المجلة الأردنية في العلوم التربوية، مجلد 4، عدد 1، 2008، 37-48

استخدام نموذج المحاولات ذات الحدين في فحص تقديرات المحكمين لدرجة القطع لاختبار مرجعي المحك في الرياضيات

زايد بني عطا* و يوسف سوالمه*

تاريخ تسلم البحث 2007/1/15 تاريخ قبوله 2008/2/6

ملخص: هدفت الدراسة إلى استخدام نموذج المحاولات ذات الحدين في فحص جودة تقديرات المحكّمين وفق طريقة أنجوف لدرجة القطع لاختبار مرجعي المحك في الرياضيات. ولتحقيق ذلك حللت علامات 110 طالباً وطالبة من طلبة الصف العاشر الأساسي على اختبار محكي في الرياضيات يتكون من 30 فقرة من نوع الاختيار من متعدد ، كما حللت تقديرات 15 محكماً لدرجات القطع في ثلاث جولات للتحكيم. وكشفت نتائج الدراسة عن وجود فروق ذات دلالة إحصائية (α=0.05) في درجات القطع المقدرة تعزى إلى جولة التحكيم. وبعد استبعاد تقديرات المحكمين الذين لم تتطابق تقديراتهم مع نموذج المحاولات ذات الحدين، قلّت الفروق بين درجات القطع المقدرّة في الجولات جميعها، وارتفعت قيمة معامل دقة القرار لتصنيف الطلبة. (**الكلمات المفتاحية:** طريقة أنجوف، درجة القطع، اختبار مرجعي المحك، نموذج المحاولات ذات الحدين، مستوى الأداء).

Using Binomial Trials Model for Examining the Judges' Ratings of Cut Scores for a Criterion-Referenced Mathematics Test

Zaid Bani Ata & Yousef Swalmeh, *Faculty of Education, Yarmouk University Irbid, Jordan.*

Abstract*: The study aimed at investigating the quality of judges' rating of the cut score for a criterion-referenced test according to Angoff Method using Binomial Trials model. To achieve this purpose, marks of 110 tenth grade students on a 30- item criterion- referenced test in mathematics were analyzed. 15 judges' ratings were also analyzed in three rounds of judgments. The results of the study indicated that there were significant differences (α = 0.05) among cut score estimates which can be attributed to judgment round. After eliminating the judges' ratings which did not fit the model expectations, the differences among the estimated cut scores decreased in all rounds while there was an increase in the hit rate for students' classification. (**Keywords**: Angoff Method, Cut Score, Criterion- Referenced Test, Binomial Trials model).*

مقدمه: أثمرت جهود علماء القياس و ظهور أنظمة المساءلة في النظام التربوي خلال النصف الثاني من القرن العشرين، في بلورة القياس مرجعي المحك Criterion Referenced الذي يفسر أداء الفرد من خلال مقارنة علامته مع معايير أداء Performance Standards محددة مسبقاً، وفي الاعتماد على مستويات الأداء في عملية اتخاذ القرارات التربوية المناسبة المتعلقة بالفرد. وأدى ذلك إلى ظهور ما يعرف بالاختبارات مرجعية المحك التي تستخدم لتفسير أداء الفرد بالنسبة لمجال سلوكي محدد بدقة (علام، 2001). وأطلق ثورندايك (Thorndike, 1997) عليها اسم اختبارات الكفاية الدنيا Minimum Competency Tests، ولها مسميات أخرى منها الاختبارات مرجعية الهدف Objective Referenced Tests والاختبارات مرجعية المجال Domain Referenced Tests واختبارات التمكن أو الإتقان Mastery Tests ((Popham, 2000.

ومع تزايد أهمية تحديد مستويات الأداء، زاد الاهتمام بطرق تحديد الحد الأدنى المقبول من الأداء المطلوب لعمل ما. و يطلق على ذلك الحد تسميات مختلفة مثل: درجة القطع Cutting Score ودرجة التمكنMastery Score ودرجة الاجتيازPassing Score ومستوى الكفاية الدنيا Minimum Competency Level ودرجة المحك Criterion Score (علام، 2001؛ Cizek, 1993). وتشير درجة القطع إلى المعيار، أو المحك، أو النقطة، أو العلامة التي تفصل بين الطلاب الذين حققوا مستوى الأداء المطلوب، والذين لم يحققوه (Brandon, 2002).

