تحديد درجة القطع لاختبار محكي المرجع في الرياضيات باستخدام نموذجي "أنجوف" و " ندلسكي" : دراسة مقارنة بمعرفة صعوبة الفقرات وعدم معرفتها*

أحمد الشريم** ويوسف سوالمه***

تاريخ تسلم البحث 2005/9/18 تاريخ قبوله 2006/3/9

ملخص: هدفت هذه الدراسة إلى مقارنة نموذجي أنجوف و ندلسكي لتقدير درجة القطع لاختبار محكي المرجع في الرياضيات، و ذلك من خلال وجود مؤشرات عن صعوبة الفقرات أو عدم وجودها. وقد تكون الاختبار من ثلاثين فقرة من نوع الاختيار من متعدد لكل منها أربعة بدائل. وتكونت عينة الدراسة من ثمانين محكماً ومحكمة. وقد تم تقسيم المحكمين بطريقة المزاوجة العشوائية إلى أربع مجموعات متساوية حددت كل منها درجة القطع للاختبار مرتين وفق الاسلوب المعين لها. وأشارت نتائج الدراسة إلى أن درجة القطع للاختبار تتراوح بين 0.62 و 0.68 باستخدام نموذج أنجوف و بين 0.49 و0.57 باستخدام نموذج ندلسكي،وأنه لا يوجد فرق دال إحصائياً عند مستوى الدلالة الإحصائية (*α=0.05*) بين معاملي ثبات النموذجين. كما بينت النتائج أنه لا يوجد فرق دال إحصائياً بين معاملي الارتباط لصعوبة الفقرات وتقديرات المحكمين لها في النموذجين عندما لا يزود المحكمين بصعوبة الفقرات. بينما يوجد فرق دال إحصائياً بين معاملي الارتباط عند معرفة المحكمين لصعوبة الفقرات.وأظهرت النتائج أن تصنيفات الطلبة جميعها كانت دقيقة. (**الكلمات المفتاحية:** درجة القطع، اختبار محكي المرجع، اختبار تحصيلي في الرياضيات، معيار الأداء، نموذج أنجوف، نموذج ندلسكي).

A Comparative Study of Angoff's and Nedelsky's Models to Determine the Cut-off Score for a Criterion-Referenced Test in Mathematics

Ahmad Shreim, *Ministry of Education, Jordan.*

Yousef Swalmeh, *Faculty of Education, Yramouk University, Irbid, Jordan.*

***Abstract**: This study aimed at comparing Angoff's and Nedelsky's models to determine the cut-off score for a criterion-referenced test in mathematics with and without item difficulty indices. The test consisted of 30 multiple-choice items,with four alternatives for each. The sample consisted of 80) male and female raters. The raters were distributed randomly into four equal separate groups. The results of the study indicated that the cut-off score ranged from 0.62 to 0.68 using Angoff's model and from 0.49 to 0.57 using Nedelsky's model. Furthermore, the differences between the reliability coefficients of Angoff's and Nedelsky's models with or without the raters' knowledge of the values of the item difficulty coefficients were not statistically significant at (α = 0.05). (**Keywords**: Cut –off Score , Criterion – Referenced Test, Performance Standard, Angoff's Model , Nedelsky's Model, Mathematics Achievement Test)*

خلفية الدراسة: نالت مسألة تصنيف الطلبة حسب مستوى التمكن من الاهداف التدريسية اهتماماً كبيراً من المهتمين بالقياس والتقويم التربوي والنفسي. و يساعد تصنيف الطلبة المربين على اتخاذ قرارات تعليمية هامة تمس حاضر الطلبة ومستقبلهم، مثل التعرف على الذين يحتاجون إلى مزيد من العناية الفردية أثناء عملية التعليم، أو إتاحة الفرصة للذين حققوا درجة عالية من الكفاية الإنتقال إلى دراسة الوحدات الدراسية التالية، أو إعفاء بعض الطلبة من متطلبات دراسة بعض الوحدات الدراسية إذا تبين أنهم أتقنوا المهارات المرتبطة بها، أو إنتقاء الطلبة وتوجيههم إلى مجال الدراسة الذي يناسبهم، واختيار الطلبة المتقدمين للمعاهد والجامعات والبرامج المختلفة، أو منح الشهادات والتراخيص لممارسة مهن مختلفة أو أنشطة معينة(علام، 1985).

ويتطلب تصنيف الطلبة تحديد الكفايات أو المهارات التي يجب أن يتقنها أو يتمكن منها الطالب، وتحليلها وصياغتها في صورة أهداف تعليمية إجرائية قابلة للقياس، وتقديم هذه الأهداف للطالب قبل بدء العملية التعليمية، وتقدير مدى تحققها باستخدام ما يسمى بالاختبارات محكية المرجع (Criterion Referenced Tests) التي يعتمد على نتائجها في تصنيف الطلبة إلى مجموعتين إحداهما متمكنة (Masters)، والأخرى غير متمكنة (Non-Masters) بالاعتماد على درجة قطع تمثل الحد الأدنى المقبول من الأداء المطلوب (الجبة، 1998).

وتستخدم الاختبارات محكية المرجع في تقدير أداء الفرد بالنسبة إلى محك أو مستوى أداء محدد مسبقاً، دون الحاجة إلى مقارنة أدائه بأداء زملائه، أي أن هذه الاختبارات تقيس مستويات يمكن تفسيرها في ضوء مستويات محددة تتطلب تحديداً دقيقاً للمجال السلوكي الذي يقيسه الاختبار. ويتكون الاختبار من عينة عشوائية من الأسئلة الممثلة للمجال السلوكي المطلوب بحيث نتمكن من معرفة ما يستطيع الفرد أداءه وما لا يستطيع (علام،1985).

---

* البحث مستل من رسالة ماجستير في جامعة اليرموك للباحث الأول بإشراف الباحث الثاني.

** وزارة التربية والتعليم، الأردن.

*** كلية التربية، جامعة اليرموك.

ويمكن تعريف درجة القطع بأنها نقطة على متصل درجات الاختبار، وتستخدم لتقسيم الطلبة إلى مجموعتين (المتمكنين وغير المتمكنين، أو الناجحين و الراسبين، أوالجيد وغير الجيد) بمستويات كفاية مختلفة بالنسبة للأهداف التي يقيسها الاختبار، أي هي الدرجة التي يمكن أن تدل على الحد الأدنى للأداء المقبول لمهارة ما، والتي ينبغي أن يمتلكها الطالب كحدٍ أدنى ليكون ناجحاً او متفوقاً في هذه المهارة(عبدالله،1990 ؛ علام،1991 ؛ Shepard , 1984).

ويشير الأدب التربوي إلى وجود عدة نماذج لتقدير درجة القطع، تعتمد على تقديرات المحكمين للحد الأدنى للنجاح في الاختبارات محكية المرجع، مثل نموذج ندلسكي، وأنجوف، وجاجير، وايبل. و تتطلب هذه النماذج أن يكون لدى المحكمين مفهوم واضح ومشترك حول الحد الأدنى المقبول في الأداء أو الكفاية في السمة موضوع القياس. وتتطلب عملية تحديد النموذج المناسب منها مقارنتها ببعضها بعضاً في ضوء بعض المعايير. ودراسة إمكانية تطويرها لتتناسب مع أهمية القرارات التربوية المتعلقة بقضية التصنيف واستخداماتها الواسعة في مختلف مجالات الحياة (Shepard, 1984).

وتهتم الدراسة الحالية بنموذجي أنجوف وندلسكي. ويتلخص نموذج ندلسكي الذي يناسب فقرات الاختيار من متعدد بعرض فقرات الاختبار على مجموعة المحكمين ويطلب من كل محكم أن يحدد من بين بدائل الفقرة البدائل التي يمكن أن يستبعدها الطالب الذي يمتلك الحد الأدنى المقبول من الكفاية في المجال الذي يقيسه الاختبار، وبذلك يكون الحد الأدنى لاحتمال الإجابة الصحيحة عن الفقرة هو مقلوب عدد البدائل الباقية؛ فمثلاً إذا كانت الفقرة تشتمل على خمسة بدائل للإجابة، ورأى المحكم أن الطالب الذي وصل إلى الحد الأدنى للكفاية المطلوبة يمكن أن يستبعد اختيار ثلاثة منها، عندئذ يكون الحد الأدنى لاحتمال الإجابة الصحيحة هو 0.50، وتكون درجة القطع التي يحددها كل محكم هي مجموع تقديراته للحد الأدنى لاحتمال الإجابة الصحيحة لجميع فقرات الاختبار، وتكون درجة القطع للاختبار بمثابة الوسط الحسابي لدرجات القطع المقدرة من قبل جميع المحكمين(Reilly, Zink, and Israelski,1984).

