختيار من متعدد، وتختبر الدراسة الحالية أيضا بالإضافة إلى هذه العوامل تأثير التحليل البياني للمفردات. وفيما يلي نبذة مختصرة عن كل منها.

صعوبة المفردة Item Difficulty: كلاسيكيا تشير صعوبة المفردة من نوع الاختيار من متعدد إلى نسبة الطلاب الذين أجابوا صوابا على المفردة صلاح علام، ٢٠٠٠).

تكرارات البدائل والنسب المئوية لاختيار كل بديل: وذلك بالنسبة لمفردات الاختيار من متعدد، حيث يتم حساب النسبة المئوية للطلاب الذين يختارون كل بديل، وكذلك يمكن حساب نسبة الطلاب الذين تركوا السؤال دون اختيار أي بديل. وتتخذ من النسبة المئوية لكل بديل مؤشرا على تمييز السؤال وصعوبته. حيث تشير النسبة المئوية للاستجابة الصحيحة على مستوى صعوبة المفردة. والمفردة الجيدة يكون فيها البديل الصحيح أكثر البدائل اختيارا، ويكون ارتباطه موجبا مع الدرجة الكلية للاختبار في حين يكون ارتباط البدائل غير الصحيحة بالدرجة الكلية سالبا. كما أن كل بديل يجذب نسبة من الطلاب بحيث لا يكون أي بديل منها مهملا. كما يكون من الضروري ألا تزيد نسبة الاستجابة المتروكة عن ٠.٢ من النسبة الكلية للاستجابات (Vukmirovic,2009).

- التحليل البياني للمفردات Graphical of Item Analysis (GIA)

يرجع الفضل لإرساء طريقة التحليل البياني للمفردات إلى Batenburg & Laros (2000). واستخدم التحليل البياني للمفردات في سياق الاختبارات القومية الكبرى.

ويمثل التحليل البياني للمفردات وسيلة بصرية توضح العلاقة بين الدرجة الكلية للاختبار ونسب الاستجابة للبديل الصحيح والبدائل غير الصحيحة للمفردات الاختبارية من نوع الاختيار من متعدد. وتوفر معلومات أساسية سهلة التفسير عن خصائص المفردة (الصعوبة، التمييز، التخمين). كما أن المفردات قليلة الجودة يمكن التعرف عليها بسهولة.

وبصفة عامة فإن للتحليل البياني للمفردات اثنين من التطبيقات المهمة تتمثل في: إمكانية أن يستخدمها الباحثون في عملية تحديد المفردات التي يجب استبعادها من التحليلات التالية أو تعديلها. وكذلك إمكانية استخدامها من قبل معد الاختبار في تحسين جودة مفردات الاختبار.

بالإضافة إلى أنها تمكن من الفهم الأفضل لنظرية الاختبار وبنائه؛ حيث يمكن أن تسهم في تقليل

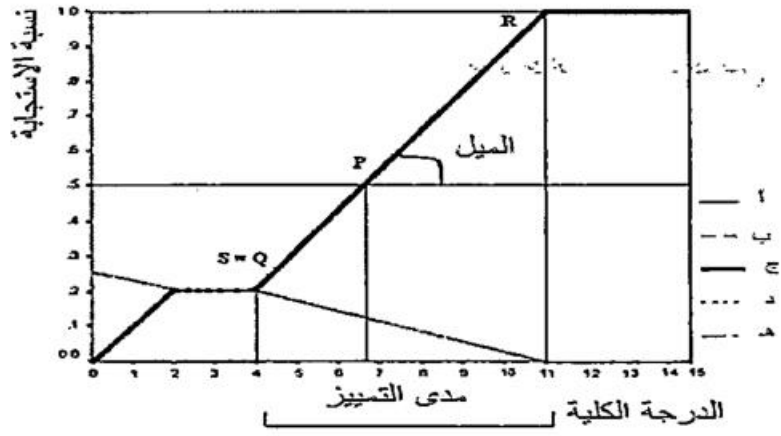
الفجوة بين عالم القياس المجرد والواقع العملي لبناء الاختبارات التحصيلية.

ويسعى معدو الاختبارات إلى بناء مفردات الاختيار من متعدد عالية الجودة. وعادة ما يعبر عن جودة المفردة من خلال قوة تمييزها. وبمعنى آخر المدى الذي يمكن أن تميز المفردة خلاله بين الطلاب ذوي القدرات المختلفة. وفي إطار النظرية الكلاسيكية فإن قوة تمييز المفردة Discrimination Power of an Item عادة ما يتم الحكم عليها باستخدام معامل الارتباط الثنائي الأصيل Point Biserial Correlation Coefficient بين درجة المفردة والدرجة الكلية للاختبار. ومن الضروري أن تكون قيمة هذا المعامل موجبة لتشير إلى أن الطلاب الذين يجيبون على المفردة صوابا غالبا ما يحصلون على درجة كلية أعلى. كما أن هذا المعامل يمكن حسابه لكل بديل من البدائل غير الصحيحة أيضا. وهذه من الضروري أن تكون قيمته سالبة، حيث تعني أن الطلاب الذين يختارون البدائل غير الصحيحة يحصلوا في الغالب على درجات كلية أقل. مما يشير إلى أن الطلاب الذين يجيبون صوابا على عدد أكبر من المفردات يعرفون أكثر من أولئك الذين يجيبون صوابا على عدد أقل من المفردات.

وعلى ذلك فإنه في التحليل البياني للمفردات GIA فإن نسبة الطلاب الذين يختارون الإجابة الصحيحة من المفترض أن يزداد مع زيادة الدرجة الكلية ونسبة الطلاب الذين يختارون البدائل غير الصحيحة (المشتتات) من المفترض أن تتناقص مع زيادة الدرجة الكلية. وبمقدار السرعة في زيادة نسبة الطلاب الذي يجيبون صوابا بزيادة الدرجة الكلية تكون الجودة في قوة تمييز المفردة. وعلى الجانب الآخر فإن المفردات قليلة الجودة تظهر انخفاضا في نسبة الطلاب الذين يختارون الإجابة الصحيحة بزيادة الدرجة الكلية.

نموذج نظري لمفردة جيدة وفقا للتحليل البياني A Theoretical Example of a Good Item

في التمثيل البياني للمفردات تمثل بيانيا العلاقة بين نسبة استجابات الطلاب الذين يختارون كل بديل – بما في ذلك الإجابة الصحيحة_ على المحور الرأسي Y، والدرجة الكلية على المحور الأفقي X. ويوضح شكل (٣) التالي مثالا افتراضيا لمفردة جيدة . وهذه المفردة جزء من اختبار مكون من ١٥ مفردة من نوع الاختيار من متعدد. ومن ثم فإن الدرجة العظمى على هذا الاختبار تساوي ١٥. تم رسم الخطوط للبدائل أ-ب -ج -د -هـ لهذه المفردة.



شكل (٣): مثال افتراضي لمفردة جيدة

الخط الأسود السميك يمثل البديل الصحيح وهو البديل ج في هذا المثال. يمكن ملاحظة أن الطلاب الذين تتراوح درجتهم الكلية من درجتين إلى أربع درجات قد أظهروا نسبة استجابة للبدائل الخمسة عند مستوى الصدفة ٠.٢٠. والطلاب ذوي الدرجة الكلية صفر أجابوا على هذه المفردة خطأ، حيث إنهم قد أجابوا خطأ على كافة مفردات الاختبار.

عند كل مستويات الدرجات فإن إجمالي النسب المئوية يصل إلى الواحد الصحيح. على سبيل المثال بالنسبة لمجموعة الطلاب ذوي الدرجة الكلية ٥ فإن ٣٢% منهم قد اختاروا البديل الصحيح؛ في حين أن البدائل أ، ب، د، هـ تم اختيار كل منها بنسبة ١٧% من الطلاب. وفي مدى الدرجة الكلية من ٠ إلى ٢ ، ومن ٤ إلى ١٥ فإن الخطوط الخاصة بالبدائل الخطأ تنطبق. وفي الفترة من ٢- ٤ تنطبق كل الخطوط وتبدأ نسبة اختيار الإجابة الصحيحة في الزيادة عند النقطة S، وذلك عند الدرجة الكلية ٤. وعند النقطة R وذلك عند الدرجة الكلية ١١، ويصل الخط السميك إلى نسبة الاستجابة ١، والتي تعتبر القيمة العظمى.ويبدأ مدى التمييز عند النقطة Q والتي تنطبق في هذا المثال النظري على النقطة S، وتنتهي عند النقطة R. وفي هذه المفردة النظرية فإن مدى تمييز المفردة يمتد بين الدرجة الكلية ٤ و ١١. وبمعنى آخر فإن هذه المفردة لها القدرة على التمييز بين الطلاب الذين لديهم درجة كلية أكبر من ٤ وأقل من ١١. وداخل مدى التمييز فإن نسبة الطلاب الذين أجابوا المفردة صوابا يزداد مع زيادة الدرجة الكلية.ولا يحدث هذا خارج مدى التمييز. ويمثل ميل الخط السميك الذي يمثل الإجابة الصحيحة قوة تمييز المفردة. وكلما اتسع مدى التمييز كلما انخفضت قوة تمييزها. فعندما يصبح الميل صفر فإن نسبة الأفراد الذين استجابوا صوابا على المفردة يكون ثابتا ولا يعتمد على الدرجة الكلية. وعندما يصل الميل إلى ٩٠ درجة فإن الخط

الممثل للإجابة الصحيحة يصبح مماثلا لنموذج جتمان Guttman Model. وبالنسبة للطلاب ذوي الدرجة الكلية من ٢ إلى ٤ ومن ١١ إلى ١٥ فإن الخط الممثل للاستجابة الصحيحة يكون أفقيا و يعني ذلك أن القوة التمييزية لهذه المفردة لهذه المجموعة يساوي صفرا. فهذه المفردة لا تميز بين الطلاب ذوي الدرجة الكلية المنخفضة التي تقع في المدى (٢: ٤) أو ذوي الدرجة الكلية المرتفعة التي تقع في المدى (١١: ١٥).

وحصلت المجموعة ذات الدرجة الكلية من صفر: ٢ على درجات أقل من مستوى الصدفة. وفي البيانات الفعلية فإن هذه المجموعة قليلة جدا. النسبة ٠.٥٠ تقطع الخط السميك معبرا عن نسبة الاستجابة الصحيحة عند النقطة P والتي تمثل درجة كلية تقريبا عند ٦.٨ ويمكن أن تعتبر الدرجة الكلية هذه معبرة عن صعوبة المفردة.

وتعد طريقة التحليل البياني للمفردات من الطرق المهمة في الحكم على القوة التمييزية للمفردة، أي المدى الذي تستطيع المفردة من خلاله التمييز بين مستويات المتغير موضع القياس؛ حيث تقدم رسماً بيانياً لمحتوى كل مفردة يوضح العلاقة بين الدرجة الكلية والنسب المئوية للاستجابة لكل من البديل الصحيح والبدائل غير الصحيحة للمفردة. عند كل مستوى من مستويات الدرجة الكلية والتي تعد مؤشرا عن مستويات التحصيل. كما تقدم أيضاً معلومات توضيحية سهلة عن مواصفات المفردة مثل الصعوبة، التمييز، نسبة التخمين، ووصف البدائل (على سبيل المثال بديل مستبعد من قبل كل التلاميذ، بديل جذاب،...)، مما يتيح الفرصة لاختيار مفردات الاختبار من متعدد عالية الجودة. كما يوفر التحليل البياني للمفردات نظرة مبدئية تساعد في التعرف على المفردات الجيدة والمفردات غير المناسبة لأسس القياس؛ وذلك على أساس أن نسبة التلاميذ الذين يختارون الإجابة الصحيحة من المفترض أن تزداد مع زيادة الدرجة الكلية، كما أن نسبة التلاميذ الذين يختارون البدائل غير الصحيحة من المفترض أن تقل مع زيادة الدرجة الكلية. وفي ضوء ذلك فإنه كلما ازداد معدل التلاميذ الذين يختارون الإجابة الصحيحة مع زيادة الدرجة الكلية، كلما كانت القوة التمييزية للمفردة أفضل، وكلما زادت جودة المفردة. وعلى الجانب الآخر فإن المفردات الأقل جودة ستظهر انخفاض في نسبة التلاميذ الذين يختارون البديل الصواب مع زيادة الدرجة الكلية أو ارتفاع نسب التلاميذ الذين يختارون المشتتات مع ارتفاع الدرجة الكلية (Batenburg & Loros, 2000, 309:334).

ويعتمد في الحكم على جودة المفردة في إطار التحليل البياني للمفردات على عدد من المؤشرات من أهمها: شكل البدائل، وتمييز المفردة، شكل البديل الصحيح وصعوبته، التخمين، الحكم على جودة المفردة. ويسترشد في ذلك على عدد مرات اختراق البديل الصحيح لشكله

النموذجي، وعدد مرات اختراق المشتتات لشكلها الافتراضي مثل: ارتفاع نسبة اختيار أحد المشتتات مع ازياد القدرة. أو وجود مشتت يجذب التلاميذ في المستويات العليا أكثر من البديل الصحيح. وكذلك عدد مرات تقاطع البديل الصحيح مع المشتتات بعد النقطة (بداية منحنى التمييز) Q.

خطوات إعداد الرسم البياني:

ولإجراء التحليل البياني للمفردات تتبع الخطوات التالية:

- إدخال البيانات على الحاسب الآلي باستخدام برنامج SPSS.
- تصحيح المفردات الاختبارية في نفس ملف البيانات. مع الاحتفاظ بالبيانات الخام غير المصححة كما هي في الملف حيث تستخدم الاستجابات غير المصححة في رسم بدائل الإجابة.
- حساب الدرجة الكلية لكل طالب على كل مفردة.
- حساب نسبة اختيار كل بديل لكل مفردة.
- استخدام الأمر التالي لرسم كل مفردة مع تغيير كود المفردة في كل مفردة ويتضمن ملحق (١) الكود المستخدم في الرسم البياني كاملا.

```
graph line=mean(altA_1 altB_1 altC_1 altD_1 SC1014K070056) by
Sumscore/TEMPLATE='E:\Analysis\SPSS\Templetes\CHR_TEMP2.sgt .'
```

- رسم الأشكال البيانية لكافة المفردات.
- استخدام المفردات في الحكم على جودة كل مفردة.

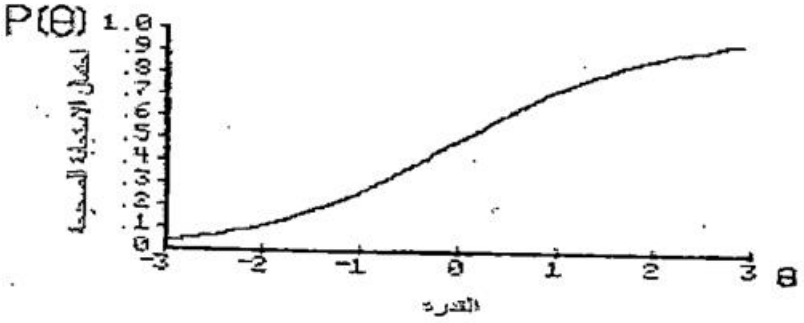
٢- نظام القياس الموضوعي/ نظرية الاستجابة للمفردة Objective Measurement

ظهر هذا النظام كرد فعل للنقد الذي وجه إلى فلسفة القياس التربوي بعامة والنظام جماعي المرجع بخاصة. وارتبط بمدخل جديد أطلق عليه مدخل السمات الكامنة في القياس، ثم أطلق عليه بعد ذلك نظرية الاستجابة للمفردة بما يشتمل عليه من نظريات ونماذج سيكومترية مستحدثة. إذ يُؤخذ على الاختبارات والمقاييس المقننة اعتمادها على مقارنة أداء المتعلم بأداء أقرانه وفقاً للمجموع الكلى لدرجاتهم في الاختبار. وسواء أُجريت هذه المقارنة على أساس الدرجات الخام أو الدرجات المعيارية بأنواعها المختلفة، فإن هذه الدرجات تخضع لخصائص عينة الأفراد التي تُستخدم في اشتقاق معايير الاختبار، كما تخضع لخصائص عينة المفردات التي يشتمل عليها الاختبار، ويؤثر اختلاف هذه الخصائص باختلاف عينة الأفراد المختبرين وعينة المفردات في صدق المقارنات التي تتم في ضوئها، وبذلك يصعب تعميم نتائج الاختبارات أو الاستفادة العملية منها. فتدريج صعوبة مفردات الاختبار يتباين بتباين قدرات الأفراد الذين يُختبرون بهذه المفردات،

وقياس قدرات أو سمات الأفراد يتباين بتباين صعوبة المفردات. كما يتأثر ثبات درجات الاختبار تأثراً كبيراً بتباين قدرات الأفراد الذين يؤدونه.

لذلك اهتم علماء القياس النفسي والتربوي بمواجهة هذه المشكلات والتوصل إلى نماذج جديدة تجعل القياس أكثر موضوعية. حيث تتحرر فيه تقديرات الأفراد من صعوبات المفردات التي اشتقت منها هذه التقديرات، كما تتحرر فيها تقديرات المفردات من قدرات الأفراد التي اشتقت منها هذه التقديرات.

عرفت بنظرية المنحني المميز للمفردة Item Characteristic Curve Theory(ICC) حيث يمثل هذا المنحني كما هو موضح بشكل (٤) التالي العلاقة بين احتمال الاستجابة الصحيحة لمفردة من مفردات الاختبار والسمة أو القدرة موضوع القياس ومن ثم يمكن استخدامه في التنبؤ بكيفية استجابة الفرد على مفردة ما (أمينة كاظم، ٢٠٠٠، ٣٢١: ٣٤٢) (Baker,2001,7).



شكل (٤): المنحنى المميز للمفردة وفق نظرية الاستجابة للمفردة

٢-١ نماذج نظرية الاستجابة للمفردة: انبثقت من نظرية الاستجابة للمفردة عدد من النماذج الرياضية، منها النماذج ثلاثية المعلم وتهتم بدراسة صعوبة المفردة والتمييز والتخمين، والنماذج ثنائية المعلم وتهتم بدراسة صعوبة المفردات وتمييزها، والنماذج أحادية المعلم وأشهرها نموذج راش وهو النموذج الذي تتبناه الدراسة الحالية في تدريج مفردات اختبار التحصيل الدراسي وتقدير قدرات الأفراد لذا تتناوله الباحثة فيما يلي بشيء من الاختصار.

٢-٢ نموذج راش اللوغاريتمي أحادي المعلم

يرجع الفضل في بناء هذا النموذج إلى عالم الرياضيات الدنماركي جورج راش. حيث وضعه ليحقق به العلاقة بين قدرة الفرد وصعوبة المفردة والاستجابة الملاحظة، كما حقق به متطلبات القياس الموضوعي للسلوك، وطوعه للتطبيق العملي العالم الأمريكي بنجامين رايت. واستخدم نموذج راش بصفة أساسية في تحليل مفردات الاختبارات الثنائية التي تُعطي فيها درجة خام واحدة

عن كل استجابة صحيحة على المفردة ودرجة خام صفر لكل استجابة خاطئة. ويفترض نموذج راش انعدام التخمين، وتساوي معاملات التمييز، ويقوم بتقدير معلم واحد للمفردة هو الصعوبة

ويستند نموذج راش إلى افتراض أساسي يؤكد على أنه كلما ازدادت قدرة الفرد عن صعوبة المفردة ازداد احتمال حدوث الاستجابة الصحيحة للمفردة والعكس صحيح، وإذا تساوت قدرة الفرد مع صعوبة المفردة فإن احتمال حدوث الاستجابة الصحيحة يتساوى مع احتمال حدوث الاستجابة الخطأ (أمينة كاظم،١٩٩٤: ١١٦ -١٢٥).

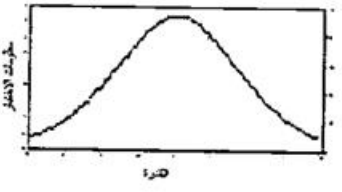
٢-٣ افتراضات نظرية الاستجابة للمفردة: تقوم نظرية الاستجابة للمفردة على مجموعة من الافتراضات تتيح عند تحققها الاستفادة مما توفره النظرية من مميزات. ومن أهم هذه الافتراضات: افتراض أحادية البعد Unidimensionality حيث تفترض معظم نماذج نظرية الاستجابة للمفردة أن هناك سمة أو قدرة واحدة فقط كافية لتفسير وتوضيح الفروق بين أداءات الأفراد على الاختبار. وافتراض التحرر من عامل السرعة في الإجابة Speediness بمعنى أن إخفاق بعض الأفراد في الإجابة على بعض مفردات الاختبار بصفة أساسية يرجع إلى عدم امتلاكهم المعرفة أو المهارة المطلوبة، وليس إلى تأثير عامل السرعة في إجاباتهم. وافتراض استقلالية الموضع (LII) Local Item Independence أي أن تقدير صعوبة أي مفردة لا يعتمد على تقديرات صعوبة المفردات الأخرى ولا يعتمد على قدرة الأفراد الذين يجيبون عليها، وكذلك لا يعتمد على تقدير قدرة أي مجموعة أخرى من الأفراد الذين يطبق عليهم الاختبار ولا يعتمد على قيم صعوبة المفردات التي يجيبون عنها. بالإضافة إلى افتراض توازي المنحنيات اللوغاريتمية التراكمية المميزة لمفردات الاختبار (أمينة كاظم،١٩٩٤، ١١٤)(Vukmirovic,2009) .