وتجدر الإشارة إلى أن هناك الكثير من الطرق لتحديد درجة القطع. وتعد طريقة أنجوف Angoff من أشهر تلك الطرق لبساطة إجراءاتها وسهولة استخدام المحكمين لها. وهي من الطرق التي تهتم بالحكم على فقرات الاختبار من خلال عرضها على مجموعة من المحكمين المختصين؛ إذ يطلب منهم تصور مجموعة من المفحوصين الذين وصلوا إلى الحد الأدنى من الكفاية لمقبولة في المجال الذي يقيسه الاختبار، ثم تقدير نسبة المفحوصين منهم الذين يحتمل أن يجيبوا عن كل فقرة من فقرات الاختبار إجابة

* كلية التربية، جامعة اليرموك، اربد، الأردن.

صحيحة، ويمثل متوسط هذه النسب درجة القطع في الاختبار (Cizek, 1993).

فقد أشارت نتائج دراسة كروس وامبارا وفراري وجيجر (Croos, Impara,Frary & Jager, 1984) إلى أن طريقة أنجوف تتفوق على طريقتي ندلسكي وجيجر وتنتج مستوى أداء يمكن الدفاع عنه، وقد برر الباحثون ذلك بأن طريقة أنجوف تنتج معاملات ارتباط عالية بين تقديرات المحكمين لمستوى الأداء للفقرات ومعاملات الصعوبة الفعلية لهذه الفقرات. وقد أكدت دراسة شين وهيرتز (Chinn & Hertz, 2002) أن تقديرات مستويات الأداء باستخدام طريقة أنجوف اكثر استقراراً منها باستخدام طريقة جيجر.

وقد أشار بيرك (Berk, 1996) من خلال مراجعته للعديد من الدراسات التي تناولت طرق تحديد مستويات الأداء إلى أنه يوجد ثلاثة مؤشرات للحكم على ثبات تقديرات المحكمين، وهي: الثبات الداخلي للمحكم بين الخطوات ntrajudge Reliability Between Steps، الثبات الداخلي للمحكم داخل الخطوات Intrajudge Reliability Within Steps والثبات بين المحكمين Interjudge Reliability. وأشار بيرك أيضا بأن التباين لأخطاء القياس يمكن حسابه في دراسات القابلية للتعميم، ويمكن عزل مكونات الثبات وحساب المؤشرات السابقة ومعامل القابلية للتعميم Generalizability Coefficients المحسوب من تحليل التباين بين المحكمين والذي يعزز الثقة بعلامة القطع.

وأظهرت نتائج دراسة جانج (Change, 1999) بأن طريقة أنجوف أكثر اتساقاً بين المحكمين من طريقة ندلسكي بالاعتماد على مؤشر معدل الفروق المطلقة بين الصعوبة المقدرة والحقيقية. كذلك أشارت نتائج الدراسة أن طريقة ندلسكي أعطت درجة قطع اقل من طريقة أنجوف وكانت اقرب لمتوسطات الصعوبة للفقرات في حال تقديرها للطالب الذي يملك الحد الأدنى من الكفاية المقبولة. وأشارت النتائج كذلك إن درجة القطع المقدرة باستخدام أسلوب أنجوف اقرب لمتوسطات الصعوبة للفقرات في حال تقديرها للطالب في مستوى التحصيل المتوسط.

وقد بينت كل من دراسة بيوكندل وسميث وامبارا وبليك (Bukendahl, Smith, Impara & Plake, 2000) ودراسة امبارا وبليك (Impara & Plake, 2000) أن طريقة أنجوف تنتج تقديرات قريبة جداً من التقديرات الني تنتج باستخدام كل من طريقة العلامة الفارقة Bookmark Method وطريقة المجموعة الحدية (أو الحد الفاصل) وطريقة ديلون Dilon لكن بصورة أسهل. كما فحصت دراسة بليك وامبارا (Plake & Impara, 2001) ثبات المحكمين في تقدير مستويات الأداء باستخدام طريقة أنجوف، وقد أشارت نتائجها إلى استقرار تقديرات المحكمين لعامين متتاليين وقد كان هناك توافق بين تقديرات المحكمين ومعاملات الصعوبة الفعلية للفقرات.