أما نموذج أنجوف فيتلخص بعرض فقرات الاختبار على مجموعة من المحكمين ويطلب من كل محكم تقدير احتمال إجابة الفرد الذي يمتلك الحد الأدنى المقبول من الكفاية للفقرة بصورة صحيحة، أي على المحكم أن يتصور مجموعة من الطلبة الذين وصلوا الحد الأدنى من الكفاية المطلوبة في المجال الذي يقيسه الاختبار، ثم يقدر نسبة الطلبة منهم الذين يحتمل أن يجيبوا كل فقرة إجابة صحيحة، ويمثل مجموع النسب للفقرات جميعها درجة القطع للاختبار(Shepard, 1984).

وقد أشار كل من بورس وشيندول (Bowers & Shindoll, 1989) إلى أن هناك عدداً من القضايا في تقدير درجة القطع ما زالت بحاجة إلى البحث والدراسة، تتلخص القضية الأولى في أن الخبراء المؤهلين لتقدير درجة القطع لاختبار ما، غالباً ما يفكرون بالطالب الوسط أو فوق الوسط أكثر من تفكيرهم بالطالب الذي يمتلك الحد الأدنى من الكفاية، بل إن أغلبهم يميل إلى تصنيف معظم الأفراد في هذين المستويين، إذ إن مفهوم الحد الأدنى من الكفاية غير واضح تماماً لكل المحكمين، وبالتالي فإن المحكم الذي لا يستطيع أن يصنف فرداً ما في المستوى الأدنى سوف يلجأ إلى تصنيفه في مستوى أعلى (وسط أو فوق الوسط)، وبالتالي فإن درجة القطع للاختبار ستكون مرتفعة. ويقترح الباحثان لحل هذه القضية أن يتم تدريب المحكمين على مفهوم الحد الأدنى من الكفاية قبل عملية التقدير وخلالها، مما يوفر صدقاً مقبولاً لتقديرات درجة القطع. أما القضية الثانية فتتعلق بالمؤشرات الإحصائية عن الفقرات والمفحوصين، هل من الضروري إطلاع المحكمين عليها أم لا ؟ إن أحد الافتراضات الهامة في تحديد درجة القطع هو أن هذه الدرجة يتم تقديرها على أساس أن الاختبار محكي المرجع، وهي 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 عن الفقرات أو معلومات عن المفحوصين، وأخذ المحكمين لهذه المعلومات بعين الاعتبار سيؤثر في تقديراتهم، وبالتالي فإن درجة القطع ستتأثر بتلك المعلومات التي جمعت من المجموعة، وهي في الواقع مجموعة مرجعية، أي أن عاملاً أو أكثر من العوامل معيارية المرجع قد أثر في تقدير درجة القطع.

ولمعرفة ما إذا كان تزويد المحكمين بمعلومات عن الفقرات والمفحوصين سوف يؤثر في تقديراتهم؛ لا بد من إجراء دراسة تجريبية على عينة مماثلة ليقوم المحكم بتقدير درجة قطع لاختبار ما مرتين، الأولى بغياب أية معلومات عن الفقرات، والثانية بوجود مؤشرات إحصائية عنها.

وفي الواقع لا يوجد اتفاق على أن درجة القطع الأعلى هي الأفضل أو العكس، فقد تكون في موقف ما الأعلى هي الأنسب، وقد تكون في موقف آخر الأقل هي الأنسب، فدرجة قطع 0.85 مناسبة لاختبار قيادة السيارة، ولكنها تعد منخفضة جداً لقيادة طائرة، والطالب الذي يحصل على 0.70 في تخصص أكاديمي يمكن اعتباره متمكناً، بينما مستوى 0.70 في الطب يعد غير متمكن، وكذلك قد نقبل أن يتقن الطالب 0.50 مما تعلمه في مادة اللغة الأجنبية في الصف الخامس الأساسي لاعتباره ناجحاً، ولكن هذه العلامة تعد منخفضة جداً إذا حصل عليها الطالب نفسه في اللغة الرسمية (اللغة الأم)، وذلك لأن لغته الرسمية هي اللغة التي سيتعلم بواسطتها بقية العلوم والمواد الدراسية الأخرى، فإذا كان الطالب غير متمكن منها بدرجة جيدة، فإنه سينعكس سلباً على بقية المواد الأخرى، أي أن هناك اعتبارات عديدة لمستوى التمكن المطلوب تتعلق بالنتائج المختلفة المترتبة على نوعية وكفاية الفئة التي سيتم اعتبارها متمكنة.

ويتضح من خلال مراجعة بعض الدراسات أن درجة القطع للاختبار محكي المرجع تعد من أهم العوامل التي تؤثر في حساب

معامل الثبات له (عبد الله،1990)، وأن مستوى التمكن يجب أن لا يزيد على 0.80 بالنسبة لاختبارات المعلمين (عبد السلام، 1992)، و أنه لا توجد فروق ذات دلالة إحصائية في تقديرات مستويات الأداء للاختبار بين مجموعات المحكمين المتباينة في مستوى التمكن باستخدام أي من نموذجي أنجوف وندلسكي (علام،1991). كما أشارت دراسة أخرى (Cross, Impara, Frary, & Jaeger,1984) إلى أن نموذج ندلسكي يعطي تقديرات أقل لمستوى النجاح من نموذج أنجوف، كما توجد ارتباطات عالية بين تقديرات المحكمين لصعوبة الفقرات ومعاملات الصعوبة الفعلية لها. وتبين دراسة بورز وشيندول (Bowers & Shindoll, 1989) أن تقديرات المحكمين باستخدام نموذج أنجوف في حال معرفتهم لمؤشرات إحصائية عن الفقرات ترتبط ارتباطاً موجباً مع صعوبة الفقرات وتكون أقل منها في حالة عدم معرفتهم لها.

وأشارت دراسة فيرهوفن وزملائه (Verhoeven et al., 1999) إلى أن نموذج أنجوف يعد ملائماً لتقدير درجة القطع للاختبار كونه يتمتع بمؤشرات ثبات وصدق مقبولة. وقد بَينت معظم الدراسات أن درجة القطع للاختبار باستخدام نموذج ندلسكي أقل منها في نموذج أنجوف (Chang,1999; Croos et al., 1984).

وتبين دراسة انجلهارد واندرسون (Engelhard & Anderson, 1998) أن ثبات الاستقرار لتقديرات المحكمين لدرجة القطع للاختبار أعلى في حال معرفتهم لمعاملات الصعوبة للفقرات، وأن كلاً من كفاية المحكمين وتجانس صياغة الفقرات يؤثر على درجة القطع الكلية للاختبار. وقد خلصت دراسة شانج و زملائه(Chang et al., 1996) إلى أن المحكمين يميلون لأن يضعوا تقديراً أعلى للفقرات التي يجيبونها إجابة صحيحة من الفقرات التي يجيبونها إجابة خاطئة.

وتتميز الدراسة الحالية عن غيرها من الدراسات أنها تقارن بين نموذجي أنجوف وندلسكي في حال تزويد المحكمين بمؤشرات عن صعوبة الفقرات وعدم تزويدهم بها، و تسعى لمعرفة الفرق بين معاملات الثبات ومعاملات الصدق للنموذجين في تلك الحالات، كما أنها تبحث في قدرة النموذجين على التنبؤ بالتحصيل المدرسي للطلبة، وكذلك تصنيف الطلبة إلى متمكنين وغير متمكنين؛ فهذه المواضيع لم تغطيها الدراسات السابقة بشكل واضح ومفصل، وهو ما حاولت هذه الدراسة القيام به.