٢-٤ إسهامات جوهرية لنظرية الاستجابة للمفردة في سياق تطوير مفاهيم القياس: قدمت نظرية الاستجابة للمفردة عددا من المفاهيم المهمة المميزة لهذه النظرية، ومن أهمها دالة المعلومات، والخطأ المعياري للقياس، وفيما يلي عرض موجز لكل منها.

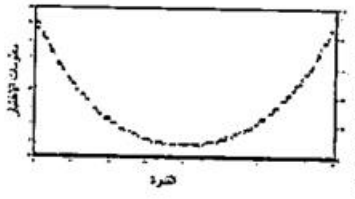
٢-٤-١ دالة المعلومات Information Function: يمثل مفهوم دالة المعلومات أحد المفاهيم المميزة لنظرية الاستجابة للمفردة، ومن أكثرها فائدة، حيث تساعد في تحديد قدر المعلومات التي تقدمها المفردة عن فرد ما، وذلك من خلال تحديد أقصي ارتفاع للمنحني الممثل لدالة المعلومات لتلك المفردة على متصل القدرة، وبهذا يمكن تحديد أي المفردات الاختبارية تقيس المتغير المراد قياسه بدرجة أفضل عند مستويات محددة للقدرة (Baker,2008,107:110) (Banks, 2015). وعرفها فيشر Fisher

على أنها مقلوب قياس دقة تقدير البارامترات ويعني التباين بين عناصر التقدير، ويعني ذلك أنه كلما كان قدر المعلومات كبيراً كلما انخفض الانحراف المعياري بما يعني ارتفاع دقة التقديرات، أي اقترب تقدير قدرة الأفراد من تقديراتها الحقيقية عند هذا المستوى من القدرة، ولكن إذا كان قدر المعلومات قليلاً فإن هذا يعني ضعف دقة التقديرات وابتعادها عن التقديرات الحقيقية لها. وتوفر نظرية الاستجابة للمفردة إمكانية الحصول على منحنى دالة المعلومات لكل مفردة اختبارية، بالإضافة إلى منحنى دالة المعلومات للاختبار ككل.

٢-٤-٢ الخطأ المعياري للقياس Standard Error of Measurement: يأتي مفهوم الخطأ المعياري للقياس في إطار نظرية الاستجابة للمفردة متكاملا مع مفهوم دالة المعلومات حيث يعد كل منهما وجها من أوجه التعبير عن دقة وثبات القياس، وإن كان رياضيا كل منهما مقلوب للآخر كما يعبر عنه بشكل (٥، ٦) التاليين.



شكل (٥): دالة المعلومات للاختبار



شكل (٦): الخطأ المعياري للاختبار

ويقوم على مفهوم دالة المعلومات والخطأ المعياري للقياس وفق نظرية الاستجابة للمفردة عدد من التطبيقات الحديثة منها الاختبارات المواءمة بالكمبيوتر Computerized Adaptive Test (CAT).

٢-٥ البرمجيات المطورة في إطار نظرية الاستجابة للمفردة:

طور في إطار نظرية الاستجابة للمفردة عدد من البرمجيات التي تلائم النماذج المختلفة للنظرية ومنها ما هو مخصص للتحليل وفق نموذج راش أحادي المعلم ومن أشهرها برنامج WINSTEPS ويوفر هذا البرنامج عدد من المميزات من أهمها: سهولة استخدامه، كما أدرج عليه عدد من أوجه التطوير حيث يتيح حاليا التحقق من أحادية البعد للمقاييس المستخدمة، كما يمكن من خلاله تدريج المفردات متعددة الاستجابة باستخدام نموذج مقياس التقدير Rating Scale Method. كما يقدم البرنامج معلومات تشخيصية هائلة حول استجابات كل من الأفراد والمفردات.

الدراسات السابقة:

في إطار دراسة تأثير بعض بيانات التغذية الراجعة على تقديرات المحكمين وممارساتهم خلال تحديد درجة القطع جاءت دراسة جيورووارد (Giraud, 1999) عقدت مناقشات مع المشاركين في

تحديد مستويات الأداء للوقوف على العوامل المؤثرة على قراراتهم حول كيف يؤدى الطلاب الجيدون على مفردات الاختبار. وتوصلت الدراسة إلى أن مناقشة المشاركين مع المعلمين الآخرين حول مفهوم الطالب عند حد الإتقان، والتفكير حول واحد أو أكثر من الطلاب الملائمين لهذا التعريف والاعتبارات السياسية والاقتصادية للولاية التي تقع بها المدرسة كانت أهم العوامل المؤثرة على أحكامهم.

وهدفت دراسة " هامبلتون وسيرسي" (Hambleton, &Sireci,2001) إلى تتبع أفكار المحكمين وتصورات أثناء مشاركتهم في عملية التحكيم لتحديد درجة القطع في مهارات القراءة، وطلبوا من المحكمين وصف مستويات أداء الطلاب قبل وبعد مشاركتهم في عملية التدريب على تحديد درجات القطع، وتوصلت نتائج الدراسة إلى أن تحديد درجة القطع تتأثر بمستوى تدريب المحكم ومدى انخراطه في عملية التحكيم، والمناقشة الجماعية المفتوحة مع زملائه من المحكمين.

وتناولت دراسة سكوربسكي وهامبلتون (Skorupski and Hambleton, 2003) العمليات المعرفية التي يمارسها المحكمون عندما يصفون مستويات الأداء-حيث تم مقابلة المحكمين على خمس مراحل خلال عملية وضع مستويات الأداء: قبل أي عرض لمحتوى الاختبار، وفى نهاية مرحلة التدريبات، وعند نهاية الجولة الأولى للتقديرات وبعد مناقشة نتائج تعليمات الجولة الأولى وعند انتهاء الجولة الثانية للتقييمات. وأشارت نتائج الدراسة إلى أن المحكمين يأتون إلى ورش تحديد مستويات الأداء بأفكار ورؤى مختلفة. وبصفة عامة كان المحكمون على درجة عالية من الثقة من فهمهم لمستويات الأداء. كما أظهر المحكمون أن التغذية الراجعة ساعدتهم في فهم المواصفات القياسية للأداء.

وفى دراسة إيجان وجرين (Egan and Green, 2003) تم تحديد العوامل المؤثرة على قرارات المحكمين في طريقة تحديد مواصفات الأداء المعروفة بطريقة "بوك مارك" Book mark حيث طلب من المحكمين بعد كل دورة التصويت والمناقشة حول خمسة عشر عاملا يمكن أن تؤثر على تقديراتهم. وأظهرت النتائج أن تقديرات المحكمين الآخرين، وخبرة المحكم وتعريف الطالب المستهدف كانت أكثر ثلاثة عوامل من بين خمسة عشر عاملاً تؤثر على قرارات المحكمين.

واهتمت دراسة فيردوس وبلاك (Ferdous and Plake, 2005) بالعوامل المؤثرة على قرارات المحكمين في طريقة أنجوف لوضع المواصفات القياسية للأداء كدالة لتقديراتهم للأداء على المفردة. حيث تم تقسيم المشاركين إلى ثلاث مجموعات بناء على تقديراتهم للأداء في الجولة الأولى (منخفض-متوسط-مرتفع). وتوصلت النتائج إلى أن المشاركين في المجموعات المرتفعة والمتوسطة في تقدير درجات القطع أخذوا في اعتبارهم متغيرات عديدة خلال تقديراتهم لمفردات

الاختبار أكثر من أقرانهم في المجموعة المنخفضة التقديرات؛ حيث فكرت المجموعة المنخفضة التقديرات من منظور جماعي المرجع في حين فكرت المجموعتين المرتفعة والمتوسطة من منظور محكي المرجع.

وهدفت دراسة أحمد الشريم ويوسف السوالمة (٢٠٠٦) إلى مقارنة نموذجي أنجوف وندلسكي لتقدير درجة القطع لاختبار محكي المرجع في الرياضيات وذلك عند وجود مؤشرات عن صعوبة الفقرات وعدم وجودها. وذلك على اختبار مكون من ٣٠ مفردة من الاختيار من متعدد. وتكونت عينة الدراسة من ٨٠ محكما ومحكمة تم تقسيمهم بطريقة المزاوجة العشوائية إلى أربع مجموعات متساوية حددت كل مجموعة منها درجة القطع للاختبار مرتين وفق الأسلوب المحدد لها. وأشارت نتائج الدراسة إلى أن درجة القطع للاختبار باستخدام نموذج أنجوف جاءت أعلى منها باستخدام نموذج ندلسكي. جاءت تقديرات المحكمين عينة الدراسة أكثر ارتباطا بصعوبات المفردات بعد تزويدهم بصعوبات المفردات.

وهدفت دراسة فردوس وبوكيندهت **(Ferdous and Buckendaht, 2013)** إلى تحديد ما يفكر فيه المحكمون عندما يصدرون تقديراتهم على المفردات في مختلف مستويات الأداء (ضعيف- مقبول- مرض- متقدم) باستخدام طريقة "نعم/ لا" لأنجوف وذلك في اختبارين أحدهما في الرياضيات والآخر في اللغة الإنجليزية بالصف الخامس الابتدائي وتكونت عينة الدراسة من ١٣ معلما للرياضيات، و١٧ معلما للغة الإنجليزية، حيث تم تقديم مجموعة من الأسئلة للمحكمين عينة الدراسة حول جميع خطوات عملية تحديد درجات القطع. حيث تم تزويدهم ببيانات الأثر بين الجولتين الأولى والثانية وشملت بيانات الأثر معامل صعوبة مفردات الاختبارات الفعلية، كما اشتقت من التطبيق الميداني للاختبار بالإضافة إلى التوزيع التكراري المتجمع الصاعد للطلاب في كل مستوى من مستويات الأداء وفق درجات القطع التي يحددها المحكمون. وأشارت النتائج إلى أن المحكمين قد اكتسبوا خبرة حول إجراءات تحديد مستويات الأداء، وفهم مواصفات مستويات الأداء لكل مستوى معرفي، وإصدار قرارات أكثر اتساقاً حول مفردات الاختبار.

كما أشارت النتائج إلى أن معظم المحكمين لم يتأثروا بقيم صعوبة المفردة وبيانات الأثر أثناء إجراء تقديرات الجولة الثانية. كما أشار المحكمون إلى أن بيانات التغذية الراجعة قد ساعدتهم في الفهم الأفضل لفئات مستويات الأداء الأربعة كما زودتهم بمراجعة واقعية لتقديرات أدائهم في الجولة الأولى.

وهدفت دراسة هسيه **(Hsieh,2013)** إلى وضع إطار مرجعي لتقييم المحكمين خلال عملية تحديد المواصفات القياسية للأداء. حيث طبقت نموذج راش متعدد الأوجه Multifaceted

Rasch Model كأداة لتقييم جودة عملية تحديد المواصفات القياسية للأداء في سياق تقييم اللغة. وتوصلت الدراسة إلى أن نموذج راش متعدد الأوجه يوفر مدخلا مهما لاختبار التباين في إجراءات عملية تحديد المواصفات القياسية للأداء. بالإضافة إلى قدرة النموذج على التعرف على القرارات الاعتباطية بالنسبة لكل محكم، والتي يمكن الاعتماد عليها في تقديم التغذية الراجعة لكل من القائمين على تقييم عملية تحديد درجات القطع والمحكمين أنفسهم.

وهدفت دراسة مارجوليوس وكلاسر (Margolis & Clauser,2014) إلى دراسة أثر معرفة المحكمين بعض المعلومات عن أداء الممتحنين على درجات القطع التي يقومون بتحديدها وفق طريقة أنجوف المعدلة. وتضمنت المعلومات المقدمة للمحكمين النسب المئوية للطلاب في مستويات الأداء المختلفة عند تطبيق درجة القطع المحددة في الجولة الأولى على بيانات البيانات الفعلية وصعوبات المفردات وكذلك التوزيع التكراري لاختيارات الطلاب لكل بديل من بدائل أسئلة الاختبار. وتم جمع البيانات من عينة قدرها (١٨) محكما مستقلا شاركوا في تحديد مستويات الأداء عبر ثلاثة اختبارات للترخيص للمهن الطبية، تم اختيارها لفحص ما إذا كانت تزويد المحكمين ببيانات عن أداء الأفراد يمكن أن تؤثر على قرارات المحكمين ودرجات القطع الناتجة وكيف يكون ذلك. وتباينت النتائج باختلاف مجموعات المحكمين، وأشارت النتائج بصفة عامة إلى تباين تأثير البيانات الخاصة بالطلاب باختلاف المحكمين، كما تأثرت درجات القطع الناتجة بالبيانات المقدمة للمحكمين. كما أدى تزويد المحكمين ببيانات عن أداء الطلاب إلى اختلاف درجات القطع لكل المشاركين بعد تزويدهم بالبيانات عنها قبل تزويدهم ببيانات التغذية الراجعة. كما أسفرت النتائج عن التباين بين المحكمين. وانخفاض التجانس _إلى حد ما_ بين المحكمين بعد مراجعتهم للبيانات، كما اختلفت درجات القطع لكل المحكمين وبالنسبة لكافة الاختبارات اختلافا جوهريا قبل وبعد تزويد المحكمين ببيانات التغذية الراجعة.

وفي إطار مقارنة طريقة أنجوف بغيرها من طرق تحديد درجة القطع جاءت دراسة هاراسيم (Harasym,1981) وهدفت إلى مقارنة طريقتي أنجوف المعدلة، وطريقة نيدلسكاي في تحديد درجة القطع لأحد الاختبارات الطبية. وشملت عينة الدراسة (٢١٢) طالبا من طلاب الفرقة الثانية بكلية الطب. وأشارت نتائج الدراسة إلى أن استخدام طريقة أنجوف المعدلة يؤدي إلى الحصول على درجات قطع أكثر اتساقا من تلك الدرجات الناتجة باستخدام طريقة نيدلسكاي.

وتناولت دراسة هالبن وسيجمون وهالبن (Halpin; Sigmon & Halpin,1983): فحص صدق ثلاث طرق لتحديد درجة القطع (إيبل- ندلسكاي- أنجوف) عند استخدام مجموعات مختلفة من المحكمين باستخدام اختبار موحد. وتكونت عينة الدراسة من خمسة عشر محكماً،

وزعوا على ثلاث مجموعات، تتكون كل منها من خمسة أعضاء، وتشتمل المجموعة الأولى على خمسة من طلاب درجة الدكتوراه في تعليم اللغة الإنجليزية، وتشتمل المجموعة الثانية على خمسة من معلمي اللغة الإنجليزية بالمرحلة الثانوية، وتشتمل المجموعة الثالثة على خمسة أعضاء من الكلية لديهم خبرة وتدريبات في تعليم اللغة الإنجليزية، وتكونت عينة الطلاب من (١٧٣) طالباً من طلاب الجامعة. وتمثلت أهم نتائج الدراسة في: وجود فروق في درجات القطع الناتجة باستخدام الطرق المختلفة. جاءت درجة القطع لعينة المحكمين من المعلمين مختلفة بشكل ملحوظ عنها بالنسبة لباقي المحكمين من أعضاء هيئة التدريس ومعاونيهم.

وهدفت دراسة **بوهينيك وآخرون (Behuanik, et.al.1983)** إلى دراسة صدق تفسير درجة التمكن على اختبار للقراءة وآخر للرياضيات من نوع الاختيار من متعدد باستخدام طريقتي أنجوف ونيدلسكاي. من خلال دراسة تأثير بعض العوامل على درجة القطع منها الخصائص الديموغرافية للمحكمين، طريقة تحديد درجة القطع. وتكونت عينة الدراسة من (٢٧) عضوا من أعضاء هيئة التدريس. وتكونت عينة الطلاب من (٤٦٠) طالب. وقد أسفرت نتائج الدراسة عن عدم وجود علاقة بين **الخصائص الديموغرافية** للمحكمين وبين درجات القطع التي قاموا بتحديدها، أدت طريقتي أنجوف ونيدلسكاي إلى درجات قطع مختلفة إحصائيا وعمليا. يميل المحكمون في المجموعة الواحدة الذين يستخدمون نفس الطريقة إلى إعطاء تقديرات أكثر اتساقا من المحكمين الذين ينتمون إلى مجموعات مختلفة.

دراسة صلاح الدين علام (١٩٩١): هدفت إلى المقارنة بين أربع طرق لتحديد مستويات الأداء في الاختبار اثنتان من الطرق المطلقة (أنجوف وندلسكاي) واثنتان من الطرق النسبية (المجموعة الحدية، المجموعات المتناقضة). وتكونت عينة المحكمين من(٢٦) محكماً من ثلاث مجموعات متباينة في مستوى خبرتها في مجال القياس والتقويم، اشتملت المجموعة الأولى على(١٤) معلماً من معلمي المدارس الثانوية العامة الحاصلين على دبلوم خاص في التربية أو على دورة تدريبية في مجال القياس والتقويم، واشتملت المجموعة الثانية على ستة مدرسين مساعدين لديهم خبرة لا تقل عن ثلاثة أعوام في تدريس مقرر الاختبارات والمقاييس، واشتملت المجموعة الثالثة على ستة من أساتذة علم النفس التعليمي. وبلغت عينة الطلاب الذين أجرى تحكيم أدائهم في الاختبار (٣٨٢) طالباً ممن يدرسون مقرر الفروق الفردية، ويدرسه لهم المحكمون عينة الدراسة. وأوضحت نتائج الدراسة أن استخدام الطرق الأربع المختلفة لتحديد درجات القطع أدى إلى الحصول على درجات قطع مختلفة نسبيا. كما أشارت النتائج إلى عدم وجود فروق دالة إحصائيا لتقديرات درجة القطع باختلاف خبرة المحكم.

وأجرى عصام الجبة (١٩٩٨) دراسة لبحث مدى فاعلية طريقة أنجوف فى تحديد درجة القطع، وذلك من خلال مقارنة نتائجها مع نتائج طريقة هوفستى واستخدم اختبار في مادة الجبر للصف الأول الثانوى، يتضمن ٢٠ مفردة من نوع الاختيار من متعدد ذات الأربعة بدائل. واستعان الباحث بعدد (٦) محكمين من موجهي الرياضيات بالإضافة إلى أربعة مدرسين أوائل وتكونت عينة الدراسة من الطلاب من (٦٥) طالبا. وتوصلت الدراسة إلى فاعلية طريقة أنجوف في تحديد درجة القطع، حيث لم تختلف نتائجها مع نتائج طريقة هوفستى.

وهدفت دراسة خالد التميمي (١٩٩٩) إلى تحديد أثر كل من نوع المحكم وطول الاختبار على تحديد درجة القطع لاختبار محكي المرجع يقيس الكفايات الرياضية. واستخدمت أربع طرق لتحديد درجة القطع تمثلت في طريقة أنجوف، و نيدلسكاي، والمجموعات المتضادة، والمجموعات المحكية. شملت الدراسة عدة عينات تمثلت في عينة من الطلاب وبلغ عددها (٦٥٩) طالباً، وعينة من معلمي مادة الرياضيات بلغ عددها(٢٠) معلما، وعينة من مشرفي مادة الرياضيات بلغ عددها (١٠) مشرفاً، وعينة من أعضاء هيئة التدريس بلغ عددها(١٠) أعضاء. وقد أسفرت نتائج الدراسة عن: عدم وجود فروق دالة إحصائياً بين درجات قطع الاختبار الناتجة باختلاف نوع المحكم وفق طريقة أنجوف، وإن كان التقارب بين درجات القطع أكبر بين كبر بين تقديرات المحكمين من المعلمين والمشرفين عنه بالنسبة لأعضاء هيئة التدريس حيث جاء تقديرات أعضاء هيئة التدريس أقل. وارتفاع تقديرات المحكمين لدرجة قطع الاختبار وفق طريقة أنجوف مقارنة ببقية الطرق.

واهتمت دراسة محمد فراج (٢٠٠٦) بتقصي فاعلية بعض النماذج الإحصائية المنفصلة متمثلة في نموذج ذي الحدين والمتصلة متمثلة في طريقة أنجوف في تحديد درجة القطع لاختبار مرجعي المحك، باستخدام اختبار مكون من ٢٠ مفردة من نوع الاختيار من متعدد في مادة علم النفس التربوي على طلاب كلية المعلمين بالرياض. وتوصلت نتائج الدراسة إلى ارتفاع درجة القطع باستخدام نموذج أنجوف عن نظيرتها المقدرة باستخدام نموذج ذي الحدين.

وبمراجعة الدراسات السابقة يتضح ما يلي:

- بالنسبة لمقارنة طريقة أنجوف بغيرها من طرق تحديد درجة القطع تبين اختلاف درجة القطع باختلاف الطريقة المستخدمة في تقديرها، وتميل طريقة أنجوف إلى إعطاء تقديرات أعلى لدرجة القطع مقارنة بطريقة نيدلسكاي، وقد يرجع ذلك إلى إجراءات طريقة نيدلسكاي ذاتها في تحديد درجة القطع.
- أشارت النتائج إلى أن طريقة أنجوف أكثر ثباتا واستقرارا مقارنة بباقي الطرق.