وقد أشارت نتائج دراسة الشريم (2003) إلى أن درجة القطع للاختبار الناتجة عن استخدام طريقة أنجوف في حالة عدم معرفة المحكمين لمعاملات الصعوبة تساوي 67 % في الجولة الأولى، و 68% في الجولة الثانية. وفي حالة معرفة المحكمين لمعاملات الصعوبة تساوي 63% في الجولة الأولى، و62% في الجولة الثانية. وتبين هذه النتائج أن تقديرات المحكمين لتقدير درجة القطع للاختبار تتأثر بمعرفتهم لمعاملات صعوبة الفقرات.

ويشير بيرك (Berk,1996) إلى أن الطريقة التي تقوم على التحكيم في تقدير درجة القطع للاختبارات تتأثر بنوعية المحكّمين و يجب أن تكون ملائمة فنيا بحيث يمكن الدفاع عنها. أي تسمح بالتصنيف الثنائي، وحساسة لخصائص الفقرات المختلفة، وتستخدم إجراءات إحصائية سهلة، وملائمة، وتأخذ بالحسبان أخطاء القياس، وتعطي دليلا على صدق القرارات، وقابلة للتطبيق وسهلة، ويتم إنجازها ضمن فترة زمنية معقولة.

واقترح هامبيلتون (Hamblton cited in Keller & Zaanetti, 2000) القيام بعدة جولات للتحكيم من أجل الوصول إلى تقديرات دقيقة؛ إذ يتم توفير تقديرات الجولة الاولى للمحكمين ليتم دراستها والاستفادة منها في تقدير درجة القطع في الجولات التالية من جولات التحكيم. وقد أكد ميلز وميلكان والويلا (Mills, Melican & Alluwalia, 1991) أهمية اختيار المحكمين ذوى المعرفة التامة في محتوى الاختبار، وتدريبهم على الطريقة المستخدمة، لأن هذا التدريب يؤدي للوصول إلى نتائج دقيقة في تحديد علامة القطع. وأشار بليك ومليكان وميلز (Plake,Melican & Mills, 1991) إلى أن اتساق المحكمين في تقدير درجة القطع يتأثر بنوعية المحكمين ونوعية الأسئلة في الاختبار والطريقة المستخدمة في تقديرها.

وقد اقترح كين (Kane cited in Plake, Impara & Irwin, 1999) أسلوبين لبيان صدق علامة القطع. يقوم الأول على مقارنة تقديرات المحكمين لصعوبة الفقرة مع النسبة الحقيقية للذين أجابوا عن الفقرة إجابة صحيحة. ويقوم الثاني على تحديد مجموعتين من الأفراد الذين تقع علاماتهم في جوارعلامة القطع المحددة، ومن ثم إجراء المقارنة بينهما من حيث الأداء، تتكون المجموعة الأولى من الأفراد الذين تزيد علاماتهم قليلاً على علامة القطع المحددة وتتكون المجموعة الثانية من الأفراد الذين تقل علاماتهم قليلاً عن علامة القطع المحددة. فإذا كانت نسبة الذين أجابوا على الفقرة إجابة صحيحة من المجموعة الأولى أعلى من نسبة الذين أجابوا إجابة صحيحة من المجموعة الثانية، فان ذلك يعد دليلاً آخر على صدق علامة القطع.

وقد كشفت دراسة بيهونياك وارشامبولت وجابل (Behuniak,Archambault & Gable, 1982) عن أنه لا يوجد أثر لخصائص المحكم الديمغرافية على تقدير علامة القطع. بينما بينت دراسة ميورر والكسندر (Maurer & Alexander, 1991) أنّ لخبرة المحكمين في تقدير علامة الاجتياز، أثراً في دقة التقديرات

وثباتها، مما يؤكد ضرورة استخدام خبراء في الموضوع عند استخدام طريقة أنجوف في تقدير علامة القطع.