مشكلة الدراسة:

تحظى عملية التقويم والاختبارات بشكل خاص باهتمام الكثير من العلماء والباحثيين، إذ قاموا بتطوير العديد من النماذج الخاصة بتحديد مستوى الأداء لأغراض مختلفة. وعلى الرغم من ذلك، إلا أننا لا نجد أي استخدام يذكر لمثل هذه النماذج في تحديد مستوى الأفراد، ولم تحظ بالاهتمام الكافي في الاردن، إذ ما زالت تستخدم العلامة 50 للنجاح في المواد المدرسية المختلفة، والمعدل 65 في الثانوية العامة كحد أدنى للقبول في الجامعات الحكومية دون وجود أدلة علمية أو عملية تدعم ذلك. و كذلك فإن الكتب والمراجع العربية لم تتناول نماذج تحديد مستوى الأداء بالقدر الكافي مما يسبب عائقاً أمام استخدام المعلمين والتربويين لها. وفيما يتعلق بالنماذج نفسها، فإنه لا يوجد اتفاق على كفاية نموذج محدد أو أكثر لتحديد درجة القطع للاختبار، وبالتالي فإنها بحاجة إلى الدراسة والمقارنة فيما بينها للتعرف على فاعليتها في تحديد درجات القطع من حيث الصدق والثبات وملاءمتها للاختبار والمفحوصين وغير ذلك من القضايا المتعلقة بتحديد مستويات الأداء.

وعلى الرغم من الاهتمام الكبير الذي تحظى به الأساليب والنماذج الخاصة بتحديد درجة القطع في الاختبارات محكية المرجع في العالم الغربي، إلا أنها لم تلق الاهتمام نفسه في الأردن. وتأتي هذه الدراسة على أنها محاولة من الباحثين لتطبيق نموذجي أنجوف و ندلسكي في تقدير درجة القطع لاختبار محكي المرجع في الرياضيات، و إجراء مقارنة بين االنموذجين وفق معايير محددة إحصائياً و تجريبياً.

أسئلة الدراسة:

تسعى هذه الدراسة للإجابة عن الأسئلة التالية:

1. هل الفرق دال إحصائياً عند مستوى الدلالة الإحصائية (α= 0.05) بين معامل الثبات لنموذج أنجوف ومعامل الثبات لنموذج ندلسكي،في حال تزويد/عدم تزويد المحكمين بمؤشرات الصعوبة ؟
2. هل يختلف معامل الثبات لنموذجي أنجوف وندلسكي في حالة عدم وجود مؤشرات عن صعوبة الفقرات عن معامل الثبات لهما في حالة وجود مؤشرات عن صعوبة الفقرات ؟
3. هل يختلف معامل الارتباط بين معاملات الصعوبة للفقرات وتقديرات المحكمين لها في نموذج أنجوف عن معامل ارتباطها مع تقديرات المحكمين لها في نموذج ندلسكي عند مستوى الدلالة الإحصائية (α= 0.05)، في حال تزويد أو عدم تزويد المحكمين بمؤشرات الصعوبة ؟
4. ما درجة الفاعلية لنموذجي أنجوف وندلسكي في تصنيف الطلبة والتنبؤ بمستوى التحصيل الدراسي؟

الطريقة والإجراءات

عينة الدراسة: تكونت عينة الدراسة من 80 محكم ومحكمة و 120 طالباً.وتألفت عينة المحكمين من أربعة مشرفين تربويين لمبحث الرياضيات، وستة وسبعين معلماً ومعلمة رياضيات ممن يدرّسون الصف العاشر الأساسي، ولا تقل الخبرة العملية لكل منهم في تدريس الصف العاشر عن سنتين، وقد تم اختيارهم بطريقة عشوائية من مدارس الذكور والاناث التي تحتوي على الصف العاشر في محافظة جرش. وقد اختيرت عينة الطلبة بالطريقة العشوائية العنقودية من مدارس الذكور في مدينة جرش للعام الدراسي 2003/2002م، إذ كانت الشعبة هي وحدة الاختيار، وقد تم اختيار ثلاث مدارس بطريقة عشوائية من بين المدارس التي تحوي

الصف العاشر، واختيرت شعبة واحدة عشوائياً من كل مدرسة. وقد بلغ العدد الإجمالي للطلبة 120 طالباً.

أداة الدراسة: تكونت أداة الدراسة من اختبار محكي في موضوع معادلة الخط المستقيم للصف العاشر الأساسي، وقد تضمن الاختبار ثلاثين فقرة من نوع الاختيار من متعدد، ولكل فقرة أربعة بدائل واحد منها يمثل الإجابة الصحيحة، وقد روعيت الشروط الواجب توافرها في مثل هذا النوع من الفقرات، وقد حرص الباحثان على أن تقيس كل فقرة هدفاً محدداً، وفقاً لقائمة الأهداف الخاصة بالوحدة الدراسية، إذ تم تحليلها تحليلاً دقيقاً ومفصّلاً، وقد أعدت قائمة مكونة من سبعة وعشرين هدفاً تفصيلياً شاملة لموضوع معادلة الخط المستقيم.

وللتأكد من شمول الأهداف للوحدة الدراسية وتمثيلها للمستويات المعرفية الثلاثة (معرفة – فهم – تطبيق)، فقد تم عرضها على مجموعة من المتخصصين في مجال أساليب تدريس الرياضيات، وأصحاب الخبرة من المعلمين، لإبداء الرأي حول مدى شمول الأهداف لموضوع الوحدة الدراسية وطريقة صياغتها، وقد تمت دراسة ملاحظاتهم بالتفصيل وأجريت التعديلات المناسبة بناءً عليها.

بعد تحديد جميع الأهداف التفصيلية للوحدة، تم كتابة مجموعة من الفقرات لكل هدف بشكل منفصل، وقد تراوح عدد الفقرات المحتملة لكل هدف بين اربع إلى ست فقرات وقد اختيرت فقرة واحدة بشكل عشوائي لكل هدف من الأهداف التفصيلية باستثناء الهدف الأول والهدف الرابع والهدف العشرين، فقد تم اختيار فقرتين لكل منها. وقد أعدت استبانة للحكم على الاختبار تكونت من الفقرات التالية :

1. مدى ارتباط الفقرة بالهدف الذي تقيسه من حيث المحتوى والمستوى المعرفي.
2. الفقرة مصاغة بلغة واضحة وسليمة ومفهومة.
3. لغة الفقرة تناسب مستوى طلبة الصف العاشر الأساسي.
4. الفقرة تخلو من أية إشارات لفظية للإجابة الصحيحة.
5. إجابة الفقرة لا تؤثر في إجابة غيرها من فقرات الاختبار.
6. متن الفقرة يبرز مشكلة واضحة ومحددة.
7. مموهات الفقرة جذابة للطلبة ومناسبة.

وقد وضع أمام كل فقرة مقياس تقدير متدرج من 1- 5، ثم عرضت هذه الاستبانة مع فقرات الاختبار على عشرة معلمين من أصحاب الخبرة، واثنين من المشرفين التربويين لمبحث الرياضيات. وقد تراوحت قيم الأوساط الحسابية لتقديرات المحكمين لصياغة فقرات الاختبار بين 4.29 و 5.0 ، وهي قيم مرتفعة مما يعني أن فقرات الاختبار مصاغة بشكل سليم ومقبول. وعلاوة على ذلك طلب منهم تدوين ملاحظاتهم على الفقرات، وبالاعتماد على هذه الملاحظات أُجريت التعديلات المناسبة.

صدق الأداة:

أولاً: للتحقق من أن فقرات الاختبار تشكل عينة ممثلة لمجتمع الفقرات التي تقيس المهارات الخاصة بالوحدة موضوع الاختبار، فقد تم حساب الأوساط الحسابية لتقديرات المحكمين الخاصة بمدى ارتباط كل فقرة بالهدف الذي تقيسه بشكل منفرد، وقد بلغت قيمته 5 بانحراف معياري يساوي صفر، مما يدل على اتفاق المحكمين فيما يتعلق بمطابقة كل فقرة للهدف الذي تقيسه.