- بالنسبة لتأثير خبرة المحكم على درجات القطع أشارت بعض الدراسات إلى أن درجة القطع لا تختلف باختلاف خبرة المحكم مثل دراسة (صلاح علام ،١٩٩١)، (خالد التميمي،١٩٩٩)؛ إلا أن هناك دراسات أشارت إلى اختلاف درجة القطع باختلاف خبرة المحكم مثل دراسة (Margolis & Clauser,2014) (Halpin; Sigmon & Halpin,1983) ,

- تباينت الدراسات من حيث **طبيعة البيانات** المقدمة للمحكمين؛ حيث اقتصرت بعض الدراسات على تقديم صعوبة المفردات ومنها دراسة (أحمد الشريم، ويوسف السوالمة،٢٠٠٦)، ومنها ما امتدت لتزيد المحكمين ببيانات الأثر حول النسب المئوية للطلاب في مستويات الأداء المختلفة نتيجة لتطبيق درجة القطع بالإضافة إلى صعوبات المفردات مثل دراسة (Ferdous and Buckendaht, 2013) ، وأضافت دراسة (Margolis & Clauser,2014) إلى هذه العوامل التوزيع التكراري لبدائل كل مفردة. واستخدمت دراسة (Ferdous and Plake, 2005) بالإضافة إلى ذلك بيانات جماعية تتعلق بمتوسط ووسيط ومنوال درجات القطع للمحكمين في الجولة الأولى. ولم تتطرق أي منها لتوظيف التحليل البياني للمفردات في تقديم بيانات الأثر للمحكمين.

- تباينت نتائج الدراسات من حيث تأثير التغذية الراجعة على ما يمارسه المحكمون من عمليات معرفية أثناء تحديد درجة القطع وتوصلت بعض الدراسات إلى أن بيانات التغذية الراجعة أثرت بدرجة كبيرة على قرارتهم في الجولة الثانية مثل دراسة (Margolis & Clauser,2014)، في حين أشارت نتائج بعض الدراسات إلى أن قرارتهم لم تتأثر ببيانات التغذية الراجعة، وإنما ساعدتهم التغذية الراجعة فقط في فهم مستويات الأداء وفهم مفهوم الطالب ذوي الحد الأدنى من الأداء مثل دراسة (Ferdous and Buckendaht, 2013).

- لا توجد دراسة عربية تناولت تأثير بيانات الأثر على تقديرات درجة القطع سوى دراسة (أحمد الشريم ويوسف السوالمة،٢٠٠٦) واقتصرت فقط على تزويد المحكمين بمعامل صعوبة المفردات.

وفي ضوء ما أشارت إليه نتائج بعض الدراسات من جدوى استخدام بيانات التغذية الراجعة وفي ضوء تباين نتائج الدراسات حول أي عوامل التغذية الراجعة أكثر تأثيرا، وطبيعة هذا التأثير ومع التطور التكنولوجي والبرامج الإحصائية والتطور في مجال القياس، وظهور التحليل البياني للمفردات ومع التطور في مجال بناء الاختبارات وظهور نظرية الاستجابة للمفردة بما تسمح به من تدريج للمفردات على ميزان تدرج واحد بصفر مشترك واحد. جاءت فكرة الدراسة الحالية في دراسة أثر بيانات التغذية الراجعة وخبرة المحكم على تقديرات درجة القطع وممارسات المحكمين وفق طريقة أنجوف المعدلة، وذلك بعد التحقق من بناء الاختبار من

خلال تدريجه باستخدام نظرية الاستجابة للمفردة، وتطوير بيانات التغذية الراجعة لتشمل **بيانات جماعية** (تتمثل في درجات القطع للمحكمين الآخرين-المتوسط – الوسيط-المنوال-الانحراف المعياري-القيم العظمى-القيم الصغرى) لدرجات القطع. كما تشمل **بيانات الأثر** (النسبة المئوية للطلاب في كل مستوى من مستويات الأداء وفق درجة القطع المقدرة من المحكمين في الجولة الأولى). كما تشمل أيضا البيانات الواقعية (وتشمل معامل الصعوبة، النسب المئوية لاختيار كل بديل، والتحليل البياني للمفردات).

حدود الدراسة: اقتصرت الدراسة الحالية على:

- بالنسبة لمجال خبرة المحكم تم الاقتصار على عينة من أعضاء هيئة التدريس والمعلمين.
- بالنسبة لطريقة أنجوف تم الاقتصار على طريقة (نعم/ لا) لأنجوف المعروفة بطريقة "توكر" من خلال استجاباتهم على مفردات الاختبار.
- بالنسبة لبيانات التغذية الراجعة: تم الاقتصار على النسبة المئوية للطلاب في كل مستوى من مستويات الأداء وفق درجة القطع المقدرة من قبل المحكمين في جولتي التحكيم، والمتوسط والوسيط والمنوال والانجراف المعياري والقيم العظمى والصغرى لدرجة القطع، والتكرارات والنسب المئوية للبدائل، ومعامل الصعوبة، والتحليل البياني للمفردات.

إجراءات الدراسة:

تصميم الدراسة: استخدمت الدراسة تصميم البحوث المختلطة التلازمية Concurrent Mixed Method Research Design (Creswell, 2003) حيث تم جمع البيانات الكمية والكيفية بالتزامن في آن واحد من خلال الاستبيان، حيث تم جمع البيانات الكمية من خلال مجموعة من المفردات متدرجة الفئات باستخدام مقياس ليكرت. وجمعت البيانات الكيفية (النوعية) باستخدام عدد من المفردات مفتوحة النهاية لفهم ما يفكر فيه المشاركون عندما يقدرون الأداء على مفردات الاختبار.

عينة الدراسة: تضمنت الدراسة الحالية عينتين هما:

أ- عينة الطلاب: وتتضمن الطلاب الذين تم تطبيق الاختبار التحصيلي عليهم، وذلك بغرض تدريجه باستخدام نظرية الاستجابة للمفردة، واستخلاص المؤشرات السيكومترية للاختبار ويتضمن جدول (١) التالي وصف عينة الطلاب.

جدول (١): توزيع عينة الطلاب وفق متغير النوع

النوع	العدد	النسبة المئوية
انثى	97	48.5
ذكر	103	51.5
المجموع	200	100.0

ب- عينة المحكمين: شملت الدراسة الحالية ستة من أعضاء هيئة التدريس بالمركز القومي للامتحانات والتقويم التربوي وجامعة حلوان وجامعة القاهرة، وتسع من معلمي مادة العلوم بإدارة مدينة نصر، وإدارة ملوي وإدارة السادات التعليمية. ويتضمن جدول (٢) التالي وصفاً لعينة الدراسة وفق متغيري النوع، والوظيفة.

جدول(٢): عينة المحكمين موزعة وفق متغيري النوع والوظيفة

الوظيفة	النوع				المجموع	
	انثى		ذكر			
	العدد	النسبة	العدد	النسبة	العدد	النسبة
عضو هيئة تدريس	5	٣٣,٣٣	1	٦,٦٧	6	٤٠
معلم	5	٣٣,٣٣	4	٢٦,٦٧	9	٦٠
المجموع	10	٦٦,٦٦	5	٣٣,٣٣	15	١٠٠

أدوات الدراسة: تمثلت أدوات الدراسة في اختبار التحصيل الدراسي، والاستبيان الخاص بتأثير بيانات التغذية الرجعة على ممارسات المحكمين العملية خلال تقدير درجات القطع، وفيما يلي عرض لخطوات بناء كل منهما.

أولا: اختبار العلوم للصف الأول الإعدادي (إعداد الباحثة): يهدف هذا الاختبار إلى قياس تحصيل الطلاب في وحدة المادة، وتركيبها وتكون في صورته النهائية من (٣٠) مفردة.

وقد اتبع في بناء وتدريج اختبار التحصيل الدراسي الخطوات الآتية:

- **تحديد محتوى المقرر الدراسي المراد قياسه وتحليله إلى أوجه التعلم الرئيسة المتضمنة فيه.**

وقد انحصر مجال اهتمام الدراسة الحالية في محتوى وحدة المادة وتركيبها من محتوى مادة العلوم للصف الأول الإعدادي للعام الدراسي (٢٠١٤/٢٠١٥) وتم تحليل محتوى هذه الوحدة في ضوء أوجه التعلم المتضمنة فيها، والتي تمثل البناء المعرفي للعلم وتشمل: **الحقائق (المعرفة**

العلمية)، المفاهيم، والتعميمات والمبادئ والقوانين والنظريات، بغرض تحديد التوزيع النسبي لأوجه التعلم المتضمنة في هذه الوحدات، ثم قامت الباحثة بعرض تحليل المحتوى على مجموعة من المحكمين بغرض مراجعة تحليل المحتوى والوقوف على مدى دقته وصدقه.

- **مراجعة المعايير القومية للتعليم في جمهورية مصر العربية الخاصة بمادة العلوم.**

قامت الباحثة بالاطلاع على المعايير القومية للتعليم لمحتوى مادة العلوم (المجلد الثالث)؛ وذلك بغرض الوقوف على أهم المعايير القومية لتعليم العلوم التي ينبغي أن تتحقق من خلال الوحدة محل الدراسة، وكذا مؤشرات تحقيق كل منها؛ ومن ثم إمكانية قياسها إجرائيا من خلال صياغة الأهداف الإجرائية التي تحققها وتقيس جوانب التعلم المتضمنة في هذه الوحدة والتي تغطي ثلاثة مستويات من مستويات النشاط العقلي وهي (التذكر، الفهم، التطبيق).

وبعد تحليل المحتوى إلى أوجه التعلم المتضمنة، وتحديد المعايير المراد تحقيقها ومؤشراتها من خلال الوحدة محل الدراسة كجزء من معايير العلوم في المرحلة الاعدادية، قامت الباحثة بترجمة هذه المؤشرات إجرائيا من خلال صياغتها إلى أهداف إجرائية تقيس جوانب التعلم المتضمنة في الوحدة محل الدراسة. وتغطي ثلاثة مستويات من مستويات النشاط العقلي وهي (التذكر، والفهم، والتطبيق).

- **بناء المفردات الاختبارية:** وتضمن بناء المفردات الاختبارية عدة خطوات تمثلت في: تصنيف الأهداف السلوكية التي تم التوصل إليها في شكلها النهائي، وبناء جدول مواصفات الاختبار تبعا لمحوري المحتوى والمستوى المعرفي. ثم تحديد عدد المفردات التي ترتبط بالأهداف في موضوعات الوحدة محل الدراسة. ويوضح جدول (٣) التالي أعداد المفردات التي تقيس الأهداف السلوكية.

جدول (٣): أعداد المفردات الاختبارية والمستويات المعرفية التي يقيسها اختبار التحصيل الدراسي

م	الموضوع		المستوى المعرفي للهدف			المجموع
			تذكر	فهم	تطبيق	
١	تركيب الذرة	عدد الأسئلة	٦	٥	٧	١٨
		النسبة المئوية	٢٠	١٦.٧	٢٣.٣	٦٠
٢	خواص المادة	عدد الأسئلة	٦	٤	٢	١٢
		النسبة المئوية	٢٠	١٣.٣	٦.٧	٤٠
٤	المجموع	عدد الأسئلة	١٢	٩	٩	٣٠
		النسبة المئوية	٤٠	٣٠	٣٠	١٠٠

وبالوصول إلى هذه المرحلة يكون قد توفر قدر من الصدق الملائم للمفردات الاختبارية

أو ما يسمى بصدق المحتوى ويعني أن محتوى المفردات الاختبارية يعكس الأهمية النسبية للأهداف التعليمية ويعبر عنها تعبيراً صحيحاً.

- **صياغة المفردات الاختبارية التي تقيس الأهداف السلوكية:**

في ضوء جدول المواصفات قامت الباحثة ببناء مجموعة من المفردات الاختبارية، وقد تم اختيار شكل المفردات من نوع **الاختيار من متعدد** ذات البدائل الأربعة يمثل أحدها الإجابة الصحيحة للمفردة.

- **التحكيم على المفردات:** بعد الانتهاء من صياغة المفردات تم عرضها على عدد من المحكمين، بغرض الوقوف على مدى صحتها العلمية وعلى أسلوب صياغتها ومدى تغطيتها وشمولها للأهداف السلوكية ومدى ملاءمتها للمستويات المعرفية المحددة. وبعد إجراء التعديلات التي اقترحها السادة المحكمون فإن المفردات التي استقر الرأي عليها أخيراً هي المفردات التي تتكون منها الصورة الأولية للاختبار التحصيلي الذي تسعى الدراسة الحالية إلى تدريجه وبلغ عددها (٣٠) مفردة.

- **صياغة التعليمات:** تم صياغة تعليمات التطبيق بحيث تتضمن: الهدف من التطبيق، وتعليمات الإجابة على المفردات.

- **التجريب الاستطلاعي للمفردات:** قامت الباحثة بتجريب المفردات الاختبارية استطلاعياً بتطبيقها على مجموعة من التلاميذ بلغ عددها(٣٠) تلميذاً وتلميذة من تلاميذ الصف الأول الإعدادي بمدرسة خالد بن الوليد الإعدادية بإدارة المقطم التعليمية للوقوف على مدى وضوح التعليمات، ووضوح الصياغة، وملاءمة البدائل، والمشكلات التي تتعلق بالتطبيق حتى يمكن تداركها قبل التطبيق على عينة التدريج، وكذلك حساب الزمن اللازم للإجابة على الاختبار حيث إنه بالرغم من أن الاختبار في الدراسة الحالية اختبار قوة وليس اختبار سرعة، فقد تم حساب الزمن اللازم للإجابة على الاختبار. حيث استغرق تطبيقه حوالي (٤٥) دقيقة تقريباً حيث كان هذا الزمن كافياً لأن يحاول كل من يتعرض للاختبار الإجابة عن جميع مفرداته.

- **تجهيز البيانات للتحليل:** بعد الانتهاء من الخطوات السابقة وإجراء التعديلات المناسبة قامت الباحثة بالخطوات التالية:

- **تطبيق الاختبار** على عينة التدريج.

- **ترميز البيانات:** حيث تم تحديد رمز لكل مفردة من مفردات الاختبار.

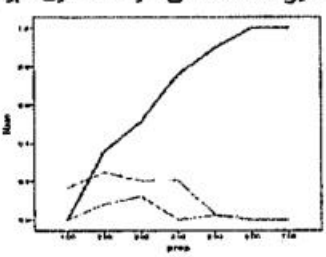
- **تجهيز ملفات الإدخال** الخاصة بالاختبار وإدخال البيانات الخاصة بأداء أفراد عينة التدريج على المفردات المختلفة على الحاسب الآلي باستخدام الحزمة الإحصائية للعلوم الاجتماعية Statistical

.Package for Social Sciences (SPSS)

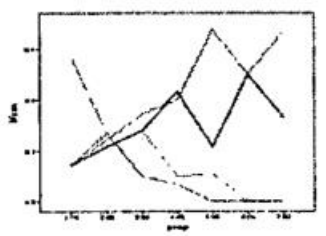
- **تصحيح المفردات** باستخدام برنامج الحاسب الآلي SPSS.

- **التحليل البياني للمفـردات:** Graphical of Item Analysis (GIA): ولإجـراء التحليـل البياني لمفردات اختبار التحصيل الدراسي تم تصحيح المفردات الاختبارية في نفس ملف البيانات السابق إعداده، ثم رسم الأشكال البيانية لمفردات الاختبار التحصيلي، ومن ثم فقد تم الحصول على عدد (٣٠) شكلا بيانيا. ثم قامت الباحثة بفحص هذه الأشكال كل على حدة للتأكد من حساسية تمييز البدائل بين التلاميذ مختلفي القدرة. وفي ضوء التحليل البياني للمفردات تبين أن هناك عدد مناسب من المفردات يتميز بخصائص سيكومترية مناسبة؛ حيث تمتعت بدرجات تمييز مناسبة والمشتتات المناسبة التي تميز بين الأفراد. كما تم تحديد عدد (٦) مفردات بها مشـكلات فـي خصائصـها السيكومترية.

ويتضمن شكل (٧، ٨) التاليان مثالا لمفردات الاختبار ذات التحليل البيـاني المناسـب ومثـالا للمفردات غير الملائمة من حيث التحليل البياني للمفردة.



شكل(٧): مفردة رقم ٢١ في الاختبار التحصيلي ذات تحليل بياني جيد.



شكل(٨): مفردة رقم ١١ في الاختبار التحصيلي ذات تحليل بياني غير ملائم

وهكذا فإنه في ضوء التحليل البياني للمفردات تم تحديد خصائص المفردات؛ كمؤشـر مبـدئي على جودتها قبل الدخول بها إلى التحليل الإحصائي بأحد نماذج نظرية الاستجابة للمفـردة؛ ممـا يساعد في اتخاذ القرار المناسب بشأن كل مفردة منها، إما بالإبقاء عليها أو تعديلها أو حذفها.

- **تحليل بيانات مفردات اختبار التحصيل الدراسي باستخدام برنامج WINSTEPS:** حللت بيانات اختبار التحصيل الدراسي وفقاً لنموذج راش باستخدام برنامج الحاسب الآلي WINSTEPS بهدف تحديد واستبعاد البيانات غير الملائمة لأسس القياس الموضوعي وتدريج المفردات المختلفة على ميزان تدرج واحد بصفر مشترك واحد، وكذلك الحصول على تقديرات القدرة للأفراد وتحديد تقدير القدرة المقابل لدرجة القطع لكل مستوى من مستويات الأداء.

وفيما يلي نتائج هذا التحليل بالتفصيل:

(١) **حذف البيانات التامة والصفرية:** يقوم برنامج WINSTEPS آليا باستبعاد البيانات التامة قبل

التحليل والتي تتضمن: كل فرد حصل على الدرجة النهائية أو الدرجة صفر على مجموعة المفردات التي أداها. وكذلك كل مفردة أجمع أفراد العينة على إجابتها سواء صواباً أو خطأ حيث تعتبر خارج نطاق القياس.

ولم يسفر التحليل الأولى لبيانات الاختبار التحصيلي الكلي عن استبعاد أي فرد أو مفردة تبعا لهذا المحك.

(٢) حذف الأفراد غير الملائمين لأسس القياس: بعد إجراء التحليل الأولى للبيانات باستخدام برنامج WINSTEPS تم تحديد وحذف الأفراد الذين تجاوزوا محكات الملاءمة الإحصائية (±٢) التي يوفرها برنامج WINSTEPS. وقد بلغ عدد الأفراد الذين حذفوا في ضوء هذه المحكات (٣٠) طالبا.

(٣) حذف المفردات غير الصادقة في تدرجها على متصل متغير التحصيل الدراسي: أعيد تحليل البيانات بعد حذف الأفراد غير الملائمين لأسس القياس الموضوعي، بهدف تحديد وحذف المفردات التي تجاوزت محكات الملاءمة الإحصائية (±٢) التي يوفرها برنامج WINSTEPS. وقد بلغ عدد المفردات التي تجاوزت القيمة (±٢) من تدريج إحصاءات الملاءمة (٦) مفردات. ويتضمن جدول (٤) التالي ملخصاً لنتائج تحليل البيانات وفقا للخطوات التي اتبعت في تدريج مفردات اختبار التحصيل الدراسي.

جدول (٤): ملخص نتائج تحليل بيانات الاختبار التحصيلي في مادة العلوم للصف الأول الإعدادي

مرحلة التحليل	عدد المفردات	عدد الأفراد	متوسط التقديرات		معامل ثبات التقديرات	
			للمفردات	للأفراد	للمفردات	للأفراد
التحليل الأولى قبل الحذف	٣٠	٢٠٠	٠.٠٠٠	-٠.٠٧	٠.٩٣	٠.٧٧
التحليل الثاني بعد حذف الأفراد	٣٠	١٧٠	٠.٠٠٠	٠.٠٦	٠.٩٥	٠.٧٦
التحليل الثالث بعد حذف المفردات	٢٤	١٧٠	٠.٠٠٠	٠.٢٨	٠.٩٤	٠.٧٧

٤- التدريج النهائي لمفردات اختبار التحصيل الدراسي: للحصول على التدريج النهائي لمفردات اختبار التحصيل الدراسي اتبعت الخطوات التالية:

- **إعادة تحليل البيانات للمرة الثالثة بعد حذف المفردات غير الملائمة بهدف:**

- تدريج مفردات اختبار التحصيل الدراسي على تدريج واحد مشترك بصفر واحد مشترك تبعاً لصعوبتها باللوجيت وتكوين التدريج النهائي لمفردات الاختبار التحصيلي.
- حساب تقديرات القدرة المقابلة لكل درجة خام كلية محتملة على الاختبار.

- تحويل تقديرات كل من الصعوبة والقدرة من وحدة (اللوجيت) إلى وحدة (المنف).

ويتضمن جدول (٥) التالي التدريج النهائي لمفردات اختبار التحصيل الدراسي في صورتها النهائية مرتبة تبعاً لصعوبتها بوحدتي اللوجيت والمنف، بالإضافة إلى الخطأ المعياري مقدراً بوحدتي اللوجيت والمنف، وذلك بعد حذف الأفراد والمفردات غير الملائمة لأسس القياس

الموضوعي.