وقد اهتم الباحثون بقضية البحث عن أدلة عملية ومنطقية تدعم جودة تقديرات المحكمين واتساقها. فقد استخدم ميورر والكسندر (Maurer & Alexander, 1991) للتدليل على دقة تقديرات المحكمين ثلاثة مؤشرات: (1) مؤشر دقة مسافة الفقرة Item Distance Accuracy حيث يدل حجم هذا المؤشر على مدى تقارب وتباعد المحكمين في تقدير درجة الصعوبة لكل فقرة، و(2) معامل الارتباط بين تقديرات المحكمين لدرجة صعوبة الفقرات ومعاملات الصعوبة الفعلية لها، و(3) معامل الارتباط بين تقديرات كل محكّم لصعوبة الفقرات ومتوسط تقديرات المجموعة الكلية لها Rater Total Correlation. كما استخدم لي سنغ (Lee-Sing, 2000) طريقة التحليل العنقودي Cluster Analysis في الكشف عن جودة تقديرات المحكمين لدرجات القطع. كما حاول كين (Kane,1987) استخدام نماذج نظرية الاستجابة للفقرة لفحص جودة علامات القطع المقدرة من قبل المحكمين من خلال الحكم على درجة مطابقة النموذج المستخدم لتلك التقديرات.

واستخدم انجلهارد واندرسون (Engelhard & Anderson, 1998) نموذج المحاولات ذات الحدين (BTM) Binomial Trials Model للحكم على جودة تقديرات المحكمين لدرجة القطع لاختبار رياضيات في ثلاث جولات للتحكيم. في الجولة الأولى لم يُعطى المحكمين أية بيانات حول أداء الطلبة على الفقرات، وقبل الجولة الثانية أعطي المحمكين درجات الصعوبة للفقرات والمتوسط الحسابي لتقديراتهم لها في الجولة الأولى، وقبل الجولة الثالثة أعطي المحمكين درجات الصعوبة للفقرات والمتوسط الحسابي لتقديراتهم لها في الجولة الثانية. وقد تم فحص جودة التقديرات من خلال مقارنة التقديرات الملاحظة مع التقديرات المتوقعة من النموذج واستخدام إحصائيات متوسط مربع الخطأ (MSE) ومؤشر ثبات الفصل (Reliability of separation index) ومربع كاي (Chi-square). وتجدر الإشارة إلى أن متوسط مربع الخطأ يوفر مؤشرات عن درجة مطابقة تقديرات المحكمين لصعوبة الفقرات مع القيم المتوقعة من نموذج المحاولات ذات الحدين. بينما يوفر مؤشر ثبات الفصل معلومات عن انتشار المحكمين والفقرات على متصل المتغير الكامن، وهو مناظر لمؤشرات الثبات التقليدية في المعنى ويعكس نسبة التباين الحقيقي إلى التباين الملاحظ، ويمكن معرفة قيمته لكل من المقدرين والفقرات. وتحسب قيمة الإحصائي مربع كاي لتحديد الدلالة الإحصائية للفروق في وجهات نظر المحكمين للكفاية الدنيا أو الفروق في درجات الصعوبة المقدرة للفقرات. أظهرت نتائج التحليل وجود فروق دالة إحصائياً (α = 0.01) في تقديرات المحكمين لصعوبة الفقرات باستخدام أسلوب أنجوف في كل جولة من جولات التحكيم كما يستدل على ذلك من قيمة معامل ثبات الفصل للفقرة (Reliability of item seperation) وقيمة اختبار مربع كاي (ChiSquare=720, R=0.92 للجولة الأولى و Chi Square=830, R= 0.93 للجولة الثانية و Chi Square=766, R= 0.93 للجولة الثالثة). وقد ازدادت درجة الاتفاق بين تقديرات المحكمين لصعوبة الفقرات مع القيم المتوقعة من النموذج من جولة إلى أخرى حيث كانت قيمة متوسط مربع الخطأ (Mean Square Error) أقل من 1.5 في 88% من الفقرات في الجولة الأولى وفي 98% من الفقرات في الجولة الثانية وفي جميع الفقرات في الجولة الثالثة. كذلك أظهرت النتائج استقراراً في معاملات الصعوبة المقدرة من قبل المحكمين حيث كانت قيمة معامل الارتباط بين التقديرات في الجولة الأولى والثانية 0.85 وبين التقديرات في الجولتين الثانية والثالثة 0.99.