ثانياً: للتحقق من قدرة الاختبار على التمييز بين المجموعات المتمايزة في السمة(موضوع الاختبار) المقاسة، فقد تم تطبيق الاختبار على عينتين، تتكون الاولى من 30 طالباً قبل أن يتعلموا الوحدة الدراسية الخاصة بالاختبار،وتتكون الثانية من 120 طالبا بعد ان تعلموا الوحدة.ولاختبار دلالة الفرق بين متوسطي المجموعتين تم استخدام اختبار (t) للبيانات المستقلة. وقد بلغت قيمة t 12.03 وهي دالة عند مستوى دلالة احصائية يقل عن (α= 0.001)،أي أن الاختبار يميز بين المجموعات المتمايزة في المعارف والمهارات الخاصة بمعادلة الخط المستقيم.

ثبات الاختبار:

أولاً: ثبات الاتساق الداخلي للاختبار: لتقدير ثبات الاتساق الداخلي للاختبار، تم تطبيق الاختبار على عينة استطلاعية مكونة من 31 طالباً من طلبة الصف العاشر بعد أن تعلموا الوحدة الدراسية (موضوع الاختبار)، وقد تم حساب معامل الثبات وفق معادلة كرونباخ (ألفا) وبلغ 0.84. وتجدر الإشارة إلى أن طريقة ألفا تناسب الاختبارات معيارية المرجع، وعند استخدامها مع اختبار محكي المرجع فإنها تعطي تقديراً منخفضاً للثبات، الأمر الذي يدعو إلى الثقة بثبات الاختبار لأغراض الدراسة الحالية. وقد تراوحت قيم معاملات الصعوبة لفقرات الاختبار بين 0.23 و0.81 ، وتراوحت قيم معاملات التمييز بين 0.12 و 0.67 ، وقد تم إعادة صياغة ست فقرات وذلك لانخفاض قيم معاملات التمييز لها عن 0.20. وعندما طبق الاختبار بصورته النهائية على عينة الدراسة المكونة من 120 طالباً ارتفعت قيمة الثبات للاختبار حيث بلغت قيمة ألفا 0.88، وتراوحت قيم معاملات التمييز بين 0.30 و 0.50، وقيم معاملات الصعوبة بين 0.41 و0.90.

ثانياً: للتحقق من اتساق تصنيف الأفراد المفحوصين إلى متمكنين أو غير متمكنين، فقد تم إعادة تطبيق الاختبار على العينة المكونة من 120 طالباً بعد مرور أسبوعين على التطبيق الأول، ولتقدير الثبات تم حساب معدل الصواب(Hit rate) عند درجة قطع محددة تم افتراضها0.70، ويقصد بمعدل الصواب نسبة الاتفاق في تصنيف الطلبة في مرتي التطبيق. كما تم استخدام معامل كابا (Coefficient Kappa)، كونه الأسلوب الإحصائي الذي يأخذ أخطاء التصنيف بعين الاعتبار، إذ كلما زاد اتساق قرار التصنيف في المرتين دل

ذلك على ارتفاع درجة الثبات للاختبار (علام،2000). وقد بلغت قيمة معامل كابا 0.65، وهي نسبة مقبولة عند درجة القطع 0.70 ، علماً أن قيم معامل كابا تكون محصورة بين -1، 1. وتتأثر قيمة معامل كابا بقيمة درجة القطع. كما بلغ معدل الصواب 0.83، وهو معدل مقبول،ويعني ذلك أن نسبة الاتفاق في تصنيف الطلبة إلى متمكنين وغير متمكنين في مرتي التطبيق تساوي 83%.

ولمعرفة درجة حساسية الفقرات لعملية التعليم، أو لفاعلية التعليم؛ فقد تم اختبار الفرق بين معاملات الصعوبة للفقرات لمجموعة الطلبة قبل التعليم (30 طالباً)، مع معاملات الصعوبة للفقرات للمجموعة التي تلقت التعليم (120 طالباً)، باستخدام مربع كاي. وتبين أن هناك فروقاً ذات دلالة إحصائية عند مستوى الدلالة الإحصائية (α= 0.01) بين معاملات الصعوبة للفقرات جميعها قبل التعليم وبعده،إذ بدت الفقرات أسهل بعد التعليم،مما يدل على حساسية الفقرات لعملية التعليم وتأثرها بها، وهذا مؤشر على صدق الاختبار محكي المرجع (علام، 2000).

إجراءات جمع البيانات:

أولاً: تطبيق الاختبار على الطلبة في المدارس: تم تطبيق الاختبار على أفراد العينة المكونة من 120 طالباً من طلبة الصف العاشر الأساسي في مدارسهم، وبشكل جماعي في الغرف الصفية، وبالاستعانة بمعلمي مبحث الرياضيات في تلك المدارس. وقد تم إبلاغ الطلبة بموعد الاختبار قبل أسبوع من تطبيقه. وكانت الإجابة على ورقة الامتحان نفسها. وبعد مرور أسبوعين تم إعادة تطبيق الاختبار نفسه على العينة نفسها، وبعد ذلك تم تصحيح الاختبار بإعطاء درجة واحدة لكل فقرة من فقرات الاختبار في حالة إجابتها إجابة صحيحة، وصفر في حالة إجابتها إجابة خاطئة. وقد تم الحصول على إحصاءات الفقرة من معاملات صعوبة وتمييز.

ثانياً: جمع البيانات من المحكمين: تم تقسيم المحكمين إلى أربع مجموعات متساوية في العدد بطريقة المزاوجة العشوائية لضمان تكافؤ المجموعات في الخبرة و الوظيفة، إذ تكونت كل مجموعة من مشرف واحد و 19 معلما. وتم تدريب أفراد كل مجموعة على إجراءات النموذج الذي يخصهم. وقام أفراد المجموعة الأولى بتقدير درجة القطع للاختبار باستخدام إجراءات نموذج أنجوف دون معرفتهم لمعاملات الصعوبة للفقرات، وتم ذلك بالطلب من كل منهم تخيل أن لديه مئة طالب من طلبة الصف العاشر ممن يمتلكون الحد الأدنى من المعرفة والقدرة التحصيلية التي تمكنهم من النجاح في موضوع الاختبار، ثم تقدير نسبة الطلبة منهم التي ستكون قادرة على إجابة كل فقرة إجابة صحيحة، أما أفراد المجموعة الثانية فقاموا بتقدير درجات القطع لفقرات الاختبار باستخدام اجراءات نموذج أنجوف، ولكن بعد تزويدهم بمعاملات الصعوبة للفقرات. وقام أفراد المجموعة الثالثة بتقدير درجة القطع للاختبار باستخدام إجراءات نموذج ندلسكي دون معرفتهم لمعاملات الصعوبة للفقرات، وتم ذلك من خلال إجابة كل محكم عن السؤال التالي: من بين البدائل الأربعة الخاصة بكل فقرة، ما هي باعتقادك البدائل التي سيتجنب اختيارها(يستبعدها)الطالب الذي يمتلك حد أدنى مقبول من المعرفة والقدرة التحصيلية تمكنه من النجاح في موضوع الاختبار ؟ أما أفراد المجموعة الرابعة، فقاموا بتقدير درجة القطع للاختبار باستخدام اجراءات نموذج ندلسكي بمعرفة معاملات الصعوبة للفقرات. وبعد مرور أسبوعين على الجولة الأولى تم إعادة الإجراءات نفسها في جولة ثانية لأفراد العينة جميعهم.