جدول (٥): التدريج النهائي لصعوبات المفردات مقدرة بوحدتي اللوجيت والمنف

رقم المفردة	تقدير الصعوبة باللوجيت	تقدير الصعوبة بالمنف	الخطأ المعياري باللوجيت	الخطأ المعياري بالمنف
6	-2.01	39.95	0.24	1.2
8	-1.49	42.55	0.21	1.05
21	-0.86	45.7	0.19	0.95
4	-0.76	46.2	0.19	0.95
14	-0.76	46.2	0.18	0.9
20	-0.55	47.25	0.18	0.9
9	-0.37	48.15	0.18	0.9
16	-0.17	49.15	0.18	0.9
7	-0.06	49.7	0.18	0.9
24	0	50	0.18	0.9
10	0.05	50.25	0.17	0.85
23	0.05	50.25	0.18	0.9
3	0.07	50.35	0.18	0.9
26	0.12	50.6	0.17	0.85
5	0.13	50.65	0.17	0.85
17	0.22	51.1	0.17	0.85
29	0.22	51.1	0.17	0.85
12	0.44	52.2	0.18	0.9
27	0.65	53.25	0.18	0.9
18	0.69	53.45	0.18	0.9
13	0.77	53.85	0.18	0.9
30	1.03	55.15	0.18	0.9
15	1.06	55.3	0.19	0.95
28	1.53	57.65	0.2	1

ويتضح من جدول رقم (٥) أن مفردات اختبار التحصيل الدراسي وعددها(٢٤) مفردة تغطى مدى الصعوبة بشكل مناسب، كما يلاحظ أن هناك عدداً مناسباً من المفردات التي تغطي المستويات المختلفة من الصعوبات على مدى متصل الصعوبة؛ حيث لا توجد فجوات على متصل هذه الصعوبات، وذلك باعتبار الخطأ المعياري على جانبي تقدير الصعوبة سوى فجوتين متطرفتين عند طرفي التدريج. كما يتضح من هذا الجدول أيضاً أن مدى الصعوبة الذي تغطيه المفردات يمتد من (-٢.٠١: ١.٥٣) لوجيت أي (٣٩.٩٥: ٥٧.٦٥) منف.

وامتدت الأخطاء المعيارية لتقديرات الصعوبة بين (٠.١٧: ٠.٢٤) لوجيت وتعتبر قيم الأخطاء المعيارية – هذه – صغيرة نسبياً مما يدل على دقة وثبات تقديرات صعوبات المفردات.

ويتضمن جدول (٦) التالي تقديرات القدرة المقابلة لكل درجة كلية محتملة على اختبار التحصيل الدراسي مقدرة بوحدتي اللوجيت والمنف.

جدول (٦): تقديرات القدرة المقابلة لكل درجة خام مقدرة بوحدتي اللوجيت والمنف

الدرجة الخام	تقدير القدرة باللوجيت	تقدير القدرة بالمنف	الخطأ المعياري باللوجيت	الخطأ المعياري بالمنف
0	-4.71	26.45	1.85	9.25
1	-3.44	32.8	1.04	5.2
2	-2.66	36.7	0.76	3.8
3	-2.18	39.1	0.65	3.25
4	-1.8	41	0.58	2.9
5	-1.5	42.5	0.53	2.65
6	-1.23	43.85	0.5	2.5
7	-0.99	45.05	0.48	2.4
8	-0.77	46.15	0.46	2.3
9	-0.56	47.2	0.45	2.25
10	-0.37	48.15	0.44	2.2
11	-0.17	49.15	0.44	2.2
12	0.02	50.1	0.43	2.15
13	0.21	51.05	0.44	2.2
14	0.4	52	0.44	2.2
15	0.59	52.95	0.45	2.25
16	0.79	53.95	0.46	2.3
17	1.01	55.05	0.47	2.35
18	1.24	56.2	0.49	2.45
19	1.5	57.5	0.52	2.6
20	1.8	59	0.57	2.85
21	2.15	60.75	0.63	3.15
22	2.63	63.15	0.75	3.75
23	3.39	66.95	1.03	5.15
24	4.64	73.2	1.84	9.2

كما يتضح من جدول رقم (٦) أن مدى تقديرات القدرة المقابلة لكل درجة كلية محتملة على اختبار التحصيل الدراسي الكلى يتراوح بين (-٤.٧١: ٤.٦٤) لوجيت أي (٢٦.٤٥: ٧٣.٢) منف، وتعتبر هذه الحدود هي أقصى حدود للقدرة يمكن تقديرها باستخدام اختبار التحصيل الدراسي أو أي صورة اختبارية فرعية مسحوبة منه.

حساب صدق وثبات القياس:

أولا: حساب الصدق: اتبع في حساب صدق القياس طريقتان هما: صدق المحتوى، وصدق التدرج (صدق الملاءمة) كما يوفره " نموذج راش".

(١) **صدق المحتوى:** يشير صدق المحتوى إلى درجة التوافق بين مفردات الاختبار والأهداف أو محتوى جدول المواصفات وللتأكد من صدق المحتوى للصورة الأولية لاختبار التحصيل الدراسي الكلي اعتمدت الباحثة على صدق البناء، وصدق المحكمين. ويرتبط **صدق المحتوى** بخطة بناء الاختبار؛ حيث تم بناء الاختبار وفق خطة محددة تشمل تحديد المجال السلوكي الذي يغطيه الاختبار وما يتضمنه من تحليل للمحتوى وتحديد لأوجه التعلم المتضمنة فيه، والأهداف السلوكية

وما يترتب على ذلك من إعداد جدول المواصفات والذي في ضوئه يتم صياغة المفردات.

وقد سارت الباحثة في بنائها للصورة الأولية لاختبار التحصيل الدراسي في هذه الخطوات. وبهذا يكون قد تحقق النوع الأول من صدق المحتوى وهو **صدق البناء** كما يتمثل في إعداد المفردات بحيث تقيس متغيراً واحداً وهو التحصيل الدراسي في مادة العلوم، وأن تكون الإجابة على كل مفردة من مفردات الاختبار مستقلة عن الإجابة عن أي من مفرداته الأخرى.

أما بالنسبة لصدق المحكمين فقد قامت الباحثة بعرض مجموعة المفردات الاختبارية على مجموعة من المحكمين وذلك لمراجعتها، والحكم على مدى تحقيقها للأهداف السلوكية التي وضعت لقياسها ومدى ملاءمتها للمستوى المعرفي الذي وضعت لقياسه. وفي ضوء آراء المحكمين تم تعديل المفردات التي تحتاج إلى تعديل لتكون صالحة لقياس الهدف الذي تقيسه. هذا وبتحقق التطابق بين المفردات الاختبارية والأهداف التي تقيسها يتحقق صدق البناء للمفردات وبالتالي يتحقق صدق المحتوى.

(٢) صدق التدرج (صدق الملاءمة كما يوفره نموذج راش): ويقوم هذا النوع من الصدق على من تدرج مفردات مجموعة من الاختبارات – يفترض أنها تقيس نفس السمة – على ميزان تدرج واحد مشترك باستخدام نموذج "راش" يعنى أن هذه المفردات تحقق شرط أحادية القياس أي أنها تعرف متغيراً واحداً.

وتتحقق أحادية البعد بتحقق صدق المحتوى، وبملاءمة كل من الأفراد والمفردات لأسس القياس الموضوعي تبعاً لمحكات الملاءمة الخاصة ببرنامج WINSTEPS المستخدم في التحليل الإحصائي. حيث توضح هذه المحكات مدى قياس المفردة لما تقيسه باقي المفردات على متصل المتغير موضوع القياس، كما تبين مدى اتساق تدرج قدرة الفرد مع تدرج قدرات باقي الأفراد على هذا المتصل.

وكما سبق فقد تم حذف المفردات غير الملائمة من اختبار التحصيل الدراسي الكلي وفقا لتلك المحكات الإحصائية. وذلك لأسباب مختلفة كأن تعتمد المفردة على غيرها من المفردات أو أن تكون المفردة ضعيفة الصياغة أو غير ذلك. كما تم حذف الأفراد غير الملائمين وفقا لتلك المحكات الإحصائية، وذلك لأسباب مختلفة أيضاً كأن يلجأ الطالب إلى التخمين أو الغش أو يتسم أداؤه بالبطء الشديد أو السرعة أو يستجيب بعكس ما هو متوقع منه.

وبحذف الحالات غير الملائمة من الأفراد والمفردات يستبقى الأفراد الصادقون وتكون المفردات المتبقية هي المفردات الصادقة في تعريف المتغير _ **التحصيل الدراسي في مادة العلوم** _ مما يوفر الموضوعية في تقدير الأفراد على متصل التحصيل الدراسي، كما يوفرها نموذج

راش من خلال شروطه. وبالإضافة إلى المفردات المستبقاة يستبقى أيضاً الأفراد الصادقون في استجاباتهم على تلك المفردات – ويعنى هذا توفر شرطي الصدق والثبات لتقديرات كل من صعوبات مفردات الاختبار، وقدرات الأفراد أي يعنى تحقق صدق وثبات القياس.

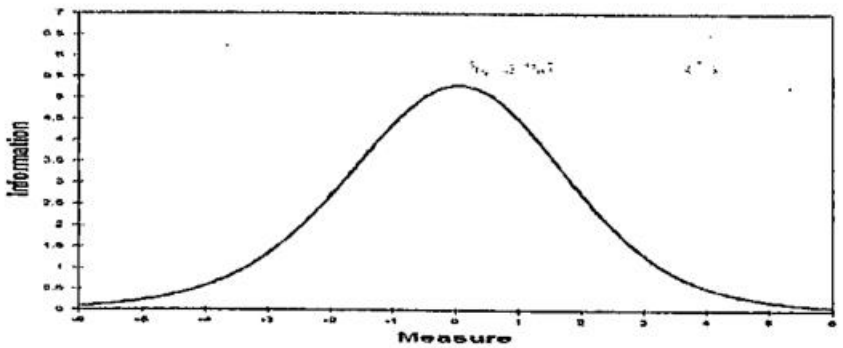
ثانياً ثبات القياس: اعتمد حساب ثبات القياس على دراسة كل من ثبات التدرج وحساب الخطأ المعياري لتقديرات كل من صعوبات المفردات وقدرات الأفراد.

- **ثبات التدرج:** إن تدرج مفردات اختبار التحصيل الدراسي الكلية معاً على ميزان تدرج واحد مشترك وفقا لنموذج "راش" – بعد حذف الحالات غير الملائمة من الأفراد والمفردات – يعنى تحقق شروط النموذج ومنها شرط استقلالية القياس، ويعنى ذلك ثبات تقديرات كل من الصعوبة والقدرة وعدم تأثرهما باختلاف الصورة الاختبارية المستخدمة والمسحوبة من الاختبار الكلي أو باختلاف عينة الأفراد المستخدمة في تدريجه (أمينة محمد كاظم، ٢٠٠٠، ٢٣١).

وقد حسب معامل ثبات تقدير كل من صعوبة المفردات وقدرة الأفراد باستخدام معامل ثبات **كيودريتشاردسون** الذي يوفره برنامج WINSTEPS. وتشير قيم معاملات الثبات الموضحة بجدول (٤) إلى أن اختبار التحصيل الدراسي الكلي يتميز بتقديرات ثابتة للصعوبة حيث بلغ معامل الثبات (٠.٧٧) تقريباً، كما يوفر تقديرات ثابتة للقدرة حيث بلغ معامل ثبات تقدير القدرة (٠.٩٤) تقريباً.

- **حساب الخطأ المعياري لتقديرات كل من صعوبات المفردات وقدرات الأفراد:** يوفر برنامج الحاسب الآلي WINSTEPS تقديرات للخطأ المعياري لتقدير قدرة كل فرد، وكذا صعوبة كل مفردة، والتي تعد مؤشر لدقة القياس. ويتضح من جدول(٥) أن قيم الخطأ المعياري لصعوبات المفردات تعتبر منخفضة بوجه عام، حيث تتراوح هذه القيم بين (٠.١٧: ٠.٢٤) لوجيت. كما يوضح جدول(٦) انخفاض قيم الخطأ المعياري لتقديرات القدرة المقابلة لكل درجة كلية محتملة على اختبار التحصيل الدراسي حيث تراوحت بين(٠.٤٣: ٠.٧٦) لوجيت، وذلك فيما عدا التقديرات المقابلة للدرجات المتطرفة جداً على المقياس (أعلى وأدنى درجة) حيث كانت قيم الخطأ المعياري المناظرة لهما(١.٨٤: ١.٨٥) لوجيت على الترتيب ولم يحصل أي من أفراد عينة التدريج على مثل هذه التقديرات المتطرفة. مما يزيد من ثبات.

كما يشير منحنى دالة المعلومات للاختبار التحصيلي شكل (٩) إلى تمتع الاختبار بدرجة عالية من الثبات.



شكل(٩): دالة معلومات اختبار التحصيل الدراسي في مادة العلوم

ثبات درجات الاختبار: بلغ معامل كرنباخ ألفا Cronbach's Alpha للاختبار ٠.٧٩٤ وتعد قيمة مرتفعة مما يشير ثبات درجات الاختبار.

الاتساق الداخلي: يتضح من جدول (١) ملحق (٥) أن مفردات الاختبار تمتعت بدرجة ارتباط مناسبة مع الدرجة الكلية للاختبار. فيما عدا ٦ مفردات؛ حيث لم تصل قيمة ارتباط هذه المفردات بالدرجة الكلية إلى ٠.٣ وهي المفردات أرقام (١-٢ -١١-١٩-٢٢-٢٥). ومما يجدر ذكره أن هذه المفردات جاءت غير ملائمة لأسس القياس الموضوعي أيضا عند التحليل الإحصائي وفق نظرية الاستجابة للمفردة.

- **صدق درجات الاختبار:** تم التحقق من صدق الاختبار بعدة طرق منها:

صدق المحكمين: بعد الانتهاء من بناء الاختبار ومراجعته تم عرضه على مجموعة من المتخصصين في مجال القياس النفسي والتربوي وتدريس العلوم. وأجريت التعديلات التي أشار إليها السادة المحكمون لبعض مفردات الاختبار.

التحقق من ملاءمة البيانات لنموذج راش الأحادي المعلم: تم الاعتماد على برنامج التحليل الإحصائي WINSTEPS في التحقق من ملاءمة البيانات لنموذج راش الأحادي المعلم. واستبعاد البيانات غير الملائمة لأسس القياس الموضوعي وتدريج المفردات الملائمة المتبقية على ميزان واحد بصفر مشترك واحد وقد تم حذف الأفراد غير الملائمين لأسس القياس الموضوعي، وكذلك المفردات غير الملائمة وفق محكات الملاءمة التي يوفرها البرنامج.

ثانيا: استبيان المحكمين حول ممارسات تحديد مستويات الأداء ودرجة القطع: تكون الاستبيان من عدة أجزاء، وجاءت عبارات الاستبيان على صيغتين الأولى مفردات تقرير ذاتي من نوع مقياس ليكرت متدرج الفئات وعددها (١٩) مفردة يستخدم لجمع البيانات الكمية لاستجابات المحكمين حول تحديد أي العوامل أكثر تأثيرا على قرارتهم. وجاء الشكل الآخر للمفردات على هيئة مفردات

مفتوحة النهاية للوقوف على ما يفكر فيه المحكمون أثناء عملية تحديد درجات القطع وبلغ عددها (٥) مفردة.

وتتناول عبارات الاستبيان ٣ محاور رئيسة تتمثل في:

المحور الأول دور المواصفات العامة والخاصة لمستويات الأداء في اتخاذ القرارات: تم تعريف المواصفات القياسية لمستويات الأداء في العلوم، والتي تميز الطالب ذا الحد الأدنى من الكفاية بحيث تكون الأساس لوضع تقديرات كل مفردة. من منظور محكي المرجع تماماً وهذا يشكل الركائز الأساسية لعمل التقدير لمستويات الكفاءة.

المحور الثاني دور طلاب المحكم في القرارات ويعتبر هذا منظور جماعي المرجع حيث يفكر المحكم من هذا المنظور ما إذا كانت المفردة الاختبارية سهلة أم تمثل تحدياً لطلابه. أكثر من التفكير في مهارات العلوم اللازمة للإجابة عن هذا السؤال وبغض النظر عما إذا كانت هذه المهارات متسقة مع مهارات العلوم اللازمة للطالب الكفء كما تم تعريفه بمواصفات مستوى الأداء.

المحور الثالث: دور بيانات التغذية الراجعة: حيث هناك ضرورة للوقوف على جدوى تزويد المحكمين ببيانات الأثر (مثل صعوبة المفردة، توزيع الطلاب وفق الدرجات) على المحكمين أثناء استخدام إجراءات طريقة أنجوف.

ويطبق الاستبيان على ثلاث مراحل، المرحلة الأولى وتتكون من ٧ عبارات يجيب عنها المحكمون بعد التدريب على إجراءات تحديد درجة القطع وقبل الجولة الأولى. المرحلة الثانية وتتكون من (٣) عبارات يتم الإجابة عنها بعد تقديرات الجولة الأولى. المرحلة الثالثة وتتكون من ١٢ عبارة يتم الإجابة عنها بعد تقديرات الجولة الثانية.

تم عرض الاستبيان على مجموعة من المحكمين للتحقق من ملاءمته لغرض القياس. وأجريت التعديلات في ضوء مقترحات السادة المحكمين.

إجراءات تنفيذ جولات تحديد درجة القطع:

١- تجهيز أدوات التدريب وتشمل:

- ✓ الإطار المرجعي للاختبار، والاختبار في صورته النهائية، ومفتاح الإجابة لكل سؤال (ملحق ١).
- ✓ صعوبات المفردات ملحق (٢).
- ✓ النسبة المئوية لاختيار كل بديل لكل مفردة (ملحق ٣).
- ✓ المنحنى البياني لكل مفردة (ملحق ٤).

✓ المواصفات القياسية العامة والخاصة للاختبار

حيث تم تحديد مستويات الأداء (العدد – مسمياتها) تم تحديد أربعة مستويات للأداء (متقدم-مرض-نام -متعثر) تفصلها ثلاث درجات قطع على الاختبار. وبمراجعة التراث النظري والأدبيات السابقة قامت الباحثة بوضع التوصيف العام لمستويات الأداء وعرضها على مجموعة من المحكمين لمراجعتها والتحقق من اتساقها مع المعايير القومية للتعليم وأهداف المرحلة الإعدادية. ويتضمن جدول (٧) التالي المواصفات العامة لمستويات الأداء في صورتها النهائية.

جدول (٧): المواصفات العامة لمستويات الأداء المستخدمة في الدراسة الحالية.

المستوى	التوصيف
يكون الطالب في الصف الأول الإعدادي من خلال دراسته لمادة العلوم قادرا على:	
متقدم	– يقدم أدلة وبراهين تؤيد آراء علمية متنوعة. – يظهر فهماً شاملاً ومتعمقاً وناقداً للمحتوى ويعطى حلولاً متنوعة للأسئلة غير النمطية. – يحل مشكلات مألوفة بطرق غير مألوفة. – يظهر مهارات تواصل فعالة وفهم متعمق يمكنه من ابتكار أفكار وحلول غير تقليدية لمدى واسع من الموضوعات.
مرض	– يظهر فهماً جيداً للمحتوى يتمثل في حل الأسئلة المألوفة التي تتطلب القدرة على الفهم والاستنتاج.
نام	– يظهر فهماً جزئياً للمحتوى يمكنه من حل الأسئلة البسيطة المعتادة.
متعثر	– يستطيع تذكر بعض المعلومات الأولية ويعجز عن حل الأسئلة البسيطة المعتادة.

في ضوء التوصيف العام لمستويات الأداء والإطار المرجعي للاختبار اشتق التوصيف الخاص لكل مستوى من مستويات الأداء الأربعة وعرضها على مجموعة من المحكمين وإجراء التعديلات المناسبة، ويتضمن جدول (٨) التالي المواصفات الخاصة لمستويات الأداء في صورتها النهائية.

جدول (٨): المواصفات الخاصة لمستويات الأداء المستخدمة في الدراسة الحالية

المستوى	التوصيف
	أن يكون الطالب في مستوى الأداء قادرا على أن:
متقدم	– يستنتج عدد مستويات الطاقة الممثلة بالإلكترونات من خلال رسم تخطيطي للذرة. – يستنتج عدد بروتونات ذرة ما من خلال رسم تخطيطي لهذه الذرة. – يستنتج عدد إلكترونات المستوى الخارجي لذرة ما بمعلومية العدد الكتلي والنيوترونات لهذه الذرة. – يستنتج تأثير الحرارة على إلكترونات المستوى الخارجي للذرة. – يتعرف تأثير الحرارة على خواص المادة. – يستنتج العدد الذري لعنصر ما بمعلومية عدد مستويات الطاقة المشغولة بالإلكترونات وإلكترونات المستوى الخارجي. – يستنتج أهمية عدد إلكترونات المستوى الخارجي في دخول الذرة في التفاعل الكيميائي.

مرض	- يستنتج خصائص المادة من بعض البيانات المعطاة عنها. - يميز بين جزيء العنصر وجزيء المركب. - يتعرف الرمز الكيميائي لبعض العناصر. - يتعرف الجسيمات المكونة للذرة من خلال بعض خواصها (مثل: شحنتها، أماكن تواجدها، كتلتها). - يستنتج عدد إلكترونات المستوى الخارجي لذرة ما من خلال بعض البيانات المعطاة (رسم تخطيطي للذرة، عدد البروتونات). - يميز بين الغازات من حيث عدد الذرات المكونة لجزيئاتها. - يستنتج أعداد بعض الجسيمات المكونة للذرة من خلال بعض البيانات المعطاة (رسم تخطيطي للذرة). - عدد الإلكترونات السالبة لذرة ما من خلال رسم تخطيطي لها. - يستنتج عدد النيوترونات المتعادلة لذرة ما من خلال رسم تخطيطي لها.
نام	- يتعرف بعض المفاهيم المرتبطة بالذرة والتركيب الذري (مثل: الجزيء – العنصر – الذرة، العدد الذري). - يتعرف بعض المفاهيم المرتبطة بتحولات المادة (مثل: الانصهار، التجمد). - يتعرف بعض الحقائق البسيطة المرتبطة بالذرة والتركيب الذري (مثل: عدد مستويات الطاقة، ترتيب مستويات الطاقة). - يتعرف تركيب بعض المركبات المعروفة (مثل تركيب جزئ الماء).