يتضح من العرض السابق أهمية دقة تقدير درجة القطع لكلٍ من الطالب والعملية التربوية ومتخذ القرار، خاصة عندما تستخدم درجة القطع لأغراض التصنيف والاختيار ومنح الشهادة وتشخيص جوانب القوة والضعف في البرامج التربوية. ونظرا لأهمية درجات القطع وحساسيتها في اتخاذ القرارات التربوية تم الاهتمام بدقة تقديرها من خلال توفير بيانات فعلية للمحكمين عن أداء الطلبة على فقرات الاختبار أو من خلال تكرارعملية التقدير اكثر من مرة. ومن الواضح أن هذه الإجراءات تتطلب الكثير من الوقت والجهد للحصول على درجة قطع دقيقة للاختبار. وهذا يبرر البحث عن طرق يمكن من خلالها فحص تقديرات المحكمين لعلامة القطع وتحديدها بصورة دقيقة دون الحاجة لتكرار عملية التقدير أو الحصول على بيانات فعلية من المفحوصين. ونظراً لندرة الدراسات العربية التي اهتمت بفحص تقديرات المحكمين _في حدود علم الباحثين_ تأتي هذه الدراسة بوصفها محاولة لاستخدام نموذج المحاولات ذات الحدين في فحص جودة تقديرات المحكمين وفقا لطريقة "أنجوف" لدرجة القطع لاختبار مرجعي المحك. ويقوم هذا النموذج على توزيع المحاولات ذات الحدين الذي يكون فيه احتمال النجاح ثابتا ومستقلا عند تكرار التجربة العشوائية اكثر من مرة، ويصف توزيع المحاولات ذات الحدين التجارب العشوائية التي يكون نواتجها ناتجين فقط، الناتج الأول يسمى نجاحاً ويرمز له بالواحد (1) والثاني فشلاً ويرمز له بالصفر (0).

وأشار انجلهارد واندرسون (Engelhard & Anderson, 1998) إلى أنه يمكن استخدام نموذج المحاولات ذات الحدين (BTM) في تحديد مستوى الأداء من خلال أن يرى المحكّم النتيجة كعملية عد لمحاولات النجاح. فعلى سبيل المثال عند استخدام طريقة أنجوف يمكن الطلب من المحكم الإجابة عن السؤال التالي "من مائة مفحوص يملكون الحد الأدنى للكفاية (أدنى مستوى مقبول للأداء من وجهة نظر المحكم) قدر عدد المفحوصين الذين يجيبون عن الفقرة إجابة صحيحة؟". وبتطبيق نموذج المحاولات ذات الحدين (Binomial Trials Model) يمكن تقدير احتمال النجاح P_{ni} على الفقرة i بالاعتماد على رؤية المحكم لمستوى الكفاية الدنيا وصعوبة الفقرة على النحو الآتي:

$$p_{ni} = \frac{\exp(\beta_n - \delta_i)}{[1+\exp(\beta_n - \delta_i)]}, \qquad (1)$$

حيث إن:-

β_n: تمثل الحد الأدنى للنجاح المقدر من قبل المحكم N

δ_i: تمثل معامل الصعوبة المقدر من قبل المحكم للفقرة I

وعندها يصبح الرقم (100) هو عدد المحاولات (m) والمطلوب من المحكم هو عد عدد المفحوصين (x) الذين هم في أدنى مستوى للأداء المقبول والمتوقع أن تكون إجابتهم عن الفقرة إجابة صحيحة. وهذا النموذج يمكن تمثيله بالمعادلة التالية: -

$$\Pr(X = x_{ni} | m, \beta_n, \delta_i) = \pi_{nix} = \binom{m}{x_{ni}} \frac{\exp[x_{ni}(\beta_n - \delta_i)]}{[1+\exp(\beta_n - \delta_i)]^m}, \qquad (2)$$

حيث إن x_{ni} = عدد المفحوصين من بين m من المفحوصين الذين يملكون الحد الأدنى من الكفاية المقدرة من قبل المحكم N الذين سيجيبون على الفقرة I التي درجة صعوبتها المقدرة δ_I إجابة صحيحة.وتجدر الإشارة إلى أن المعادلة 2 يمكن أن ترى كنموذج راش متعدد الأوجه Many-Faceted Rasch Model (FACETS) (Linacer, 2003).