النتائج

أولاً: تراوحت الأوساط الحسابية لتقديرات المحكمين لدرجات القطع للاختبار في الجولة الأولى، باستخدام نموذج أنجوف دون معرفة المحكمين لمعاملات الصعوبة للفقرات، بين 0.50 و 0.80، بمتوسط حسابي مقداره 0.67، و تعد هذه القيمة درجة قطع للاختبار، وإذا ضربت هذه النسبة بعدد فقرات الاختبار 30 فقرة سنجدها تساوي 20.1. أي أنه يمكن اعتبار الطالب متمكناً إذا أجاب عن عشرين فقرة إجابة صحيحة كحد أدنى من المجموع الكلي لفقرات الاختبار. وقد بلغت درجة القطع في الجولة الثانية 0.68. وقد لوحظ أن أعلى قيمة للفرق المطلق بين تقديرات المحكمين في الجولتين الأولى والثانية تساوي 0.07 وهي قيمة منخفضة، وقد تم حساب معامل ارتباط بيرسون بين التقديرات في جولتي التحكيم وبلغ 0.94، وهو دال إحصائي عند مستوى الدلالة الإحصائية (α= 0.05)، مما يدل على ارتفاع معامل ثبات الاستقرار للنموذج.

ثانيا: تراوحت الأوساط الحسابية لتقديرات المحكمين لدرجات القطع للاختبار في الجولة الأولى، باستخدام نموذج أنجوف بمعرفة المحكمين لمعاملات الصعوبة للفقرات بين 0.59 و 0.71، بمتوسط حسابي مقداره 0.63، أي على الطالب في الصف العاشر أن يجيب على تسع عشرة فقرة من فقرات الاختبار حتى يمكن اعتباره متمكناً من الوحدة الدراسية. وقد بلغت قيمة متوسط تقديرات المحكمين في الجولة الثانية 0.62 ، وهي مساوية تقريباً لوسط تقديراتهم في الجولة الأولى. وقد لوحظ أن أعلى قيمة للفرق المطلق بين تقديرات المحكمين في مرتي التطبيق تساوي 0.03، وهي قيمة منخفضة.هذا وقد تم حساب معامل ارتباط بيرسون بين تقديرات المحكمين في التطبيقين وبلغ 0.89،وهو دال إحصائياً عند مستوى الدلالة الإحصائية (α= 0.01) وهذه مؤشرات على ثبات الاستقرار للنموذج.

ثالثا: تراوحت الأوساط الحسابية لتقديرات المحكمين لدرجات القطع للاختبار في الجولة الأولى، باستخدام نموذج ندلسكي دون معرفة معاملات الصعوبة للفقرات بين 0.41 و 0.72، بمتوسط حسابي مقداره0.50، أي على الطالب في الصف العاشر أن يجيب على خمس عشرة فقرة من فقرات الاختبار حتى يمكن اعتباره متمكناً من الوحدة الدراسية. وقد بلغت

قيمة متوسط تقديرات المحكمين في الجولة الثانية 0.49. وقد لوحظ أن أعلى قيمة للفرق المطلق بين تقديرات المحكمين في جولتي التحكيم تساوي 0.11. وقد بلغ معامل ارتباط بيرسون بين تقديرات المحكمين في الجولتين 0.85، وهو دال إحصائياً عند مستوى الدلالة الإحصائية ($\alpha = 0.01$)، وهذا مؤشر على ثبات الاستقرار للنموذج.

رابعا: تراوحت الأوساط الحسابية لتقديرات المحكمين لدرجات القطع للاختبار في الجولة الأولى، باستخدام نموذج ندلسكي بمعرفة معاملات الصعوبة للفقرات بين 0.52 و 0.73 بمتوسط حسابي مقداره 0.57، أي على الطالب في الصف العاشر أن يجيب عن سبع عشرة فقرة من فقرات الاختبار حتى يمكن اعتباره متمكناً من الوحدة الدراسية. وبلغ متوسط تقديرات المحكمين في الجولة الثانية 0.57. وقد لوحظ أن أعلى قيمة للفرق المطلق بين تقديرات المحكمين في جولتي التحكيم تساوي 0.07. و بلغ معامل ارتباط بيرسون بين تقديرات المحكمين في الجولتين 0.80، وهو دال إحصائياً عند مستوى الدلالة الإحصائية ($\alpha = 0.01$)، وهذا مؤشر على ثبات الاستقرار للنموذج.

خامسا: يبين الجدول (1) أوساط تقديرات المحكمين لصعوبة كل فقرة من فقرات الاختبار باستخدام نموذجي أنجوف و ندلسكي في حالة معرفة معاملات الصعوبة، وفي حالة عدم معرفتها، بالإضافة إلى قيم معاملات الصعوبة المحسوبة فعلياً للفقرات من استجابات الطلبة.

وقد تم حساب معامل ارتباط بيرسون بين تقديرات المحكمين في النماذج الأربعة ومعاملات الصعوبة للفقرات. وقد بلغ معامل الارتباط بين تقديرات المحكمين ومعاملات الصعوبة للفقرات 0.98 لنموذج أنجوف في حالة معرفة المحكمين لصعوبة الفقرات، و 0.57 لنموذج أنجوف في حالة عدم معرفة المحكمين بمعاملات الصعوبة،و 0.77 لنموذج ندلسكي في حالة معرفة المحكمين لصعوبة الفقرات و0.52 لنموذج ندلسكي في حالة عدم معرفة المحكمين لصعوبة الفقرات. وتجدر الإشارة إلى أن هذه المعاملات جميعها دالة إحصائياً ($\alpha = 0.05$).

ولإجابة سؤال الدراسة الأول فقد استخدم اختبار Z. وقد بلغت قيمة Z المحسوبة 1.42، في حالة عدم معرفة معاملات الصعوبة و 0.95 في حالة معرفة معاملات الصعوبة،ولكون كل منها أقل من القيمة الحرجة 1.96 ، فإنه لا يوجد فرق دال إحصائياً عند مستوى الدلالة الإحصائية ($\alpha = 0.05$) بين معامل الثبات لنموذج أنجوف و معامل الثبات لنموذج ندلسكي سواء بوجود أو عدم وجود صعوبة الفقرات.

ولإجابة سؤال الدراسة االثاني فقد تم استخدام اختبار Z، و تبين أن قيمة Z المحسوبة تساوي 0.92 لنموذج أنجوف و 0.46 لنموذج ندلسكي، وكل منها أقل من القيمة الحرجة 1.96، أي لا يوجد فرق دال إحصائياً عند مستوى الدلالة الإحصائية($\alpha = 0.05$) بين معامل ثبات النموذج(أنجوف،ندلسكي) في حالة معرفة قيم معاملات الصعوبة للفقرات، و معامل ثباته في حالة عدم معرفة معاملات الصعوبة.

جدول (1): أوساط تقديرات المحكمين لصعوبة الفقرات حسب النماذج المختلفة ومعاملات الصعوبة للفقرات.

الفقرة	أنجوف1*	أنجوف2*	ندلسكي1	ندلسكي2	صعوبة الفقرة	الفقرة	أنجوف1	أنجوف2	ندلسكي1	ندلسكي2	صعوبة الفقرة
1	0.81	0.92	0.78	1.00	0.90	16	0.71	0.56	0.53	0.53	0.48
2	0.75	0.77	0.55	0.55	0.75	17	0.73	0.48	0.52	0.50	0.41
3	0.81	0.85	0.66	0.85	0.83	18	0.69	0.47	0.50	0.48	0.43
4	0.68	0.80	0.53	0.63	0.79	19	0.65	0.48	0.58	0.45	0.43
5	0.64	0.60	0.52	0.52	0.58	20	0.63	0.52	0.53	0.66	0.49
6	0.59	0.63	0.47	0.52	0.63	21	0.67	0.56	0.38	0.54	0.53
7	0.66	0.67	0.53	0.57	0.68	22	0.75	0.84	0.57	0.82	0.82
8	0.62	0.64	0.43	0.47	0.63	23	0.72	0.69	0.49	0.54	0.67
9	0.55	0.45	0.43	0.42	0.40	24	0.62	0.68	0.41	0.53	0.65
10	0.59	0.63	0.45	0.47	0.58	25	0.82	0.80	0.51	0.67	0.78
11	0.72	0.74	0.45	0.63	0.74	26	0.80	0.69	0.62	0.75	0.59
12	0.76	0.66	0.42	0.51	0.62	27	0.74	0.50	0.49	0.46	0.42
13	0.64	0.63	0.54	0.55	0.63	28	0.78	0.67	0.54	0.57	0.64
14	0.63	0.73	0.58	0.55	0.70	29	0.53	0.43	0.41	0.39	0.42
15	0.62	0.55	0.51	0.48	0.49	30	0.46	0.37	0.38	0.41	0.37