٢- عقد ورشة العمل لتحديد درجات القطع من خلال الخطوات التالية:

- **تدريب المحكمين على إجراءات تحديد درجة القطع:** ويبدأ ذلك بمناقشة المقصود بمعايير التعلم والمواصفات الخاصة والعامة للأداء ودرجة القطع والإجراءات العامة المتبعة لتحديد درجة القطع.
- **مراجعة توصيفات مستويات الأداء العامة والخاصة** مع المحكمين حيث تم تزويد المحكمين بتوصيف مستويات الأداء العامة والخاصة ومناقشتها.
- **تزويد المحكمين بمواد الاختبار:** وتشمل (الاختبار– نواتج التعلم التي يقيسها الاختبار) مع إتاحة الفرصة لهم بالتعرف على الاختبار من خلال قيام كل محكم بالإجابة عن كل سؤال بشكل مستقل. حيث تتم مناقشة محتوى الاختبار مع المحكمين من حيث الهدف نوع المفردات ومتطلبات الإجابة عنها وتقديم الاختبار للمحكمين للإجابة عنه حتى يستدخل كل منهم العمليات المعرفية المتطلبة للإجابة عن كل سؤال.
- الانتقال إلى مناقشة مفهوم الطالب ذي الحد الأدنى من الأداء وهو الطالب ذو الأداء المقبول فقط عندما يقاس في مقابل المواصفات محل الاهتمام. وهذه الخطوة مهمة جداً في عملية وضع المواصفات القياسية للأداء لأن التناغم مع مفهوم الطالب ذي الحد الأدنى من الأداء يعتبر أساسياً لكل العمل الذي يقوم به المحكمون وفق طريقة أنجوف لتحديد درجة القطع.
- الاتفاق على **القواعد الأساسية للعمل الجماعي وتشمل** عدم مناقشة أي من فلسفة التقييم التربوي، أو الأساس العلمي والفلسفي لمستويات الأداء، أو إجراءات تحديد المواصفات القياسية، أو الأهداف التي بني عليها الاختبار. مع مراعاة أن يكون الاسم مجهولاً أثناء العرض، ويكون هناك استقلال في الحكم، أن يعتمد الحكم على مهنية المحكم، مع مراعاة أن الإجماع ليس ضرورياً.
- تدريب المحكمين على إجراءات الحكم باستخدام ٥ مفردات تدريبية تم اختيارها لتمثل

المفردات التي سيحكم عليها المحكمون في المرحلة الأساسية من حيث نوع المفردات ومستوى صعوبتها.

- يراجع المحكمون المفردات واحدة تلو الأخرى مع تسجيل أفكارهم المتوقعة ثم تسلم تقاريرهم لقائد المجموعة ويتم تدوينها على سبورة ثم تناقش الأسباب لإعطاء التقديرات التي قاموا بها. وكان السؤال حول توقعات الأداء هو:

هل يستطيع الطالب ذو الحد الأدنى من الكفاية أن يجيب على هذه المفردة صواباً؟

ويتم تسجيل إجابة المحكم كرقم صحيح (صفر أو ١).

-بعد المرور عبر كل مفردة، وتزويد المحكمين بالأفكار المبدئية ومناقشة الأحكام مع المجموعة تم تزويد المحكمين بالبيانات التجريبية التي تشير إلى كيف يؤدى الأفراد فعليا على كل مفردة من مفردات الاختبار يتم تزويد المحكمين بنوعين من بيانات الأداء هما:

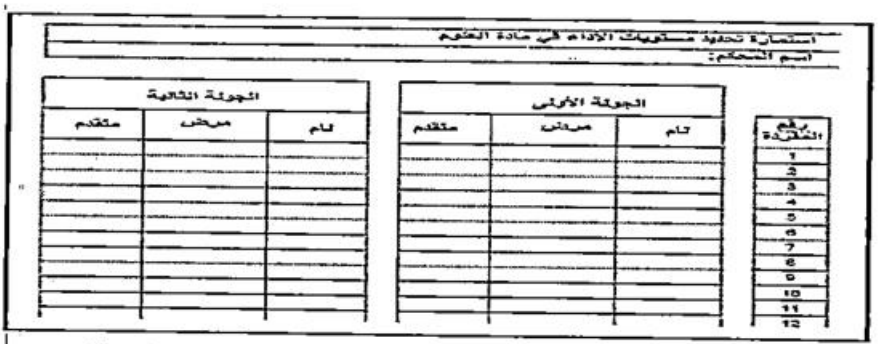
- ✓ الرسم البياني لكل مفردة ويوضح كيف يؤدى الطلاب على هذه المفردة كما يعبر عنه بالدرجة الكلية للأفراد على الاختبار.
- ✓ النسبة المئوية للممتحنين الذين اختاروا كل بديل من بدائل الإجابة لكل مفردة.

-طلب من المحكمين أن يراجعوا بيانات الأداء، ويراجعوا المفردة وأحكامهم المبدئية مرة ثانية ثم يقرروا ما إذا كانوا يعتقدون أنهم بحاجة إلى أن يغيروا أحكامهم القبلية أم لا بناء على هذه البيانات.

-بعد مراجعة البيانات للمفردة طلب من المحكمين أن يعطوا حكماً نهائياً_ قد يكون هو نفس التقدير الأول أو يختلف عنه _ ومرة ثانية فإن الأحكام يتم تسجيلها ومناقشتها.

-تطبيق الجزء الأول من الاستبيان على المحكمين.

-تنفيذ الجولة الأولى: بعد انتهاء إجراءات التحكيم على المفردات السابقة يتم تنفيذ إجراءات الجولة الأولى على عدد (٢٤) مفردة تمثل مفردات اختبار التحصيل الدراسي الذي طبق في الميدان، كل بشكل فردى مستقل دون أي مناقشات بينهم. ويطلب من كل منهم تدوين استجاباته في استمارة خاصة لذلك كما هو موضح بشكل (١٠) التالي.



شكل (١٠): الاستمارة المستخدمة في رصد تقديرات المفردات عبر جولتي التقدير

- إعلان نتائج الجولة الأولى ومناقشتها.
- تقديم بيانات التغذية الراجعة للمحكمين: من خلال استجابات الطلاب الفعلية على الاختبارات للمحكمين وتشمل نسب الطلاب المجتازين على درجة القطع المحددة بالجولة الأولى وصعوبات المفردات. وتكرارات البدائل لكل سؤال وكذلك البيانات المرتبطة بدرجة القطع لكل محكم ومتوسط ووسيط ومنوال درجات القطع، وكذلك القيم العظمى والصغرى للانحراف المعياري لدرجات القطع.
- تطبيق الجزء الثاني من الاستبيان على المحكمين.
- تنفيذ الجولة الثانية: وفيها يقوم كل محكم مرة ثانية بإصدار حكم فردي على كل مفردة من مفردات الاختبار وتحسب درجة القطع لكل محكم على الاختبار ثم إعلان نتائج الجولة الثانية ومناقشتها وتقديم التغذية الراجعة.
- تطبيق الجزء الثالث من الاستبيان على المحكمين.
- إنهاء ورشة العمل وتقديم الشكر للسادة المشاركين.
- التحليل الإحصائي للبيانات، والإجابة عن أسئلة الدراسة.

البرامج والأساليب الإحصائية المستخدمة: استخدمت الحزمة الإحصائية للعلوم الاجتماعية SPSS لتجهيز البيانات للتحليل باستخدام نظرية الاستجابة للمفردة، وإجراء الاختبارات الإحصائية الكلاسيكية اللازمة. واستخدم برنامج الأكسيل المستخدم في إجراء الإحصاءات المستخدمة في التقدير. واستخدم برنامج WINSTEPS في تحليل البيانات وفق نموذج راش أحادي المعلم.

واستخدمت مجموعة من الإحصاءات للإجابة عن تساؤلات الدراسة وتتضمن: الإحصاءات الوصفية (المتوسط، الانحراف المعياري) لبارامترات الأفراد والمفردات. والارتباط الثنائي الأصيل لحساب ارتباط كل مفردة بالدرجة الكلية للاختبار. التكرارات والنسب المئوية،

والرسم البياني للمفردات. بالإضافة إلى اختبار "مان ويتني"، واختبار ويلكوكسون.

نتائج الدراسة ومناقشتها:

يتناول هذا الجزء الإجابة عن أسئلة الدراسة ومناقشتها. وتتضمن النتائج عرض لـتأثير بيانات التغذية الراجعة على تقدير درجة القطع، وكذلك تأثيرها على ممارسات المحكمين خلال عملية تحديد درجة القطع.

أولا: النتائج الخاصة بتأثير بيانات التغذية الراجعة على تقدير درجة القطع.

الإجابة عن السؤال الأول وينص على:

ما درجات القطع المحددة من قبل المحكمين عينة الدراسة لكل مستوى من مستويات الأداء (متقدم- مرض-نام) عبر جولتي التحكيم؟

يتضمن جدول (٩) التالي درجات القطع لكل محكم من المحكمين عينة الدراسة لكل مستوى من مستويات الأداء عبر جولتي التحكيم الأولى والثانية.

جدول (٩): درجات القطع لكل محكم بالنسبة لمستويات الأداء الثلاث عبر جولتي التحكيم

العينة	كود المحكم	درجة القطع للجولة الأولى			درجة القطع للجولة الثانية		
		نام	مرض	متقدم	نام	مرض	متقدم
المعلمين	١	9	21	24	8	11	21
	٢	9	19	24	2	10	23
	٣	11	16	24	6	9	15
	٤	12	19	24	9	16	20
	٥	13	23	24	8	17	21
	٦	9	19	24	8	15	20
	٧	9	16	24	1	14	22
	٨	13	18	24	5	11	24
	٩	6	17	24	6	18	22
أعضاء هيئة التدريس	١٠	12	16	24	9	11	18
	١١	7	17	24	4	14	20
	١٢	10	19	24	3	11	18
	١٣	3	10	20	3	16	22
	١٤	9	19	24	4	10	19
	١٥	9	19	24	5	12	19

يتضح من جدول (٩): انخفاض تقديرات درجة القطع في الجولة الثانية عنها بالنسبة للجولة الأولى لكافة المحكمين وكافة مستويات الأداء. فيما عدا المحكم (كود ٩) من عينة المحكمين والمحكم (كود ١١) من أعضاء هيئة التدريس؛ حيث ظلت تقديراتهم لدرجة القطع في المستوى

"تام" كما هي دون تغيير في الجولتين.

ويتضمن جدول (١٠) التالي الوصف الإحصائي (المتوسط، الوسيط، المنوال، الانحراف المعياري، القيمة العظمى، والقيمة الصغرى) لدرجات القطع لكل مستوى من مستويات الأداء لكل من جولتي التحكيم بالنسبة لعينة الدراسة من أعضاء هيئة التدريس والمعلمين والعينة ككل.

جدول (١٠): الوصف الإحصائي لدرجات القطع عبر جولتي التحكيم لدى المحكمين عينة الدراسة

الإحصاءة	العينة	الجولة الأولى			الجولة الثانية		
		تام	مرض	متقدم	تام	مرض	متقدم
المتوسط	المعلمون	10.11	18.67	24.00	5.89	13.44	20.89
	أعضاء هيئة التدريس	8.33	16.67	23.33	4.67	12.33	19.33
	العينة ككل	9.40	17.87	23.73	5.40	13.00	20.27
الوسيط	المعلمون	9	19	24	6	14	21
	أعضاء هيئة التدريس	9	18	24	4	12	19
	العينة ككل	9	19	24	5	12	20
المنوال	المعلمون	9	19	24	8	11	21
	أعضاء هيئة التدريس	9	19	24	4	11	18
	العينة ككل	9	19	24	8	11	20
الانحراف المعياري	المعلمون	2.32	2.29	0.00	2.80	3.28	2.57
	أعضاء هيئة التدريس	3.08	3.50	1.63	2.25	2.25	1.51
	العينة ككل	2.69	2.90	1.03	2.59	2.88	2.28
القيمة الصغرى	المعلمون	6	16	24	1	9	15
	أعضاء هيئة التدريس	3	10	20	3	10	18
	العينة ككل	3	10	20	1	9	15
القيمة العظمى	المعلمون	13	23	24	9	18	24
	أعضاء هيئة التدريس	12	19	24	9	16	22
	العينة ككل	13	23	24	9	18	24

يتضح من جدول (١٠) السابق ما يلي:

١- بالنسبة لقيم درجات القطع المقدرة لمستويات الأداء (متقدم-مرض-تام) خلال جولتي التقدير الأولى والثانية:

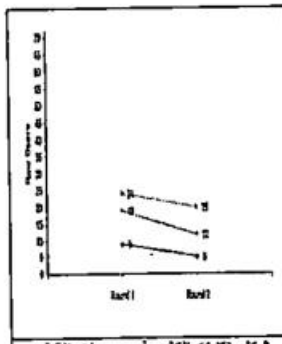
- بالنسبة لقيم درجات القطع المقدرة في الجولة الأولى: تراوحت درجة القطع لمستوى "تام" بين (٦: ١٣) بالنسبة لعينة المعلمين، وتراوحت بين (٣: ١٢) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (٣: ١٣) بالنسبة للعينة ككل. وتراوحت درجة القطع لمستوى "مرض" بين (١٦: ٢٣) بالنسبة لعينة المعلمين، وتراوحت بين (١٠: ١٩) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (١٠: ٢٣) بالنسبة للعينة ككل. وبلغت درجة القطع لمستوى "متقدم" (٢٤) بالنسبة لعينة المعلمين، وتراوحت بين (٢٠: ٢٤) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (٢٠: ٢٤) للعينة ككل.

- بالنسبة لقيم درجات القطع المقدرة في الجولة الثانية: تراوحت درجة القطع لمستوى "تام" بين

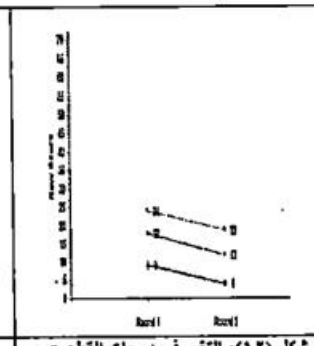
(١: ٩) بالنسبة لعينة المعلمين، وتراوحت بين (٣: ٩) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (١: ٩) بالنسبة للعينة ككل. وتراوحت درجة القطع لمستوى "مرض" بين (٩: ١٨) بالنسبة لعينة المعلمين، وتراوحت بين (١٠: ١٦) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (٩: ١٨) بالنسبة للعينة ككل. وتراوحت درجة القطع لمستوى "متقدم" بين (١٥: ٢٤) بالنسبة لعينة المعلمين، وتراوحت بين (١٨: ٢٢) بالنسبة لأعضاء هيئة التدريس، وتراوحت بين (١٥: ٢٤) للعينة ككل.

٢- **بالنسبة لوسيط درجات القطع:** استخدم وسيط درجات القطع للتعبير عن درجة القطع المعبرة عن مجموعة من المحكمين حيث إن الوسيط هو **الإحصاءة الأنسب عند التعامل مع القيم المتطرفة**. ويلاحظ من قيم وسيط درجات القطع لمستويات الأداء (نام، مرض، متقدم) ما يلي:

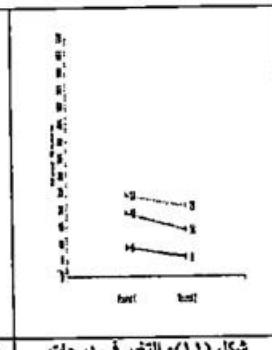
- انخفاض قيمة وسيط درجات القطع لمستويات الأداء (نام -لمرض-متقدم) بالنسبة لكل من عينة المعلمين وأعضاء هيئة التدريس والعينة ككل في الجولة الثانية عنها في الجولة الأولى. وتوضح الأشكال التالية من (١١: ١٣) التغير في وسيط درجات القطع لمستويات الأداء عبر جولتي التقدير.



شكل (١٣) التغير في درجات القطع عبر جولتي التقدير بالنسبة للعينة ككل



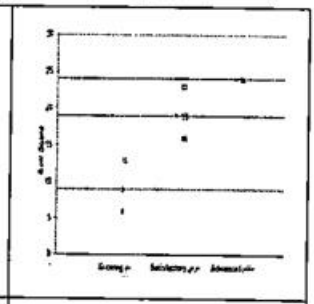
شكل (١٢): التغير في درجات القطع عبر جولتي التحكيم لعينة أعضاء هيئة التدريس



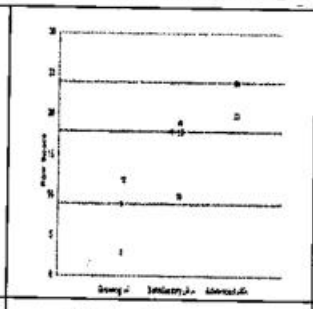
شكل (١١): التغير في درجات القطع عبر جولتي التحكيم لعينة المعلمين

- هناك اتفاق في قيمة الوسيط لمستوى الأداء "تام" ومستوى الأداء "متقدم" بين عينة المحكمين من المعلمين وأعضاء هيئة تدريس والعينة ككل في **الجولة الأولى** حيث بلغت قيمته بالنسبة لمستوى نام (٩ درجات) وبلغت قيمته بالنسبة لمستوى "متقدم" (٢٤ درجة).
- كان هناك تقارب كبير بين قيمة الوسيط لمستوى "مرض" في الجولة الأولى بين عينة المحكمين من المعلمين وأعضاء هيئة التدريس والعينة ككل حيث بلغت قيمته (١٨) درجة

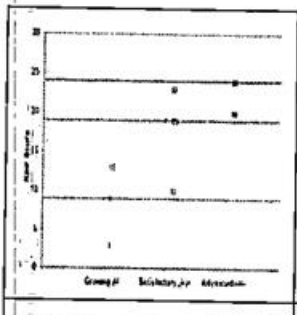
بالنسبة لعينة أعضاء هيئة التدريس و(١٩) درجة بالنسبة للمعلمين والعينة ككل. وتوضح الأشكال التالية من (١٤: ١٦) قيم الوسيط لمستويات الأداء الثلاثة في الجولة الأولى بالنسبة لعينة المحكمين وأعضاء هيئة التدريس والعينة ككل.



شكل(١٤): وسيط درجات القطع لمستويات الأداء في الجولة الأولى لعينة المعلمين.

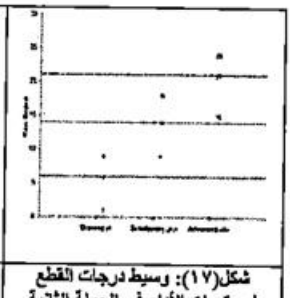


شكل(١٥): وسيط درجات القطع لمستويات الأداء في الجولة الأولى لعينة أعضاء هيئة التدريس.

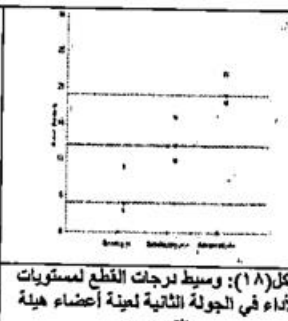


شكل(١٦): وسيط درجات القطع لمستويات الأداء في الجولة الأولى للعينة ككل

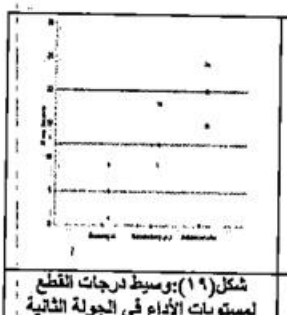
- كما جاء التقارب بين قيم وسيط درجات القطع لعينة المحكمين من المعلمين وأعضاء هيئة التدريس والعينة ككل في الجولة الثانية، حيث تراوحت قيمته بين (٥: ٦) درجات بالنسبة لمستوى نام، وتراوحت قيمته بين (١٢: ١٤) درجة بالنسبة لمستوى مرض، وتراوحت قيمته بين (١٩: ٢١) درجة بالنسبة لمستوى متقدم. وتوضح الأشكال من (١٧: ١٩) وسيط درجات القطع للجولة الثانية لعينة المحكمين من أعضاء هيئة التدريس والمعلمين والعينة ككل. وجاء وسيط التقديرات لأعضاء هيئة التدريس في الجولة الثانية أقل من وسيط تقديرات عينة المعلمين وتتفق هذه النتيجة مع نتيجة دراسة (Halpin etal, 1985)(خالد التميمي، ١٩٩٩) ، (Hurtz & Auorbach,2003).



شكل(١٧): وسيط درجات القطع لمستويات الأداء في الجولة الثانية لعينة المعلمين



شكل(١٨): وسيط درجات القطع لمستويات الأداء في الجولة الثانية لعينة أعضاء هيئة التدريس.