وقد استخدم هذا النموذج في الدراسة الحالية لكونه نموذج احتمالي له أساس رياضي ينسجم مع نماذج نظرية الاستجابة للفقرة. كما أنه يفيد في الكشف عن أنماط التقديرات الشاذة وبالتالي يكشف عن المحكمين الذين يعطون تلك التقديرات فيتم إستبعادهم من عملية التحكيم مما يساعد في إعادة حساب درجة القطع اعتماداً على تقديرات المحكمين التي تطابق النموذج. وتتميز الدراسة الحالية بأنه يتوقع منها أن توثق الأثر الايجابي لاستبعاد أنماط التقديرات الشاذة على جودة درجة القطع المقدرة. ويتمثل ذلك بالحصول على تصنيفات أكثر أستقراراً للمفحوصين في ضوء درجة القطع المقدّرة من المحكمين إلى متمكنين وغير متمكنين. ويستدل على ذلك بالتحسن الذي يحدث على قيم كل من معامل دقة القرار (نسبة الاتفاق في قرارات تصنيف الطلبة إلى متمكنين وغير متمكنين عند استخدام علامة القطع وإعادة استخدامها مرة أخرى مع نفس الطلبة عند تكرار عملية القياس) ومعامل كابا (نسبة الاتفاق في قرارات تصنيف الطلبة إلى متمكنين وغير متمكنين بعد أخذ العشوائية بالاعتبار) بالاعتماد على علامة قطع مقدرة من المحكمين بعد فحص تقديراتهم باستخدام نموذج المحاولات ذات الحدين مقارنة مع قيمها قبل فحصها. ويقصد بدرجة القطع في الدراسة الحالية العلامة المقدرة من المحكمين باستخدام أسلوب أنجوف، والتي ينبغي أن يحصل المفحوص على علامة تساويها أو أعلى منها لكي يعد متمكناً من المجال السلوكي المحدد للاختبار، أي هي الحد الأدنى لمستوى الأداء المقبول.

مشكلة الدراسة وأسئلتها

تعد مسألة تحديد درجة القطع من القضايا الأساسية للاختبارات التي تستخدم في تصنيف الطلبة، واتخاذ قرارات مهمة، تتعلق بتحديد موقع الفرد بالنسبة إلى الحد الأدنى من مستوى الكفاية المقبول. وينبغي أن تتصف درجة القطع بالصدق والاتساق في تصنيف أداء الأفراد وتفسيره، وعند الاعتماد على المحكمين لتقدير درجة القطع، فإن جودة تقديرات المحكمين واتساقها ضرورية جدا لصدق درجة القطع الناتجة عن تلك التقديرات. الأمر الذي يحتم فحص تقديرات المحكمين لضمان جودتها واتساقها، بهدف الوصول إلى درجة قطع صادقة. وتسعى الدراسة الحالية إلى فحص تقديرات المحكمين لدرجة القطع لاختبار مرجعي المحك في الرياضيات، لقياس كفايات طلبة الصف العاشر الأساسي في المعارف والمهارات الأساسية في وحدة الهندسة التحليلية في ثلاث جولات للتحكيم. وستحاول هذه الدراسة على وجه التحديد الإجابة عن الأسئلة التالية:-

- هل تختلف تقديرات المحكمين لدرجة القطع للاختبار باختلاف جولة التحكيم عند مستوى الدلالة الإحصائية (α=0.01)؟
- ما جودة تقديرات المحكمين لدرجة القطع قبل فحصها باستخدام نموذج المحاولات ذات الحدين وبعده ؟