* أنجوف 1 بدون معرفة المحكمين لصعوبة الفقرات، أنجوف 2 بمعرفة المحكمين لصعوبة الفقرات، ندلسكي 1 بدون معرفة المحكمين لصعوبة الفقرات، ندلسكي 2 بمعرفة المحكمين لصعوبة الفقرات

ولإجابة سؤال الدراسة الثالث، فقد تم استخدام اختبارZ ، وتبين أن قيمة Z المحسوبة للفرق بين معاملي الارتباط بين النموذجين قد بلغت 0.21 في حالة عدم معرفة المحكمين لمعاملات الصعوبة و 3.76 في حالة معرفة المحكمين لمعاملات الصعوبة، أي لا يوجد فرق دال إحصائياً ($\alpha = 0.05$) بين معاملي الارتباط لتقديرات المحكمين ومعاملات صعوبة الفقرات للنموذجين في حالة عدم معرفة المحكمين لمعاملات الصعوبة،بينما يوجد فرق دال إحصائياً بين معاملي الارتباط لتقديرات المحكمين ومعاملات صعوبة الفقرات للنموذجين.إذ يتبين أن تقديرات المحكمين في نموذج أنجوف أكثر ارتباطاً بالصعوبة عندما يتم توفيرها للمحكمين.

ولإجابة سؤال الدراسة الرابع، فقد قدر ثبات التصنيف كمؤشر لفاعلية التصنيف باستخدام درجات القطع التي تم الحصول عليها من الدراسة للنموذجين، إذ إن لكل نموذج تقديرين لدرجة القطع في كل جولة من جولات التحكيم(تقدير دون معرفة الصعوبة،

وتقدير بمعرفة الصعوبة)، أي لدينا ثماني درجات قطع للاختبار تم تقديرها من المحكمين في المجموعات الأربع ؛ تقسم كل منها المفحوصين إلى مجموعتين (متمكنين، وغير متمكنين). ونظراً لكون درجة القطع في الجولة الثانية مساوية أو قريبة جداً من درجة القطع في الجولة الأولى، فقد اعتمدت درجات القطع في الجولة الأولى لأغراض حساب ثبات التصنيف للطلبة في مرتي التطبيق. و يبين الجدول (2) تصنيفات الطلبة المختلفة في ضوء درجة القطع في مرتي تطبيق الاختبار عليهم.

جدول (2): درجات القطع للاختبار وتصنيفات الطلبة المختلفة في مرتي التطبيق، ومعدلات الصواب لها.

النموذج	درجة القطع	عدد المتمكنين في التطبيقين	متمكن في 1 وغير متمكن في 2	متمكن في 2 وغير متمكن في 1	غير متمكن في التطبيقين	معدل الصواب	معامل كابا
أنجوف بدون الصعوبة	0.67	53	8	9	50	0.86	0.72
أنجوف بمعرفة الصعوبة	0.63	63	4	10	43	0.88	0.77
ندلسكي بدون الصعوبة	0.50	80	3	7	30	0.92	0.81
ندلسكي بمعرفة الصعوبة	0.57	72	1	11	36	0.90	0.78

يتبين من الجدول (2) أن معدلات الصواب لتصنيف الطلبة جميعها معدلات مرتفعة، وأن أعلى معدل صواب كان لنموذج ندلسكي دون معرفة الصعوبة 0.92، ويليه معدل الصواب لنموذج ندلسكي بمعرفة الصعوبة 0.90، ثم معدل الصواب لنموذج أنجوف بمعرفة الصعوبة 0.88، وكان أقل معدل صواب لنموذج أنجوف بدون معرفة الصعوبة 0.86 .

وقد تم حساب قيم معامل كابا، إذ تبين أن أعلى قيمة لمعامل كابا كانت لنموذج ندلسكي دون معرفة الصعوبة 0.81، ويليه معامل كابا لنموذج ندلسكي بمعرفة الصعوبة 0.78، ثم معامل كابا لنموذج أنجوف بمعرفة الصعوبة 0.77، وكانت أقل قيمة لمعامل كابا لنموذج أنجوف دون معرفة الصعوبة 0.72 . ويلاحظ أن جميع القيم مرتفعة، وتدل على انخفاض قيم أخطاء التصنيف.

ولحساب قدرة النماذج على التنبؤ بمستويات التحصيل المدرسي في الرياضيات، فقد تم الحصول على العلامات النهائية للطلبة من جداول العلامات النهائية للمدرسة، وحساب معامل كابا لكل درجة قطع من الدرجات الأربع على الاختبار باعتماد محكات مختلفة لجودة التحصيل المدرسي(علامة التمكن) في الرياضيات تتمثل في المستويات التالية : 50، 60، 70. ويشير محك الجودة إلى الحد الأدنى للإتقان. والجدول (3) يبين قيم معامل كابا للنماذج الأربعة مع العلامات المدرسية للطلبة.

جدول (3): قيم كابا لتصنيف الطلبة حسب درجات القطع الناتجة من تقديرات المحكمين والعلامات المدرسية.

النموذج	درجة القطع للاختبار	علامة التمكن في التحصيل المدرسي		
		50	60	70
أنجوف دون معرفة الصعوبة	0.67	0.47	0.83	0.86
أنجوف بمعرفة الصعوبة	0.63	0.61	0.96	0.66
ندلسكي دون معرفة الصعوبة	0.50	0.87	0.77	0.53
ندلسكي بمعرفة الصعوبة	0.57	0.77	0.96	0.67

ويتبين من الجدول (3) أن أعلى قيمة لمعامل كابا عند استخدام العلامة 50 كمحك للجودة في العلامات المدرسية تساوي 0.87 لنموذج ندلسكي دون معرفة الصعوبة للفقرات، وأن أعلى قيمة لمعامل كابا عند استخدام العلامة 60 كمحك للجودة في العلامات المدرسية تساوي 0.96 لكل من نموذج ندلسكي بمعرفة الصعوبة للفقرات ونموذج أنجوف بمعرفة الصعوبة، وأن أعلى قيمة لمعامل كابا عند استخدام العلامة 70 كمحك للجودة في العلامات المدرسية، كانت لنموذج أنجوف بدون معرفة الصعوبة للفقرات وقد بلغت 0.86.

0.67 لنموذج أنجوف دون معرفة صعوبة الفقرات، و0.63 لنموذج أنجوف بمعرفة صعوبة الفقرات، و 0.50 لنموذج ندلسكي دون معرفة صعوبة الفقرات، و 0.57 لنموذج ندلسكي بمعرفة صعوبة الفقرات.

وقد نتج عن نموذج ندلسكي في الحالتين درجة قطع أقل للاختبار من نموذج أنجوف. وقد أكدت هذه النتيجة عدة دراسات سابقة (علام،1991؛Croos, et al.,1984; Chang, 1999،إذ أشارت إلى أن طريقة انجوف تعطي غالباً درجات قطع أعلى من طريقة ندلسكي. وقد يعزى انخفاض درجة القطع في نموذج ندلسكي مقارنة بنموذج انجوف إلى الأسباب التالية (Chang, 1999): السبب الأول أن طريقة ندلسكي تضع قيوداً على المحكمين فيما يتعلق بعدد القيم الاحتمالية التي تنتج عن تقديراتهم، إذ إنها تؤدي إلى قيم منفصلة والفروق بينها غير متساوية، إذ إن احتمال إجابة الفقرة يساوي 0.33 عندما يقدر المحكم أن الطالب قادر على استبعاد بديل واحد من أصل أربعة بدائل، و 0.50 إذا

مناقشة النتائج

يتضح من النتائج السابقة أن الطرق الأربعة التي استخدمها المحكمون في تقدير درجة القطع لاختبار الرياضيات في موضوع معادلة الخط المستقيم من وحدة الهندسة التحليلية للصف العاشر الأساسي، أدت إلى درجات قطع متباينة نسبياً فقد كانت اوساطها كما يلي :