شكل(١٩):وسيط درجات القطع لمستويات الأداء في الجولة الثانية للعينة ككل

ويتضمن جدول (١١) التالي وسيط درجات القطع لمستويات الأداء خلال جولتي التحكيم بالنسبة لأعضاء هيئة التدريس والمعلمين والعينة ككل، وما يقابله من تقديرات وفق نموذج راش.

جدول (١١): تقديرات القدرة بالمنف المقابلة لوسيط درجات القطع لمستويات الأداء خلال جولتي التحكيم

العينة	الجولة الأولى						الجولة الثانية					
	نام		مرض		متقدم		نام		مرض		متقدم	
	درجة خام	تقدير القدرة بالمنف	درجة خام	تقدير القدرة بالمنف	درجة خام	تقدير القدرة بالمنف	درجة خام	تقدير القدرة بالمنف	درجة خام	تقدير القدرة بالمنف	درجة خام	تقدير القدرة بالمنف
المعلمون	9	٤٧.٢	19	٥٧.٥	24	٧٣.٢	6	٤٣.٨٥	14	٥٢	21	٦٠.٧٥
هيئة التدريس	9	٤٧.٢	18	٥٦.٢	24	٧٣.٢	4	٤١	12	٥٠.١	19	٥٧.٥
العينة ككل	9	٤٧.٢	19	٥٧.٥	24	٧٣.٢	5	٤٢.٥	12	٥٠.١	20	٥٩

يتضح من جدول (١١) السابق أن تقديرات القدرة المقابلة لوسيط درجات القطع قد امتدت لتغطي معظم ميزان تقدير القدرة بوحدة المنف حيث تراوحت تقديرات القدرة لوسيط درجات القطع بين (٤١: ٧٣.٢) منف.

٣- بالنسبة للتجانس بين المحكمين:

- يتضح التجانس بين المحكمين من عدة زوايا منها تقارب قيم مقاييس: المتوسط – الوسيط- المنوال – الانحراف المعياري لكل مستوى من مستويات الأداء في كل جولة من جولتي التقدير كل على حدة.
- رغم انخفاض قيم الانحراف المعياري في الجولة الأولى لعينة المعلمين عنها بالنسبة لعينة أعضاء هيئة التدريس؛ إلا أن الوضع تغير قليلا في الجولة الثانية بعد تقديم بيانات الأثر؛ حيث ارتفعت قليلا قيم الانحراف المعياري لتقديرات درجات القطع لكل مستوى من مستويات الأداء في جولة التحكيم الثانية بالنسبة لقيمتها قبل تقديم بيانات الأثر بالنسبة لعينة المعلمين؛ وعلى النقيض من ذلك فقد انخفض إلى حد ما الانحراف المعياري لدرجات القطع لأعضاء هيئة التدريس بعد تقديم بيانات التغذية الراجعة عنه قبل تقديمها. وقد يرجع ذلك إلى أن المعلمين _ بحكم طبيعة عملهم_ أقل تعاملا مع البيانات الإحصائية فكان الارتفاع قليلا في الانحراف المعياري.
- ولاختبار الاتساق بين المحكمين إحصائيا استخدمت الباحثة اختبار الاتساق بين المحكمين Interclass Correlation كما هو موضح بجدول رقم (١٢) التالي.

جدول(١٢): الاتساق بين المحكمين في كل جولة من جولتي التحكيم وعبر مستويات الأداء

الجولة ومستوى الأداء	التربويين		المعلمين		العينة ككل	
	معامل ألفا كرونباخ	معامل الاتساق بين المحكمين	معامل ألفا كرونباخ	معامل الاتساق بين المحكمين	معامل ألفا كرونباخ	معامل الاتساق بين المحكمين
مستوى الأداء التام في الجولة الأولى	.824	.438**	.913	.538**	.940	.513**
مستوى الأداء التام في الجولة الثانية	.631	.222**	.749	.249**	.828	.244**
مستوى الأداء المرضي في الجولة الأولى	.634	.224**	.791	.297**	.845	.267**
مستوى الأداء المرضي في الجولة الثانية	.763	.349**	.806	.316**	.864	.297**
مستوى الأداء المتقدم في الجولة الثانية*	.866	.518**	.818	.334**	.910	.402**

يتضح من جدول (١٢) ما يلي:

- جاء الاتساق بين المحكمين في تقدير درجة القطع مرتفعا كما يشير إليه معامل ألفا؛ حيث تراوحت قيمته بين (٠.٨٢٨: ٠.٩٤٠) بالنسبة للعينة ككل، وتراوحت قيمته بين (٠.٦٣١:٠.٨٦٦) بالنسبة لعينة أعضاء هيئة التدريس، وتراوحت قيمته بين (٠.٧٤٩:٠.٩٣١) بالنسبة لعينة المعلمين.

- كما جاءت قيمة معامل الاتساق بين المحكمين Intraclass Correlation دالة إحصائيا عند مستوى ٠.٠١ بالنسبة لتقديرات درجات القطع في المستويات الثلاثة، وذلك بالنسبة لعينة الدراسة ككل، وأعضاء هيئة التدريس، والمعلمين خلال جولتي التقدير الأولى والثانية.

الإجابة عن السؤال الثاني وينص على:

هل يختلف وسيط درجات القطع لمستويات الأداء الثلاث باختلاف خبرة المحكم في الجولة الأولى قبل تقديم بيانات التغذية الراجعة؟

للإجابة عن هذا السؤال استخدم اختبار "مان ويتني" للفروق بين وسيط تقديرات درجات القطع لعينة المعلمين وعينة أعضاء هيئة التدريس بالنسبة لكل مستوى من مستويات الأداء في الجولة الأولى. ويتضمن جدول (١٣) الفروق في وسيط درجات القطع بين تقديرات أعضاء هيئة التدريس والمعلمين في جولة التقدير الأولى.

* تعذر حساب الارتباط في هذا المستوى نظرا لعدم وجود تباين بين المقدرين في درجات القطع المقدرة في هذا المستوى في الجولة الأولى.

جدول (١٣): الفروق في درجات القطع بين أعضاء هيئة التدريس والمعلمين في جولة التقدير الأولى

مستوى الأداء	المجموعة	متوسط الرتب	مجموع الرتب	قيمة مان ويتني U	قيمة Z	الدلالة
نام	أعضاء هيئة التدريس	6.58	39.50	18.5	-1.037	.328
	المعلمين	8.94	80.50			
مرض	أعضاء هيئة التدريس	6.83	41.00	20	-0.856	.456
	المعلمين	8.78	79.00			
متقدم	أعضاء هيئة التدريس	7.25	43.50	22.5	-1.225	.607
	المعلمين	8.50	76.50			

يتضح من جدول (١٣): عدم وجود فروق دالة إحصائيا في تقديرات درجة القطع كما حددها أعضاء هيئة التدريس وتقديرات درجة القطع كما حددها المعلمون في أي من مستويات الأداء الثلاثة (نام-مرض-متقدم) في جولة التقدير الأولى قبل تقديم بيانات التغذية الراجعة.

<u>الإجابة عن السؤال الثالث وينص على:</u>

هل يختلف وسيط درجات القطع لمستويات الأداء الثلاث باختلاف خبرة المحكم في الجولة الثانية بعد تقديم بيانات التغذية الراجعة؟

ويتضمن جدول (١٤) الفروق في وسيط درجات القطع بين تقديرات أعضاء هيئة التدريس والمعلمين في جولة التقدير الثانية بعد تقديم بيانات التغذية الراجعة.

جدول (١٤): الفروق في درجات القطع بين أعضاء هيئة التدريس والمعلمين في جولة التقدير الثانية

مستوى الأداء	المجموعة	متوسط الرتب	مجموع الرتب	قيمة مان ويتني U	قيمة Z	الدلالة
نام	أعضاء هيئة التدريس	6.67	40.00	19	-0.95	.388
	المعلمين	8.89	80.00			
مرض	أعضاء هيئة التدريس	7.25	43.50	22.5	- 0.537	.607
	المعلمين	8.50	76.50			
متقدم	أعضاء هيئة التدريس	5.50	33.00	12	- 1.785	.088
	المعلمين	9.67	87.00			

يتضح من جدول (١٤): عدم وجود فروق دالة إحصائيا في تقديرات درجة القطع كما حددها أعضاء هيئة التدريس وتقديرات درجة القطع كما حددها المعلمين في أي من مستويات الأداء (نام-مرض-متقدم) في جولة التقدير الثانية أي بعد تقديم بيانات التغذية الراجعة. وتتفق هذه النتيجة مع نتائج دراسة خالد التميمي (١٩٩٩)، ودراسة صلاح علام (١٩٩١)، ودراسة (Behuniak,etal.,1981). وتختلف مع نتيجة دراسة (Halpin;Sigmon&Halpin,1983)

حيث أشارت إلى اختلاف تقديرات المعلمين بشكل ملحوظ عن تقديرات أقرانهم من أعضاء هيئة التدريس ومعاونيهم. وتشير هذه النتائج إلى أن تقديم بيانات التغذية الراجعة لم يكن له تأثير على التجانس بين تقديرات المحكمين من المعلمين وأعضاء هيئة التدريس. وقد يرجع ذلك إلى تقنين إجراءات التدريب من ناحية، ووضوح تعليمات استخدام التغذية الراجعة للمجموعتين من ناحية أخرى.

الإجابة عن السؤال الرابع وينص على:

هل تختلف تقديرات درجة القطع للمحكمين عينة الدراسة (المعلمون-أعضاء هيئة التدريس-العينة ككل) في الجولة الثانية عنها في الجولة الأولى بالنسبة لكل مستوى من مستويات الأداء؟

للإجابة عن هذا السؤال استخدم اختبار ويلكوكسون لفروق الرتب قبل تقديم بيانات التغذية الراجعة لدراسة الفروق في تقديرات المحكمين عينة الدراسة (المعلمون-أعضاء هيئة التدريس-العينة ككل) لدرجات القطع في الجولة الأولى عنها في الجولة الثانية بعد تقديم بيانات التغذية الراجعة. ويتضمن جدول (١٥) هذه النتائج.

جدول (١٥): الفروق في درجات القطع لمستويات الأداء لعينة المعلمين في جولتي التقدير الأولى والثانية

العينة	مستوى الأداء	الرتب	متوسط الرتب	مجموع الرتب	قيمة z	الدلالة
المعلمون	نام	الرتب السالبة	4.50	36.00	-2.530	.011
		الرتب الموجبة	.00	.00		
	مرض	الرتب السالبة	5.50	44.00	-2.549	.011
		الرتب الموجبة	1.00	1.00		
	متقدم	الرتب السالبة	4.50	36.00	-2.530	.011
		الرتب الموجبة	.00	.00		
أعضاء هيئة التدريس	نام	الرتب السالبة	3.00	15.00	-2.032	.042
		الرتب الموجبة	.00	.00		
	مرض	الرتب السالبة	3.60	18.00	-1.572	.116
		الرتب الموجبة	3.00	3.00		
	متقدم	الرتب السالبة	4.00	20.00	-2.003	.045
		الرتب الموجبة	1.00	1.00		
العينة ككل	نام	الرتب السالبة	7.00	91.00	-3.191	.001
		الرتب الموجبة	.00	.00		
	مرض	الرتب السالبة	8.58	111.50	-2.929	.003
		الرتب الموجبة	4.25	8.50		
	متقدم	الرتب السالبة	7.85	102.00	-3.116	.002
		الرتب الموجبة	3.00	3.00		

يتضح من جدول (١٥) ما يلي:

- وجود فروق دالة إحصائيا في قيمة وسيط درجة القطع المقدرة من قبل المعلمين لكل مستوى من مستويات الأداء في الجولة الأولى عنها في الجولة الثانية لصالح تقديرات الجولة الأولى؛ حيث جاءت تقديرات درجة القطع في الجولة الأولى أعلى. وتشير هذه النتائج إلى تأثير بيانات التغذية الراجعة على تقديرات عينة المعلمين بالنسبة لجميع مستويات الأداء.

- وجود فروق دالة إحصائيا في قيمة وسيط درجة القطع المقدرة من قبل أعضاء هيئة التدريس لمستوى نام ومستوى متقدم من مستويات الأداء في الجولة الأولى عنها في الجولة الثانية لصالح تقديرات الجولة الأولى حيث جاءت تقديرات درجة القطع في الجولة الأولى أعلى، في حين لم تصل الفروق في التقديرات إلى حد الدلالة الإحصائية بالنسبة لمستوى الأداء "مرض".

- وجود فروق دالة إحصائيا في قيمة وسيط درجة القطع المقدرة من قبل العينة ككل لكل مستوى من مستويات الأداء في الجولة الأولى عنها في الجولة الثانية لصالح تقديرات الجولة الأولى؛ حيث جاءت تقديرات درجة القطع في الجولة الأولى أعلى. وتشير هذه النتائج إلى تأثير بيانات التغذية الراجعة على تقديرات العينة ككل بالنسبة لجميع مستويات الأداء.

وتشير هذه النتائج إلى التأثير البالغ لبيانات التغذية الراجعة على تقديرات درجة القطع في الجولة الثانية. وتتفق هذه النتائج مع دراسة (Hurtz& Auerbach,2003) وتختلف مع دراسة (Hambleton,2001) حيث أشارت إلى أن الأثر الرئيس لبيانات التغذية الراجعة قد يكون نفسيا أكثر منه سيكومتريا.

<u>الإجابة عن السؤال الخامس وينص على:</u>

ما نسبة الطلاب في كل مستوى من مستويات الأداء عند استخدام درجة القطع المحددة لكل مستوى في جولتي التحكيم الأولى والثانية؟

للإجابة عن هذا السؤال تم حساب النسبة المئوية للطلاب في كل مستوى من مستويات الأداء عندما تطبق درجة القطع التي أسفرت عنها كل جولة من جولتي التحكيم بالنسبة لعينة الدراسة من المعلمين وأعضاء هيئة التدريس والعينة ككل وذلك على استجابات الطلاب الفعلية لكل من البنين والبنات والعينة ككل، كما هو موضح بجدول (١٦).

جدول(١٦): النسبة المئوية للطلاب في مستويات الأداء المختلفة في جولتي التحكيم.

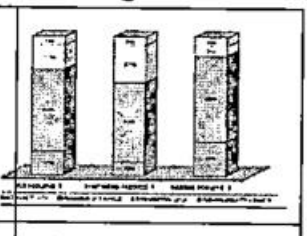
الجولة	المجموعة	العينة	متطر	نام	مرض	متقدم
الجولة الأولى	المعلمون	الكل	17%	60%	17%	7%
	أعضاء هيئة التدريس		17%	53%	24%	7%
	العينة ككل		17%	60%	17%	7%

	المعلمون	البنات	10%	57%	27%	6%
	أعضاء هيئة التدريس		10%	50%	34%	7%
	العينة ككل		10%	57%	27%	6%
	المعلمون	البنين	25%	62%	7%	6%
	أعضاء هيئة التدريس		25%	57%	12%	6%
	العينة ككل		25%	62%	7%	6%
الجولة الثانية	المعلمون	ككل	4%	48%	33%	16%
	أعضاء هيئة التدريس		2%	38%	37%	24%
	العينة ككل		4%	37%	41%	19%
	المعلمون	البنات	2%	38%	40%	20%
	أعضاء هيئة التدريس		1%	27%	39%	33%
	العينة ككل		2%	26%	47%	25%
	المعلمون	البنين	5%	59%	26%	10%
	أعضاء هيئة التدريس		3%	49%	34%	13%
	العينة ككل		5%	47%	35%	12%

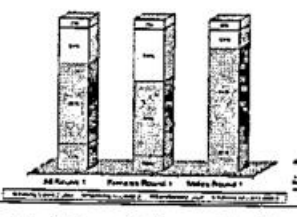
يتضح من جدول (١٦) ما يلي:

- **انخفاض** نسبة الطلاب المصنفين في كل من مستوى الأداء "**متعثر**" ومستوى **الأداء "تام"** في الجولة الثانية عنها بالنسبة للجولة الأولى؛ وذلك نتيجة للانخفاض الذي طرأ على درجة القطع لمستوى الأداء "تام" في الجولة الثانية عنها بالنسبة للجولة الأولى.
- **ارتفاع** نسبة الطلاب المصنفين في كل من مستوى الأداء "**مرض**" ومستوى **الأداء "متقدم"** في الجولة الثانية عنها بالنسبة للجولة الأولى؛ وذلك نتيجة للانخفاض الذي طرأ على درجة القطع لمستوى الأداء "مرض" و"متقدم" في الجولة الثانية عنها بالنسبة للجولة الأولى.

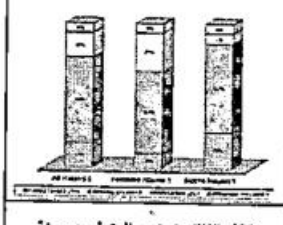
وذلك كما هو موضح بالأشكال من (٢٠: ٢٥) التالية.



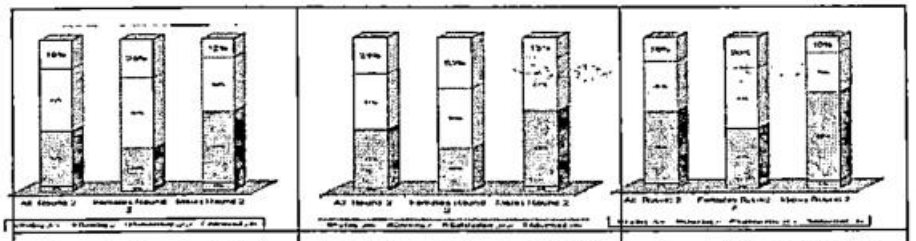
شكل(٢٠): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الأولى وفق عينة المعلمين.



شكل(٢١): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الأولى وفق عينة أعضاء هيئة التدريس.



شكل(٢٢): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الأولى وفق العينة ككل.



شكل(٢٣): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الثانية وفق عينة المعلمين.

شكل(٢٤): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الثانية وفق عينة أعضاء هيئة التدريس.

شكل(٢٥): توزيع المتعلمين عينة الدراسة على مستويات الأداء بالجولة الثانية وفق العينة ككل.

ثانيا: النتائج الخاصة بتأثير بيانات التغذية الراجعة على الممارسات العملية التي يمارسها المحكمون خلال تحديد مستويات الأداء ودرجة القطع:

في هذا الجزء من نتائج الدراسة يتم عرض ومناقشة استجابات المحكمين عينة الدراسة من المعلمين وأعضاء هيئة التدريس على أسئلة الاستبيان بشقيها الكمي والكيفي. والوقوف على العمليات التي يفكر فيها المحكمون خلال كافة إجراءات تحديد درجة القطع وحتى الوصول إلى درجات القطع النهائية. وعلى ذلك يتم تناول استجابة المحكمين على ما يتعلق بعملية تحديد درجة القطع من ممارسات من خلال:

- الوقوف على مدى وضوح المتطلبات القبلية لتحديد درجة القطع.
- العوامل المؤثرة في تقديرات المحكمين في الجولة الأولى.
- الممارسات التي يتبعها المحكمون لاتخاذ قراراتهم حول المفردات في الجولة الأولى.
- العوامل المؤثرة في قرارات المحكمين في الجولة الثانية.
- الممارسات التي يتبعها المحكمون لاتخاذ قراراتهم حول المفردات في الجولة الثانية.

وذلك من خلال الإجابة عن أسئلة الدراسة التالية.

<u>الإجابة عن السؤال السادس وينص على:</u>

ما مدى وضوح المواصفات القياسية للأداء والمفاهيم المرتبطة بها قبل إجراء الجولة الأولى من التقديرات من وجهة نظر المحكمين عينة الدراسة؟

للإجابة عن هذا السؤال تم حساب التكرارات والنسب المئوية لاستجابات المحكمين عينة الدراسة على عبارات هذا المحور كما هو مبين بجدول (١٧).