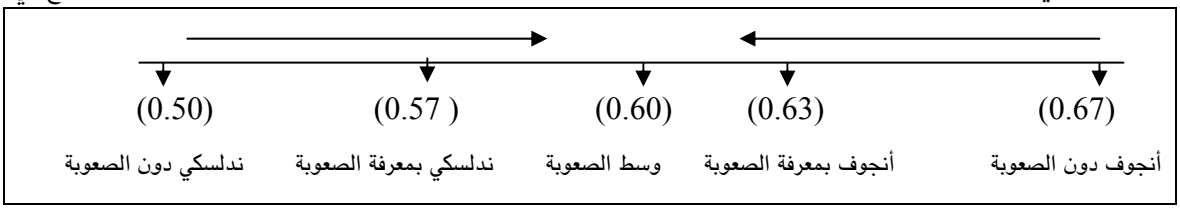
استبعد بديلين، و1.00 إذا استبعد ثلاثة بدائل، و0.25 إذا لم يستطع استبعاد أي بديل، وبالنظر إلى هذه القيم نجد أن ثلاثاً منها لا تتجاوز القيمة 0.50، أي أن النموذج لا يسمح بإعطاء قيم مرتفعة. ولهذا السبب جاءت درجة القطع للاختبار منخفضة.

والسبب الثاني هو صعوبة تحديد البدائل التي يرى المحكمون أن الطلبة الذين يمتلكون الحد الأدنى من الكفاية قادرون على استبعادها كونها غير صحيحة، إذ إن هذا القرار أصعب من القرار الذي يتخذه المحكم في طريقة أنجوف، التي يعطي المحكم فيها حكماً مباشراً يتعلق بمستوى صعوبة الفقرة بالنسبة للطالب الذي يمتلك الحد الأدنى من الكفاية. وتزداد صعوبة تحديد البدائل في طريقة ندلسكي بازدياد تجانسها، فإذا تشابهت البدائل لفقرة ما في طريقة ندلسكي فإن عملية التقدير تزداد صعوبة، والمحكم في هذه الحالة يعطي الفقرة تقديراً أقل من التقدير الحقيقي للفقرة، لأن التشابه الكبير بين البدائل يجعله يعتقد أن الطالب غير قادر على استبعادها وبالتالي سينعكس على قيمة صعوبة الفقرة، في حين أن مثل هذه القضية لا تواجه المحكم في طريقة أنجوف، فهو يركز انتباهه بشكل أكبر على متن الفقرة والبديل الصحيح أكثر من البدائل الأخرى.

والسبب الثالث هو أن بعض المحكمين غالباً ما يفكرون بالطالب في مستوى الوسط أو فوق الوسط، وقليلاً ما يفكرون بالطالب الذي يمتلك الحد الأدنى من الكفاية، ولهذا فإن تقديرات المحكمين في طريقة أنجوف غالباً ما تكون أكبر من 0.50، وهذا ما يجعلها تنتج درجات قطع مرتفعة.

وبالإضافة إلى ذلك تشير نتائج الدراسة إلى أن قيم معاملات ثبات الاستقرار للنماذج الأربعة جميعها كانت مرتفعة، مما يدل على فاعلية النماذج ومناسبتها لتقدير درجة القطع لاختبار الرياضيات في وحدة الهندسة التحليلية، وقد أكد على فاعلية استخدام نموذج أنجوف لتحديد درجة القطع الكثير من الدراسات (علام، 1991؛ Plake Hertz & Hertz, & Kane, 1991; Shepard, 1984; Chang, 1999 ; 1999).

وفيما يتعلق بأثر معرفة المحكمين لصعوبة الفقرات على تقديراتهم لدرجة القطع، فيلاحظ أن تقديرات المحكمين لدرجات القطع باستخدام طريقة أنجوف قد انخفضت بعد معرفتهم بصعوبة الفقرات. إذ بلغت درجة القطع للاختبار قبل معرفتهم للصعوبة 0.67، ولكن عند إطلاعهم على صعوبةالفقرات انخفضت إلى 0.63، مع العلم أن وسط صعوبة الفقرات للاختبار بلغ 0.60. ونلاحظ كذلك أن تقديرات المحكمين لدرجات القطع باستخدام طريقة ندلسكي قد ارتفعت بعد معرفتهم لقيم الصعوبة، فقد بلغ متوسط تقديراتهم قبل معرفة الصعوبة 0.50، ثم ارتفعت إلى0.57 بعد معرفتهم لها. ومن ذلك يتبين أن معرفة المحكمين بمؤشرات الصعوبة تؤثر بشكل كبير على تقديراتهم في النموذجين، أي أنها تجذب تقديرات المحكمين نحوها كما هو موضح في الشكل 1.



الشكل 1: تأثير معرفة الصعوبة على تقديرات درجة القطع

ونلاحظ كذلك أن قيمة معامل الارتباط لتقديرات المحكمين باستخدام نموذج أنجوف بمعرفة معاملات الصعوبة 0.98 أعلى منها في حال عدم معرفة المحكمين لمعاملات الصعوبة 0.57، أي أن معرفة المحكمين لقيم معاملات الصعوبة قد أثر وبشكل واضح على تقديراتهم مما جعلها أكثر ارتباطاً بقيم صعوبة الفقرات، وهذا يؤثر على صدق تقديرات المحكمين لمستويات النجاح بكل تأكيد، إذ كلما ارتفعت قيمة معامل ارتباط تقديرات المحكمين بصعوبة الفقرات ارتفع صدق التقديرات لمستويات النجاح، وهذا ما أشارت إليه دراسة بورز و شندول (Bowers & Shindoll, 1989).

وكذلك بالنسبة لمعامل الارتباط لتقديرات المحكمين باستخدام نموذج ندلسكي بمعرفة معاملات الصعوبة 0.77 فإنه أعلى منه في حال عدم معرفة المحكمين لمعاملات الصعوبة 0.52، أي أن معرفة المحكمين لقيم معاملات الصعوبة قد أثر وبشكل واضح على تقديراتهم مما جعلها أكثر ارتباطاً بقيم صعوبة الفقرات مما يؤكد ضرورة تزويد المحكمين بمؤشرات إحصائية عن الفقرات عند استخدام أي من النموذجين في تحديد درجة القطع للاختبار.

وقد أشارت نتائج الدراسة إلى عدم وجود فرق بين معامل الارتباط لمعاملات الصعوبة وتقديرات المحكمين في نموذج أنجوف، ومعامل الارتباط لمعاملات الصعوبة وتقديرات المحكمين في نموذج ندلسكي، في حال عدم وجود مؤشرات عن صعوبة الفقرات، إذ بلغ معامل الارتباط في نموذج أنجوف 0.57، وبلغ معامل الارتباط في نموذج ندلسك 0.52، وهما قيمتان منخفضتان والفرق بينهما قليل وغير دال إحصائياً.

بينما يوجد فرق دال إحصائياً بين معامل الارتباط لمعاملات الصعوبة وتقديرات المحكمين في نموذج أنجوف، ومعامل الارتباط لمعاملات الصعوبة وتقديرات المحكمين في نموذج ندلسكي، في حال وجود مؤشرات عن صعوبة الفقرات، إذ بلغ معامل الارتباط في نموذج أنجوف 0.98، ومعامل الارتباط في نموذج ندلسكي 0.77، ومع أن القيمتين مرتفعتان إلاّ أن الفرق بينهما كبير وهو دال إحصائياً. وهذا يؤكد أنه في حال استخدام أي من النموذجين لتقدير درجة القطع لاختبار ما فإنه من الممكن توفير مؤشرات عن صعوبة الفقرات وعندها يكون نموذج أنجوف الأكثر تأثراً بمعاملات

الصعوبة، وقد أشارت إلى ذلك دراسة بورز و شندول (Bowers & Shindoll, 1989).