جدول (17): التكرارات والنسب المئوية لاستجابات عينة الدراسة على العبارات الخاصة بوضوح المواصفات التفاضلية للأداء

العبارة	المكون	موافق بشدة		موافق		غير موافق		غير موافق بشدة		المتوسط	الانحراف المعياري	درجة الموافقة
		ك	%	ك	%	ك	%	ك	%			
1-أهمية وضع معايير للأداء حول هذا الاختبار واضحة.	أعضاء هيئة التدريس	6	100	-	-	-	-	-	-	4	.000	كبيرة جدا
	المعلمين	6	66.7	3	33.3	-	-	-	-	3.67	.50	كبيرة جدا
	الهيئة ككل	12	80	3	20	-	-	-	-	3.80	.414	كبيرة جدا
2-تعريف الدراسات السابقة لمستويات الأداء الأربعة واضح.	أعضاء هيئة التدريس	6	100	-	-	-	-	-	-	4	.000	كبيرة جدا
	المعلمين	6	66.7	3	33.3	-	-	-	-	3.67	.50	كبيرة جدا
	الهيئة ككل	12	80	3	20	-	-	-	-	3.80	.414	كبيرة جدا
3-إجراءات " عم – لا" للحكم على الدرجات واضحة.	أعضاء هيئة التدريس	4	66.7	2	33.3	-	-	-	-	3.67	.516	كبيرة جدا
	المعلمين	2	22.2	6	66.7	-	-	1	11.1	3.25	.463	كبيرة جدا
	الهيئة ككل	6	40	8	53.3	-	-	1	6.7	3.43	.514	كبيرة جدا
4-خطوات وضع معايير الأداء واضحة.	أعضاء هيئة التدريس	3	50	2	33.3	1	16.7	-	-	3.33	.517	كبيرة جدا
	المعلمين	1	11.1	6	66.7	-	-	2	22.2	3.14	.378	كبيرة
	الهيئة ككل	4	26.7	8	53.3	1	6.7	-	-	3.23	.599	كبيرة
5- التمييز بين مستويات الأداء الأربعة واضح.	أعضاء هيئة التدريس	5	83.3	1	13.7	-	-	-	-	3.83	.408	كبيرة جدا
	المعلمين	4	44.4	4	44.4	1	11.1	-	-	3.33	.707	كبيرة جدا
	الهيئة ككل	9	60	5	33.3	1	6.7	-	-	3.55	.640	كبيرة جدا
6- يمكن التمييز بين الطلاب وفقا لمستويات الأداء الأربعة.	أعضاء هيئة التدريس	4	66.7	2	33.3	-	-	-	-	3.67	.516	كبيرة جدا
	المعلمين	3	33.3	5	55.6	1	11.1	-	-	3.22	.667	كبيرة
	الهيئة ككل	7	46.7	7	46.7	1	6.7	-	-	3.40	.632	كبيرة جدا
7- سيتوقع الطالب ذو الحد الأدنى من الكفاية في كل مستوى من مستويات الأداء الأربعة واضح.	أعضاء هيئة التدريس	3	50	3	50	-	-	-	-	3.50	.548	كبيرة جدا
	المعلمين	2	22.2	7	77.8	-	-	-	-	3.22	.441	كبيرة
	الهيئة ككل	5	33.3	10	66.7	-	-	-	-	3.33	.488	كبيرة جدا

يتضح من جدول (١٧) ما يلي:

- بالنسبة لعينة الدراسة من أعضاء هيئة التدريس تحققت كافة عبارات المحور الأول والمرتبطة بوضوح مستويات الأداء بدرجة " كبيرة جدا" وذلك من وجهة نظرهم.
- بالنسبة لعينة الدراسة من المعلمين تراوحت درجة تحقق عبارات المحور الأول والمرتبطة بوضوح مستويات الأداء بين " درجة كبيرة جدا" و "درجة كبيرة" وذلك من وجهة نظرهم.
- بالنسبة لعينة الدراسة ككل تحققت كافة عبارات المحور الأول والمرتبطة بوضوح مستويات الأداء بدرجة " كبيرة جدا"، فيما عدا العبارة رقم(٤) والمتعلقة بوضوح خطوات وضع معايير الأداء، حيث تحققت بدرجة كبيرة. وذلك من وجهة نظرهم.

وبصفة عامة فإن عبارات المحور الأول والمرتبطة بوضوح مستويات الأداء العامة والخاصة تحققت من وجهة نظر المحكمين بدرجة تراوحت بين "كبيرة جدا" و"كبيرة"، وتعد هذه الخطوة أساسية لاستكمال باقي الخطوات. ويشير ذلك إلى وضوح الإجراءات اللازمة لتحديد درجات القطع بالنسبة لجميع أفراد العينة من أعضاء هيئة التدريس والمعلمين وتتفق هذه النتيجة مع نتيجة دراسة (Ferdous &Buckendaht,2013).

الإجابة عن السؤال السابع وينص على:

ما الممارسات التي يتبعها المُحكمون عندما يصدرون أحكامهم على المفردات في الجولة الأولى قبل تقديم بيانات التغذية الراجعة من وجهة نظرهم؟

للإجابة على هذا السؤال تم حساب التكرارات والنسب المئوية والمتوسط والانحراف المعياري ودرجة الموافقة لكل عبارة من عبارات هذا المحور كما هو موضح بجدول (١٨) التالي.

جدول(18): التكرارات والنسب المئوية للاستجابات على العبارات المرتبطة بالعوامل المؤثرة في تقديرات الجولة الأولى.

العبارة		موافق بشدة		موافق		غير موافق		غير موافق بشدة		متردد		المتوسط	الانحراف المعياري	درجة الموافقة
		ك	%	ك	%	ك	%	ك	%	ك	%			
8- الجولة الأولى لتقدير المفردات مشتقة مع فهمي لمستويات الأداء الأربعة.	أعضاء هيئة التدريس	6	100	-	-	-	-	-	-	-	-	4	.000	كبيرة جدا
	المعلمين	4	44.4	5	55.6	-	-	-	-	-	-	3.44	.527	كبيرة جدا
	العينة ككل	10	66.7	5	33.3	-	-	-	-	-	-	3.67	.488	كبيرة جدا
9- مستويات أداء الطلاب في فصلي ومدرستي تؤثر على قراراتي حول تصنيف أسئلة الاختبار.	أعضاء هيئة التدريس	5	83.3	1	16.7	-	-	-	-	-	-	3.83	.408	كبيرة جدا
	المعلمين	5	55.6	3	33.3	1	11.1	-	-	-	-	3.44	.726	كبيرة جدا
	العينة ككل	10	66.7	4	26.7	1	6.7	-	-	-	-	3.60	.632	كبيرة جدا
10- أثق في تقديراتي للمفردات في الجولة الأولى بدرجة كبيرة.	أعضاء هيئة التدريس	1	16.7	5	83.3	-	-	-	-	-	-	3.167	.408	كبيرة
	المعلمين	1	11.1	7	77.8	1	11.1	-	-	-	-	3	.50	كبيرة
	العينة ككل	2	13.3	12	80	1	6.7	-	-	-	-	3.07	.458	كبيرة

يتضح من جدول (١٨) ما يلي:

- إن تقديرات المحكمين في الجولة الأولى جاءت متسقة بدرجة كبيرة مع فهمهم لمستويات الأداء العامة والخاصة وذلك بالنسبة لعينة الدراسة من المعلمين وأعضاء هيئة التدريس والعينة ككل. وتتفق هذه النتيجة مع نتائج دراسة (Ferdous &. Buckendahl, 2013) (Skorupski & Hambleton,2003).

- تأثرت قرارات المحكمين حول مستويات تصنيف الطلاب بدرجة كبيرة من وجهة نظرهم بأداء طلابهم في المدرسة. وتختلف هذه النتيجة مع دراسة (Ferdous &. Buckendahl, 2013) حيث جاء تأثير أداء الطلاب في فصول المحكمين بدرجة أقل.

- تحققت ثقة المحكمين في تقديرات الجولة الأولى متحققة بدرجة كبيرة وذلك من وجهة نظرهم، وتتفق هذه النتيجة مع نتائج دراسة كل من (Ferdous &. Buckendahl, 2013) (Skorupski&Hambleton,2003).

وهذا ما يفسره استجاباتهم على الأسئلة المفتوحة المرتبطة بمدى ثقتهم في حكمهم في الجولة

الأولى، وكذلك تساؤلاتهم عندما يتشككون في تقديرهم كما هو موضح بجدول (19) التالي.

جدول (19): استجابات عينة الدراسة على الأسئلة المرتبطة بـــــــالعوامل المؤثرة في تقديرات الجولة الأولى.

العبارة	استجابات أعضاء هيئة التدريس	استجابات المعلمين
١٣-عندما تكون غير متأكد من قرارك صف كيف تقرر بين اختيارك تقدير "نعم" أو "لا" للمفردة.	- الاعتماد على المواصفات الخاصة والمستوى المعرفي للسؤال. - أقرر في ضوء تقديري للصعوبة المتوقعة للسؤال - مراجعة المواصفات الخاصة للأداء بشكل جيد. - مراجعة مواصفات الأداء العامة والخاصة ومفهوم الطالب ذي الحد الأدنى من الأداء والتفكير في طلاب الصف الأول الإعدادي الذين أعرفهم. - اعتمد على خبرتي التدريسية. - الرجوع إلى توصيف مستويات الأداء الخاصة والعامة. - استخدم التقدير الأدنى.	- مراجعة المواصفات الخاصة للتأكد من قراري. - مراجعة المواصفات العامة والخاصة للأداء. - الرجوع ثانية للمواصفات الخاصة والتحقق من فهمها. - العودة إلى المواصفات الخاصة للأداء. - الخبرة الذاتية من خلال التعامل مع الطلاب بالفصول الدراسية. - العودة إلى المواصفات العامة والخاصة.
١٤-صف أي تساؤلات لديك حول تقديراتك للمفردات في الجولة الأولى.	- إلى أي مدى تتسق تقديراتي في الجولة الأولى مع تقديرات زملائي.	- الشعور بأن بعض تقديرات المفردات بعيدة عن الواقع. - شعرت بالتحيز للمستوى الأعلى من الكفاءة وليست الأدنى.

ويتضح من جدول (١٧) ما يلي:

- جاء الاعتماد على المواصفات الخاصة والعامة قاسما مشتركا بين عينة الدراسة من المعلمين وأعضاء هيئة التدريس عندما يكونون غير متأكدين من تقديراتهم وتتفق هذه النتيجة مع دراسة (Ferdous & Plake,2005) (Skoruski & (Ferdous & Buckendahl, 2013) Hambleton,2003). وترجع أهمية هذه النتيجة إلى أن المحكمين يميلون إلى إعطاء تقديراتهم باستخدام قواعد محكية المرجع، ومن ثم يمكن أن تؤدي إلى تفسير عام لمواصفات مستويات الأداء.

- كما أوضح بعض أعضاء هيئة التدريس أنهم يحكمون أيضا في ضوء المستوى المعرفي للسؤال وخبرتهم التدريسية وصعوبته المتوقعة وأنهم يلجئون أحيانا إلى استخدام التقدير الأدنى.

- وعن التساؤلات التي يفكرون فيها حول تقديرات الجولة الأولى كان أعضاء هيئة التدريس يفكرون في مدى اتساق تقديرات كل منهم مع تقديرات زملائه. أما بالنسبة لعينة المعلمين فقد كانت تساؤلاتهم حول ابتعاد بعض تقديرات المفردات عن الواقع، وكذلك تحيزهم للمستويات الأعلى من الكفاءة دون الأدنى.

<u>الإجابة عن السؤال الثامن وينص على:</u>

ما الممارسات التي يتبعها المحكمون عندما يصدرون أحكامهم على المفردات في الجولة الثانية بعد تقديم بيانات التغذية الراجعة من وجهة نظرهم؟

للإجابة على هذا السؤال تم حساب التكرارات والنسب المئوية لاستجابات عينة الدراسة من المعلمين وأعضاء هيئة التدريس والعينة ككل على العوامل التي قد يكون لها تأثير على تقديراتهم لدرجة القطع في الجولة الثانية كما مبين بجدول (٢٠).

جدول (20): التكرارات والنسب المئوية لاستجابات عينة الدراسة على العبارات الخاصة بالعوامل المؤثرة في تقديرات الجولة الثانية.

العبارة	المحكمين	مؤثر بدرجة كبيرة		مؤثر		غير مؤثر		متروك		المتوسط	الانحراف المعياري	درجة التأثير
		ك	%	ك	%	ك	%	ك	%			
15-مفهوم الطالب ذو الحد الأدنى من الكفاية	أعضاء هيئة التدريس	5	83.3	1	16.7	-	-	-	-	2.83	.408	كبيرة
	المعلمين	4	44.4	5	55.6	-	-	-	-	2.44	.527	كبيرة
	العينة ككل	9	60	6	40	-	-	-	-	2.60	.507	كبيرة
16-المناقشة بين الجولات	أعضاء هيئة التدريس	5	83.3	1	16.7	-	-	-	-	2.83	.408	كبيرة
	المعلمين	8	88.9	1	11.1	-	-	-	-	2.89	.333	كبيرة
	العينة ككل	13	86.7	2	13.3	-	-	-	-	2.87	.352	كبيرة
17-المعارف والمهارات المطلوبة للإجابة على كل سؤال	أعضاء هيئة التدريس	6	100	-	-	-	-	-	-	3	.0000	كبيرة
	المعلمين	4	44.4	3	33.3	2	22.2	-	-	2.22	.833	متوسطة
	العينة ككل	10	66.7	3	20	2	13.3	-	-	2.53	.743	كبيرة
18- درجة القطع للمحكمين الآخرين	أعضاء هيئة التدريس	1	16.7	4	66.7	1	16.7	-	-	2	.632	متوسطة
	المعلمين	1	11.1	7	77.8	1	11.1	-	-	2	0.5	متوسطة
	العينة ككل	2	13.3	11	73.3	2	13.3	-	-	2	.535	متوسطة

تابع جدول (20): التكرارات والنسب المئوية لاستجابات عينة الدراسة على العبارات الخاصة بالعوامل المؤثرة في تقديرات الجولة الثانية.

العبارة	المحكمين	مؤثر بدرجة كبيرة		مؤثر		غير مؤثر		متروك		المتوسط	الانحراف المعياري	درجة التأثير
		ك	%	ك	%	ك	%	ك	%			
19- خبرتك المهنية	أعضاء هيئة التدريس	4	66.7	2	33.3	-	-	-	-	2.67	.516	كبيرة
	المعلمين	6	66.7	3	33.3	-	-	-	-	2.67	.50	كبيرة
	العينة ككل	10	66.7	5	33.3	-	-	-	-	2.67	.488	كبيرة
20- قيمة معامل الصعوبة	أعضاء هيئة التدريس	3	50	3	50	-	-	-	-	2.50	.548	كبيرة
	المعلمين	3	33.3	6	66.7	-	-	-	-	2.33	.50	كبيرة
	العينة ككل	6	40	9	60	-	-	-	-	2.40	.507	كبيرة
21-بيانات الأثر التوزيع التكراري للطلاب على مستويات الأداء	أعضاء هيئة التدريس	3	50	3	50	-	-	-	-	2.50	.548	كبيرة
	المعلمين	3	33.3	6	66.7	-	-	-	-	2.33	.50	كبيرة
	العينة ككل	6	40	9	60	-	-	-	-	2.40	.507	كبيرة
22- تكرارات البدائل لكل سؤال	أعضاء هيئة التدريس	3	50	3	50	-	-	-	-	2.50	.548	كبيرة
	المعلمين	5	55.6	3	33.3	1	11.1	-	-	2.44	.726	كبيرة
	العينة ككل	8	53.3	6	40	1	6.7	-	-	2.47	.640	كبيرة
23- التمثيل البياني للمفردات	أعضاء هيئة التدريس	3	50	3	50	-	-	-	-	2.50	.548	كبيرة
	المعلمين	7	77.8	2	22.2	-	-	-	-	2.87	0.441	كبيرة
	العينة ككل	10	66.7	5	33.3	-	-	-	-	2.67	.488	كبيرة

يتضح من جدول (٢٠) السابق ما يلي:

- جاء تأثير مجموعة العوامل المتعلقة بإصدار الأحكام في الجولة الثانية (مفهوم الطالب ذي الحد الأدنى من الأداء، والمناقشة بين الجولات، والخبرة المهنية، وقيمة معامل الصعوبة،

وتوزيع الطلاب على مستويات الأداء وفق درجات القطع، تكرارات البدائل لكل سؤال، التمثيل البياني للمفردات) مؤثرة بدرجة كبيرة في قرارات المحكمين **من المعلمين وأعضاء هيئة التدريس**؛ فيما عدا العامل المرتبط بدرجة القطع للمحكمين الآخرين فقد جاء مؤثرا بدرجة متوسطة بالنسبة لكلا العينتين. والعامل المرتبط بالمعارف والمهارات المطلوبة للإجابة على كل سؤال حيث جاء تأثيره متوسطا بالنسبة لعينة المعلمين، وكبيرا بالنسبة لأعضاء هيئة التدريس والعينة ككل.

- جاءت "المعارف والمهارات المطلوبة للإجابة على كل سؤال" أكبر العوامل تأثيرا في قرارات أعضاء هيئة التدريس؛ في حين جاءت " المناقشة بين جولات التحكيم" أكبر العوامل تأثيرا في قرارات عينة المعلمين.
- جاءت "درجة القطع بالنسبة للمحكمين الآخرين" أقل العوامل تأثيرا في قرارات المحكمين من أعضاء هيئة التدريس والمعلمين والعينة ككل؛ مما يشير إلى أن المحكمين يميلون للاعتماد على الممارسات محكية المرجع أكثر من المعايير جماعية المرجع عند تقديرهم للمفردات.
- جاء التمثيل البياني للمفردات بالنسبة لكل من أعضاء هيئة التدريس والمعلمين أكثر بيانات الأثر تأثيرا في قرارات المحكمين والعينة ككل. يليها تكرارات البدائل لكل سؤال، ثم توزيع الطلاب على مستويات الأداء وفق درجات القطع، وقيمة معامل الصعوبة في نفس الترتيب.

وتتفق نتائج الدراسة الحالية في أهمية بيانات الأثر بالنسبة للمحكمين في مراجعة تقديراتهم مع دراسة (Egan & Green,2003) (Skoruski & Hambleton,2003) وتختلف مع نتيجة دراسة (Ferdous & Buckendahl, 2013) حيث جاء اعتماد المحكمين فيها على تقديرات التغذية الراجعة أقل.

ويتضمن جدول (٢١) التالي استجابات المحكمين من المعلمين وأعضاء هيئة التدريس للأسئلة المفتوحة المرتبطة بممارسات المحكمين في الجولة الثانية بعد تقديم بيانات الأثر.

جدول (٢١): استجابات عينة الدراسة على الأسئلة المرتبطة بـــــالعوامل المؤثرة في التقديرات في الجولة الثانية.

العبارة	استجابات أعضاء هيئة التدريس	استجابات المعلمين
٢٤-من فضلك وضح كيف أثرت بيانات التغذية الراجعة على تقديراتك "نعم – لا" في الجولة الثانية.	– أثرت بدرجة كبيرة على إعادة تقديرات الجولة الثانية. – أعادة النظر في تقييمي في ضوء الواقع الفعلي لتوزيع درجات الطلاب ومعاملات الصعوبة الفعلية للمفردات ونسب اختيار البدائل. – إعادة تقدير بعض المفردات في ضوء ما أوضحته النسب الفعلية للطلاب في كل مستوى، ومستوى صعوبة كل سؤال، وتكرارات البدائل، والرسم البياني للمفردات. – يسرت بيانات التغذية الراجعة اتخاذ القرار بشأن المفردات في الجولة الثانية. – تقدم التغذية الراجعة معلومات في منتهى الأهمية يصعب التقدير الدقيق بدونها.	– أثرت من خلال النسب المئوية لتكرارات البدائل ونسب الأفراد في المستويات المختلفة والرسم البياني للمفردات حيث تغيرت إلى حد ما تقديراتي لبعض المفردات. – الشعور بأن هناك فجوة بين قراراتي والواقع الفعلي لاستجابات الطلاب لذا تغيرت تقديراتي في الجولة الثانية لبعض المفردات. – أوضحت أن هناك عدة إحصاءات ممكن أن تسهم في تفسير إجابات المتعلمين على الأسئلة. – أثرت بيانات التغذية الراجعة بشكل كبير على تقديراتي للمفردات في الجولة الثانية. – تأثرت تقديراتي بدرجة كبيرة بالمنحنى البياني للمفردات ومعامل الصعوبة. – الفهم الأعمق لمفهوم الطالب ي الحد الأدنى من الأداء. – أثرت بدرجة كبيرة في إعادة تقييم المفردات في الجولة الثانية.
٢٥-من فضلك وضح كيف أثرت التغذية الراجعة على ثقتك في تقديراتك في الجولة الأولى.	– إعادة النظر بشكل كبير في تقديرات الجولة الأولى. – إضافة معلومات مهمة عن المفردات لإعادة التقييم على ضونها. – زادت ثقتي في نسبة كبيرة من تقديراتي، وجعلتني أفكر في مراجعة تقدير بعض المفردات.	– كان الفرق بين تقديراتي في الجولتين الأولى والثانية صغير. – كانت تقديراتي في الجولة الأولى معتمدة فقط على خبرتي المهنية. – إلى حد ما كانت متسقة مع فهمي لمواصفات الأداء إلا أنني غيرت رأي في تقدير بعض المفردات بعد الرجوع إلى بيانات التغذية الراجعة. – إعادة تقدير المفردات في ضوء التغذية الراجعة. – مراجعة تقديراتي مرة ثانية في ضوء فهم أوضح لمفهوم الطالب ذي الحد الأدنى من الأداء.

تابع جدول (٢١): استجابات عينة الدراسة على الأسئلة المرتبطة بـــــالعوامل المؤثرة في التقديرات في الجولة الثانية.

العبارة	استجابات أعضاء هيئة التدريس	استجابات المعلمين
٢٦-من فضلك وضح أي تغييرات أجريتها للطريقة التي نفذت بها تقديرات المفردات "نعم – لا " في الجولة الثانية.	– دراسة التمثيل البياني للمفردات. – في ضوء بيانات التغذية الراجعة قمت برفع تقديرات نعم لبعض المفردات التي تحققت بنسبة مرتفعة أكثر مما توقعت وفي المقابل خفض تقديرات بعض المفردات التي تبين انخفاض نسبة تحقيقها. – قمت بإعادة التقدير في ضوء التغذية الراجعة لاسيما مستوى صعوبة الأسئلة.	– إعادة القرارات ودقة التقديرات. – اعتمدت على نتائج التغذية الراجعة وكذلك على خبرتي بمستوى التحصيل الدراسي للمتعلمين في مادة العلوم وخبرتي في توقع درجة صعوبة مفردات الاختبار في إعادة تقديري لمفردات الاختبار في الجولة الثانية. – اعتمدت بشكل أساسي على معامل الصعوبة في إعادة تقديراتي في الجولة الثانية. – دراسة المنحنى البياني جيدا وكذلك معامل الصعوبة. – اعتبرت أثناء تقديري بيانات التغذية الراجعة متمثلة في الرسم البياني للمفردات ومعامل الصعوبة وتوزيع الطلاب على مستويات الأداء المختلفة كما أوضحتها بيانات الأثر. – أعدت تقديراتي للأداء على المفردات في ضوء بيانات التغذية الراجعة، حيث تغيرت تقديراتي لبعض المفردات.