وفيما يتعلق بمعدلات الصواب لتصنيفات المفحوصين في مرتي التطبيق إلى متمكنين وغير متمكنين؛ فإن أعلى معدل صواب كان باستخدام نموذج ندلسكي بدون معرفة قيم الصعوبة 0.92، والسبب في ذلك يعود لانخفاض درجة القطع الناتجة عن استخدام هذا النموذج 0.50، إذ من المعروف أن معدل الصواب في عملية تصنيف الطلبة إلى متمكنين وغير متمكنين يتأثر بشكل مباشر بدرجة القطع للاختبار، فكلما قلت درجة القطع زادت دقة التصنيف. ولذلك نجد أن أقل قيمة لمعدل الصواب كان لنموذج أنجوف دون معرفة الصعوبة 0.86، وهو صاحب أعلى درجة قطع للاختبار، ومع ذلك فهي تعد قيمة مرتفعة ومؤشراً جيداً على دقة التصنيف، ونستطيع القول أن استخدام إجراءات أي من النموذجين لأغراض التصنيف يعد مناسباً.

وفيما يتعلق بصدق التنبؤ للنماذج؛ فقد بينت النتائج وعن طريق استخدام درجات القطع الناتجة من النماذج المختلفة للتنبؤ بالتحصيل الدراسي للطلبة في الرياضيات، أن أعلى قيمة لمعامل كابا عند استخدام المستوى 50 كمحك للجودة للعلامات المدرسية كان لنموذج ندلسكي دون معرفة الصعوبة، وقد بلغ 0.87، وهو مؤشر على أن هذا النموذج مناسب جداً للتنبؤ بالتحصيل المدرسي للطلبة في مادة الرياضيات. ويعود سبب هذا الارتفاع لقيمة معامل كابا لهذا النموذج عند المحك 50 إلى أن درجة القطع للنموذج كانت أيضاً50، أي مساوية للمحك. بينما كانت أقل قيمة لمعامل كابا لنموذج أنجوف بدون معرفة الصعوبة للفقرات، وقد بلغت0.47، وهي قيمة منخفضة وتعد غير مقبولة، أي أن النموذج لا يصلح لأغراض التنبؤ بالتحصيل المدرسي للطلبة في الرياضيات، وسبب الانخفاض هذا يعود لارتفاع درجة القطع للنموذج، التي بلغت 0.67 وهي أعلى بكثير من المحك.

بينما نلاحظ تغيراً كبيراً في قيم معامل كابا عند تغير قيمة محك الجودة، إذ كانت أعلى قيم له في النموذجين عند استخدام محك الجودة 60 إذ تقترب قيمته من درجات القطع لكل من نموذجي أنجوف وندلسكي بمعرفة الصعوبة والتي بلغت لهما 0.63، 0.57 على التوالي، إذ يقع محك الجودة في منتصف المسافة بين النموذجين.وعند استخدام محك الجودة 0.70 فإن أفضل قيمة لمعامل كابا كانت لنموذج أنجوف دون معرفة الصعوبة كونه يعطي درجة قطع قريبة من محك الجودة، التي بلغت 0.67، بينما أقل قيمة لمعامل كابا كانت لنموذج ندلسكي دون معرفة الصعوبة كونه يعطي أقل درجة قطع 0.50 .

ومما تقدم نستنتج أن معدلات الصواب وقدرة النماذج على التنبؤ بمستويات التحصيل تتأثر بشكل كبير بدرجات القطع، فكلما ارتفعت درجة القطع للاختبار انخفضت معدلات الصواب، أي يزيد احتمال الخطأ في التوقع (علام،1985). وبالتالي فإن عملية اختيار النموذج المناسب للتنبؤ بالتحصيل في المستقبل يجب أن يعتمد على درجة القطع التي ينتجها النموذج للاختبار، فكلما اقتربت درجة القطع الناتجة من النموذج من درجة القطع المعتمدة لغايات النجاح والرسوب أدى ذلك إلى ارتفاع معدلات الصواب وارتفاع دقة التوقع. وبالتالي فلا يجوز الاعتماد على نموذج محدد لأغراض التنبؤ بالتحصيل دون تحديد المحك الذي سيعتمد للتصنيف، إذ إن قيمة هذا المحك هي التي تحدد أي النماذج أكثر ملاءمة وصدقاً في التنبؤ.

الاستنتاجات والتوصيات :

- إن درجة القطع الناتجة عن استخدام إجراءات نموذج انجوف تعد أعلى من درجة القطع الناتجة عن استخدام إجراءات نموذج ندلسكي بوجود وعدم وجود مؤشرات إحصائية عن فقرات الاختبار.
- تتأثر تقديرات المحكمين لتقدير درجة القطع للاختبار في كل من نموذج انجوف ونموذج ندلسكي بمعرفتهم بمعاملات الصعوبة للفقرات.
- معاملات ثبات الاستقرار للنموذجين في الحالات الأربع مرتفعة ومقبولة.
- معدلات التوافق في تصنيف الطلبة إلى متمكنين وغير متمكنين في مجال الاختبار باستخدام درجات القطع الناتجة من النموذجين مرتفعة.
- يعد نموذجا أنجوف وندلسكي بمعرفة معاملات الصعوبة الأكثر ملاءمة للتنبؤ بالتحصيل المدرسي للطلبة في الرياضيات عند استخدام محك الجودة 60 كونهما يعطيان أعلى معدل صواب.
- توفير معلومات عن صعوبة فقرات الاختبار يجعل الأسلوب المستخدم أقل استقراراً.
- من حيث الاستخدام وبساطة الإجراءات، يعد نموذج أنجوف الأسهل.

المصادر والمراجع

الجبة، عصام الدسوقي إسماعيل.(1998).مدى فاعلية نموذج "أنجوف " في تحديد المستوى لاختبار محكي المرجع.*مجلة كلية التربية/جامعة المنصورة/مصر*،36: 41 – 73.

عبد السلام، نادية محمد. (1992). مشكلات معاصرة عند بناء الاختبارات محكية المرجع "تحليل وتقويم ". *مجلة علم النفس*،21، 30 - 39.

عبد الله،محمود محمد إبراهيم.(1990). *دراسة سيكومترية مقارنة لطرق حساب معامل ثبات الاختبارات المرجعة إلى المحك*. رسالة ماجستير غير منشورة،الجامعة الأردنية، عمان.

علام، صلاح الدين محمود. (1985). استخدام النموذج ذي الحدين في تقدير درجة القطع لاختبار محكي المرجع (دراسة إحصائية وتجريبية). *المجلة العربية للعلوم الإنسانية*، 5(9): 27-43.

علام، صلاح الدين محمود. (1991). دراسة مقارنة لبعض طرق تحديد مستويات الأداء في اختبار مرجعي المحك. *المجلة المصرية للدراسات النفسية*، 1: 77 – 96.

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

Bowers, J. and Shindoll R.(1989).A Comparison of the Angoff, Beuk, and Hofstee Methods for setting a Passing Score. *ACT Research Report Series* 89-2.

Chang, L.(1999). Judgmental item analysis of the Nedelsky and Angoff standard – setting methods. *Applied Measurement in Education*, 12: 151-156.

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

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 Teacher Examinations, *Journal of Educational Measurement*, 21: 113-129.

Engelhard, G. & Anderson W. (1998). A Binomial trials model for examining the ratings of standard – setting judges. *Applied Measurement in Education* , 11: 209- 230.

Hurtz, M. & Hertz, R. (1999). How many raters should be used for establishing cut off scores with the Angoff method? Ageneralizability Theory Study. *Educational and Psychological Measurement*, 59:113-129.

Plake, Barbara & Kane, M. (1991). Comparison of Methods For Combining the Minimum Passing Levels for Individual Items Into a Passing Score for a Test. *Journal of Educational Measurement*, 28:249-256.

Rrilly, R., Zink, D., & Israelski, W.(1984).Comparison of direct and indirect methods for setting minimum passing scores. *Applied Psychological easurement,8*, 421-429.

Shepard, Lorrie. (1984). Setting performance standards, In Berk (Ed.),*A guide to criterion – referenced test construction*. Hopkins UniversityPress. Verhoeven B., Van der Steeg A., Scherpbier A., Muijtjens A., Verwijnen G.,&Van der

Vleuten C. (1999).Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education* ,33: 32-837.