فيما يتعلق بنتائج الاستجابة حول المفردات المفتوحة الإجابة يتضح من جدول (٢١) ما يلي:

- تنوع المصادر التي اعتمد عليها أعضاء هيئة التدريس عندما يكونون غير متأكدين من قرارتهم (نعم/ لا) وجاءت مراجعة المواصفات الخاصة والعامة نقطة فيصلية بالنسبة لكل من أعضاء هيئة التدريس والمعلمين.
- ركز أعضاء هيئة التدريس على المستوى المعرفي وصعوبة السؤال في حين ركزت عينة المعلمين على الخبرة الذاتية للتدريس بالفصول الدراسية.

بالنسبة لاستجابة عينة الدراسة على المفردة " من فضلك وضح كيف أثرت التغذية الراجعة على ثقتك في تقديراتك في الجولة الأولى"

- جاء تأثير التغذية الراجعة بصفة عامة بعد الجولة الأولى في تفكير المحكمين في مراجعة تقدير بعض المفردات مرة ثانية خلال تقديرات الجولة الثانية. كما زادت من ثقة المحكمين من أعضاء هيئة التدريس في تقديراتهم لنسبة كبيرة من المفردات.
- وبالنسبة لعينة المعلمين كانت موجهة لهم لمراجعة تقديراتهم في ضوء فهم أعمق لمفهوم الطالب ذي الحد الأدنى من الأداء؛ حيث اعتمدت تقديرات عدد منهم على خبرته المهنية في الجولة الأولى

بالنسبة لاستجابة عينة الدراسة على المفردة " من فضلك وضح أي تغييرات أجريتها للطريقة التي نفذت بها تقديرات المفردات "نعم – لا " في الجولة الثانية"

- أوضحت عينة أعضاء هيئة التدريس أهمية بيانات الأثر في تقدير المفردات كما أشارت إلى صعوبة التقدير بدونها. وتتفق هذه النتيجة مع دراسة (Clauser, 2009).
- وبالنسبة لعينة المعلمين جاءت بيانات الأثر باعثا لشعور بالفجوة بين تقديراتهم والواقع الفعلي لاستجابات الطلاب، مع مزيد من الفهم لمفهوم الطالب ذو الحد الأدنى من الأداء.

بالنسبة لاستجابة عينة الدراسة على المفردة " من فضلك وضح أي تغييرات أجريتها للطريقة التي نفذت بها تقديرات المفردات "نعم – لا " في الجولة الثانية".

- اعتمد المحكمون من أعضاء هيئة التدريس والمعلمين على بيانات التغذية الراجعة في تعديل تقديراتهم للجولة الأولى خاصة في ضوء التحليل البياني للمفردات وصعوبة المفردة. واهتمت عينة المحكمين من المعلمين بالنسب المئوية لتوزيع الطلاب على مستويات الأداء وفق درجة القطع المقدرة في الجولة الأولى.

خلاصة:

أسفرت نتائج الدراسة عن الانخفاض في تقديرات درجة القطع في الجولة الثانية عنها في

الجولة الأولى، وذلك بالنسبة لكافة مستويات الأداء وبالنسبة لعينة المحكمين من المعلمين وأعضاء هيئة التدريس، مع عدم وجود فروق دالة إحصائيا بين تقديرات المحكمين ترجع لمتغير خبرة المحكم. وأسفرت نتائج الإجابة عن الأسئلة الكيفية أن هناك ممارسات مشتركة يقوم بها المحكمون على اختلاف مجال خبرتهم عند تقديراتهم لدرجات القطع تمثلت بصفة أساسية في الجولتين الأولى والثانية في مراجعة المواصفات العامة والخاصة لمستويات الأداء، وكذلك مفهوم الطالب ذي الحد الأدنى من الأداء. ورغم عدم وجود فروق كمية في قيمة درجة القطع ترجع لخبرة المحكم، ومع وجود اتفاق بين عينة المعلمين والمحكمين على أهمية عوامل التغذية الراجعة، إلا أن التحليل الكيفي لاستجابات المحكمين أوضح أن أعضاء هيئة التدريس بحكم تخصصهم كانت من بين ممارساتهم مراجعة المستوى المعرفي للسؤال، وكانت الخبرة بالطلاب الذين يعرفهم المحكمون من المعلمين وخبرتهم التدريسية من الممارسات التي يقومون بها. ويحقق التكامل بين تقديرات أعضاء هيئة التدريس والمعلمين الفرصة للحكم الأدق على المفردات. كما أوضح تحليل ممارسات المحكمين أيضا تأثرهم بدرجة كبيرة بعوامل التغذية الراجعة (التحليل البياني للمفردات، صعوبات المفردات-تكرارات البدائل). كانت أكثر العوامل تأثير في قرارات أعضاء هيئة التدريس في الجولة الثانية المعارف والمهارات اللازمة للإجابة على السؤال ومفهوم الطالب ذي الحد الأدنى من الأداء وهي معايير محكية المرجع. وجاءت بيانات التغذية الراجعة مصدرا لثقة أعضاء هيئة التدريس لتقديراتهم في حين كانت مصدرا لمراجعة مفهوم الطالب ذي الحد الأدنى من الأداء بالنسبة لعينة المعلمين. وجاء في الترتيب الأخير لعوامل التغذية الراجعة المؤثرة في قرارات المحكمين درجات القطع للمحكمين الآخرين حيث تشير إلى أن خيارات المحكمين للحكم على المفردات جاءت محكية المرجع بالدرجة الأولى. وتأتي نتائج الدراسة الحالية لتشير إلى ضرورة أن تتضمن عينات المحكمين أعضاء هيئة التدريس والمعلمين. كما جاءت التغذية الراجعة ميسرة لمفهوم الطالب ذي الحد الأدنى من الأداء كمفهوم رئيس تقوم عليه طريقة أنجوف المعدلة. وفي ضوء ما تقدم من نتائج الدراسة فإنه من الممكن الاستفادة من التكامل بين خبرات أعضاء هيئة التدريس الأكاديمية وخبرات المعلمين المهنية العملية في التوصل إلى تقديرات أفضل لدرجات القطع لمستويات الأداء المختلفة للطلاب. ولما كان هناك اتساق في تقديرات درجة القطع بين أعضاء هيئة التدريس والمعلمين في وجود بيانات التغذية الراجعة فإنه من المهم العناية بتوفير بيانات التغذية الراجعة المناسبة التي تساعد في تيسير فهم المحكمين لمستويات الأداء العامة والخاصة، ومن ثم دقة الأحكام المشتقة من عملية تحديد درجة القطع.

التوصيات:

في إطار ما أسفرت عنه الدراسة من نتائج توصي الباحثة بما يلي:

- الاهتمام بتوفير المؤشرات الإحصائية المناسبة لمفردات الاختبار، والاستعانة بها كمصدر للتغذية الراجعة خلال عملية تحديد درجات القطع.
- مراعاة الدمج بين أعضاء هيئة التدريس والمعلمين في تحديد درجات القطع حيث يسمح الدمج بينهما بالاستفادة من تكامل الخبرات في مجالي التدريس والتقويم التربوي.
- ضرورة الاهتمام بتطوير طرق تحديد مستويات الأداء ودرجات القطع باستخدام نظريات القياس الحديثة ومنها نظرية الاستجابة للمفردة.

بحوث مقترحة:

في ضوء ما أسفرت عنه الدراسة من نتائج وتوصيات يقترح إجراء البحوث التالية:

- دراسة تقديرات المحكمين المتطرفة باستخدام نموذج راش.
- دراسة تأثير بيانات الأثر المرتبطة بالتوزيع التكراري لبدائل المفردات على تقدير درجة القطع باستخدام طريقة نيدلسكاي.
- تحديد مستويات الأداء ودرجات القطع لاختبارات الشهادات بجمهورية مصر العربية باستخدام طريقة أنجوف المعدلة.

مراجع الدراسة:

المراجع العربية:

أحمد الشريم ويوسف السوالمة. (٢٠٠٦). تحديد درجة القطع لاختبار محكي المرجع في الرياضيات باستخدام نموذجي "أنجوف" و"تدلسكي": دراسة مقارنة بمعرفة صعوبة الفقرات وعدم معرفتها. *المجلة الأردنية في العلوم التربوية*، مجلد ٢، عدد ١. ص ص ١-١٠.

أمينة محمد كاظم (١٩٩٤). في محمد عماد الدين إسماعيل وناهد رمزي وليلي كرم وأمينة كاظم وهدي ناشف، *معايير نمو طفل ما قبل المدرسة، المجلد الثاني*، الدراسة النفسية. القاهرة: رئاسة مجلس الوزراء، المجلس القومي للطفولة والأمومة.

أمينة محمد كاظم. (٢٠٠٠). *اتجاهات معاصرة في بناء بنوك الأسئلة*. في: الأسس التربوية لإعداد المعلم الجامعي. القاهرة: جامعة عين شمس، ط٣، ٣٢١: ٣٤٢.

خالد التميمي. (١٩٩٩). أثر كل من نوع المحكم وطول الاختبار على تحديد درجة القطع لاختبار

محكي المرجع يقيس الكفايات الرياضية في العمليات الحسابية على الأعداد بالصف السادس الابتدائي بمدينة جدة. *رسالة ماجستير غير منشورة*، كلية التربية، جامعة أم القرى.

راشد الدوسري.(٢٠١٢). تحديد درجة القطع ومستوى الأداء في اللغة العربية لطلبة الحلقة الأولى من التعليم الأساسي. *رسالة الخليج العربي*. العدد ١٢٤.ص ص ١٠٥: ١٥١.

صلاح علام.(١٩٨٥). استخدام النموذج ذي الحدين في تقدير درجة القطع للاختبار محكي المرجع (دراسة إحصائية). *المجلة العربية للعلوم الإنسانية*. جامعة الكويت.

صلاح علام.(١٩٩١).دراسة مقارنة لبعض طرق تحديد مستويات الأداء في الاختبار محكي المرجع، *المجلة المصرية للدراسات النفسية*، العدد ١، ٧٧- ٩٦.

صلاح علام.(٢٠٠٠). *القياس والتقويم التربوي والنفسي*. القاهرة: دار الفكر العربي

صلاح علام. (٢٠٠٥). *نماذج الاستجابة للمفردة الاختبارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي*. القاهرة: دار الفكر العربي.

صلاح علام.(٢٠٠٧).*الاختبارات التشخيصية محكية المرجع في المجالات التربوية والنفسية والتدريبية*. القاهرة: دار الفكر العربي.

عصام الجبة.(١٩٩٨). مدى فاعلية نموذج "أنجوف" في تحديد المستوى لاختبار محكي المرجع. *مجلة كلية التربية جامعة المنصورة*. عدد ٣٦، ص ص ٤١- ٧٣.

فوزية هادي، وصلاح مراد. (٢٠١٣). استخدام نموذج راش في إعداد وتقنين اختبار في اللغة العربية لقبول طلاب الدراسات العليا بكليات العلوم الإنسانية بجامعة الكويت. *المجلة التربوية*، العدد ١٠٦. الجزء الأول. مارس ٢٠١٣. ص ص ١٥: ٤٧.

محمد فراج. (٢٠٠٦). فاعلية بعض النماذج الإحصائية المنفصلة والمتصلة في تحديد الدرجة الفاصلة لاختبار مرجعي المحك. *مجلة مركز الخدمة للاستشارات البحثية*-كلية الآداب – جامعة المنوفية.

المراجع الاجنبية:

Baker, F.(2001). . University of Maryland: College Park, MD. ERIC Clearing house on Assessment and Evaluation.

Batenburg. T. & Laros, J.(2000).
. University of Groningen, The Netherlands, T.A.T.A. Van Batenburg@ PPSW.RVG.NL.

Berk,R.A (1986). Consumers Guide To Setting Performance Standard on Criterion Referenced Test. . 56: 137-172.

Bowers,J. and Shindoll.R.(1989).A Comparison of the Angoff, Beuk, and Hofstee Methods for Setting a Passing Score.
.89-2.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: Areview.
, 35, 167–181.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics.
17, 59–88.

Brennan,R.I.(2002).Estimated Error of a jean when there are only two Observations. Center for Advanced Studies IN Measurement and Assessment Technical Note Number 1. Iowa College of Education .University of Iowa.

Buckendahl.C.W.,Huynh.H.,Siskind.T. and Saunders.J.(2005). Acase Study of vertically Moderated Standard Setting for a state science assessment Program. Applied Measurement in Education. 18(1). 83-98.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations.
, 27, 145–163.

Chang,L.,Dziuban,C.,Hynes,M., and Olson,H.(1996).Does a Standard reflect minimal competency of examinees or judge competency?
,9:161-173.

Cizek, G. J. (1996). Setting passing scores.
, 15, 20–31.

Cizek, G. J. (2012).
. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G., & Bunch, M. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA:

Sage.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). Empirical evidence for the evaluation of performance standards estimated using the Angoff procedure. 22(1), 1–21.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., and Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. , 46, 390–407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. , 13, 65–85.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard setting procedure , 39, 269–290.

Clauser, J., & Clauser, B. (2012, April). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Creswell, J. (2003). Research design: Quantitative, qualitative, and mixed methods approaches. Thousand Oaks, CA: Sage.

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teachers Examination. , 21, 113–129.

Egan, K., & Green, D. R. (2003, April). Influence on judge's decision. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Engelhard , G.J., and Stone,G.E.(1998). Evaluating the quality of ratings obtained from standard setting judges. . 58(2). 179-196.

Ferdous,A.,& Buckendahl, C.,(2013).Evaluating Panelists' Standard Setting Perceptions in a Developing Nation. ,

13: 4-18.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard setting study. , 18(3), 257–267.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An Item Response Theory Analysis of Self-Report Measures of Adult attachment. , 78, 350-365.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis (2nd ed.). New York: Chapman & Hall.

Giraud, G. T. (1999). Making the cut: A qualitative inquiry into the setting of cut scores on school district assessments. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. , 22, 22–32.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. 'Measurement, 2, 61–103.

Halpin , G., sigmon , & halpin , G.(1983). Minimum competency standard set by three divergent groups of raters using three judgmental procedures: implications for validity. Educational & psychological measurement. 43: 185-196.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.) Standard setting: Concepts, methods and perspectives (pp. 159–173). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.). Educational measurement (4th ed.). Westport, CT: American Council on Education & Praeger.

Harasym , P.H. (1981). A comparison of the nedelsky & modified angoff standard setting proceedure on evaluation outcomes. educational & psychological measuement . 41: 725-734.

Hsieh,M.(2013). An Application of Multifaceted Rasch Measurement in the Yes/No Angoff Standard Setting Procedure. , 30(4).491-512.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrainedla tent class analysis using the Gibbs sampler and

posterior predictive checks. , 62, 171–189.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. , 63, 584–601.

Hurtz, G. M., & Jones.J.P.(2009). Innovations in Measuring Rater Accuracy in Standard Setting: Assessing "Fit" to Item Characteristic Curves. . 22(2).120-143.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. , 34, 353–366.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff Standard Setting Method. , 35(1), 69–81.

Jaeger, R. M (1989). Selection of judges for standard setting. . 10(2).6-10.

Jaeger, R. M.(1994). Setting Performance Standards through two-stage judgment. Paper presented at the annual meeting of American Educational Research Association and the National Council on Measurement in Education. New Orleans.

Jaeger, R. M. (1995). On the cognitive construction of standard-setting judgments: The case of configural scoring. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments (Vol. II, pp. 57–73). Washington, DC: U.S. Government Printing Office.

Jalili, M; Hejri,M & Norcini,J.(2011). Comparison of two methods of standard setting: the performance of the three-level Angoff method. , 45: 1199–1208. doi:10.1111/j.1365-2923.2011.04073.x

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. , 25, 285–306.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. , 64, 425–461.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Karantonis,A.,&Sireci,S.G.(2006). The bookmark Standard Setting Method: Aliterrature Review.
,25(1).4-12.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov Chain Monte Carlo in practice: A roundtable discussion.
, 52, 93–100.

Katz,I., and Tannenbaum,J.,(2014). Comparison of Web-based and Face-to-face Standard Setting using the Angoff Method.
. 15(1), 1–17.

Koffler, S. L. (1980) A comparison of Approaches for setting proficiency standards. jem . 17: 167-178.

Lewis.D.M.,Mitzel.H.C.,& Green,D.R.(1996).Standard Setting: A bookmark approach. Paper presented at the Council of Chief States School Officers National Conference on Large Scale Assessment, Boulder.

Linacre, J.M. (2011). A User's Guide to WINSTEPS & MINISTEPS: Rasch-Model Computer Programs. Chicago: MESA.

Lissitz,R.w.,& Wei,H.(2008). Consistency of Standard Setting in an augmented state testing program.
.27(2).46-56.

Margolis,J.,& Clauser,E.(2014). The Impact of Examinee Performance Information on Judges□ Cut Scores in Modified Angoff Standard Exercises. ,Spring 2014, vol.33,No.1,pp.15-22.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises.
, 32(3), 27–35.

Mills, C.N. (1983). A comparison of three methods of establishing cut off scores on criterion referenced tests. jem . 20: 283 - 292.

Nedelsky, L. (1954). Absolute grading standards for objective tests.
. 14: 3-19.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Washington, DC: National Academy Press.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! Educational

Measurement: Issues and Practice, 27(1), 3–9.

Plake, B. S., Buckendahl, C. W., & Ferdous, A. A. (2005). Setting multiple performance standards using the yes/no method: An alternative item mapping method. Paper presented to the Meeting of the National Council on Measurement in Education, Montreal, Canada.

Plake, B. S., Impara, J., & Potenza, M. (1994). Content specificity of expert judges in a standard setting study. Journal of Educational Measurement, 31, 339–347.

Shepard, L.(1984). Setting Performance standards, In Berk(ED.),Aguide to Criterion Referenced Test Construction. Hopkins University Press.

Sireci,S.G.,Hauger,J.B.,Wells,C.S.,Shea,C.,& Zenisky,A.L.(2009). Evaluation of the standard setting on the 2005 grade 12 National Assessment of Educational Progress Mathematics test. Applied Measurement in Education.22(4).339-359.

Skorupski, W. P., & Hambleton, R. K. (2003, April). What are panelists really thinking when they set performance standards? Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. Human Performance, 9, 275–295.

Vukmirovic, Z. (October, 200٦). Standard Setting. Workshop for NCEEE new Staff. Funded by World Bank-PPMU and provided by American Institutes for Research. International Development Division. USA.

Vukmirovic, Z. (October, 2009). Item Response Theory. Conceptual Introduction. Training Material for Psychometric Workshop. Presented on Psychometric Workshop for NCEEE new Staff. Funded by World Bank-PPMU and provided by American Institutes for Research. International Development Division. USA.

The Impact of Feedback Data and Judges' Experience to Estimate Cut Scores and Judges' Practices Using Modified Angoff

by
Ekram Hamza Al-Sayed Sahwan,
PhD NCEE

Summary:

The study aimed to study the effect of the feedback data and the judges' experience in estimating cut scores and the judges' practices according to modified Angoff method. In order to achieve the objective of the study, the researcher built an academic achievement test in the science and calibrated using Rasch Model. A questionnaire also constructed on the practical practices of the judges while estimating cut scores and the factors affecting their decision when using the (yes / no) method of Angoff. The sample consisted of fifteen judges (teachers and faculty).

A workshop was held to estimate the cut scores to classify students into four levels (stumbling-growing-satisfactory-advanced) through two rounds of judging providing the judges with three types of feedback data: Normative data, Impact data and Empirical data.

The most important finding were:

- The median of the final cuts cores estimated by teachers were less than cut scores estimated by faculty. Consistency was found within each group across the rounds.
- There are no statistically significant differences in the estimation of cut scores (growing-satisfactory-advanced) in the first and second rounds due to judges' experience.
- The estimates of the cut scores in the second round were statistically significant less than in the first round for all levels of performance (growing-satisfactory-advanced) on teachers' sample and the overall sample. The estimates of the cut scores in the second round were not statistically significant less than in the first round for the levels of performance (growing-advanced) on faculty's sample.
- Review of the specific and general descriptors of performance levels were the most common practice on which judges relied on their final decisions to judge the items in the first and second rounds.

- A number of factors influenced the decisions of the faculty and teachers in the second round. These factors were the concept of the minimally qualified student, discussion between rounds, professional experience, difficulty coefficients, the frequency distribution of the students' grades, and Graphical of Item Analysis. While the score for other judges was the least influential factor in the decisions of teachers and faculty members in the second round.
- The judges of faculty and teachers relied on the graphical analysis of items, the items difficulty, and the frequency distribution of students' scores on performance levels in the light of the estimated cut scores in the revision of their items estimates.