الطريقة والاجراءات

عينة الدراسة

تكونت عينة الدراسة من 110 طالباً وطالبة اختيروا عشوائيا من ست مدارس تم اختيارها عشوائيا من المدارس الحكومية التابعة لمحافظة عجلون للعام الدراسي 2004/2003 التي تتضمن الصف العاشر الأساسي. كما تمّ اختيار 15 معلماً ومعلمة بالطريقة العشوائية من بين معلمي الرياضيات للصف العاشر الأساسي الذين لديهم مؤهل تربوي بالاضافة لبكالوريوس الرياضيات كمحكمين. وقد تم اختيارعينة المحكمين من بين المعلمين الذين يقومون بتدريس المادة لكونهم أكثر معرفة بمحتوى المادة من جهة وبمستوى الطلبة من جهة أخرى.

أداة الدراسة

لتحقيق أهداف الدراسة تم بناء اختبار تحصيلي مرجعي المحك في مبحث الرياضيات لطلبة الصف العاشر الأساسي، يقيس المهارات والمفاهيم والمصطلحات والتعميمات المتعلقة بوحدة الهندسة التحليلية وفق خطوات بناء الاختبارات مرجعية المحك (علام، 2001؛ Popham, 2000).

ويعد تحديد المجال السلوكي للاختبار خطوة هامة في بناء الاختبار مرجعي المحك، ويعتمد تحديد المجال السلوكي على المحتوى الذي سيقيسه الاختبار، ويكون المجال السلوكي بمنزلة المعيار الذي ينسب إليه أداء الفرد. وبالاعتماد على التحليل التفصيلي لوحدة الهندسة التحليلية تم تحديد المجال السلوكي

للاختبار بسبع كفايات أساسية (استنتاج ميل المستقيم إذا علمت زاوية ميله، ايجاد معادلة الخط المستقيم في ظل معطيات محددة، تعرف الصيغة العامة لمعادلة الخط المستقيم، التحقق من تعامد مستقيمين وتوازيهما، ايجاد بعد نقطة عن مستقيم إذا علمت معادلته، تعرف المحل الهندسي لنقطة تتحرك وفق شروط معينة وايجاد معادلته، وايجاد معادلة الدائرة وفق شروط محددة) تمثل المجال السلوكي الذي يقيسه الاختبار، وجرى ترجمتها إلى أهداف سلوكية. وللتأكد من شمول الأهداف السلوكية للمجال السلوكي تم عرضهاعلى سبعة من الخبراء المختصين في تدريس الرياضيات لإبداء الرأي حول مدى شمولها، ووضوح صياغتها. وقد تم مناقشة ملاحظاتهم وتحليلها وإجراء التعديلات التي اقترحها الخبراء من إضافة وحذف وإعادة صياغة لبعض الأهداف السلوكية، ونتيجة لذلك تم الاتفاق على صياغة 24 هدفاً سلوكياً تغطي الكفايات المذكورة سابقاً (مثل: يجد ميل المستقيم المار بنقطتين معلومتين، يتعرف مفهوم زاوية ميل المستقيم، يجد معادلة المستقيم إذا علم ميله ونقطة عليه،...) ولكل هدف سلوكي تم صياغة فقرتين إلى أربع فقرات من نوع الاختيار من متعدد لكل منها خمسة بدائل بصورة أولية، للحصول على فقرات اكبر مما هو مطلوب نظراً لأن بعضها ربما يستبعد عند مراجعته وتدقيقه، بحيث روعي في صياغة مثل هذا النوع من الفقرات الإرشادات والمحكات والشروط الفنية لصياغتها (عوده، 1993؛ Haladyna, Downing & Rodriguez, 2002) كما روعي أن تكون الفقرة مرتبطة بالهدف الذي تقيسه من حيث المحتوى ودرجة الصعوبة وأن لا تعتمد إجابة إحدى هذه الفقرات على إجابة الفقرات الأخرى. وقد تم اختيار فقرة أو فقرتين بشكل عشوائي لكل هدف سلوكي.

بعد إخراج الاختبار بصورته الأولية، عرض على 25 محكماً من أهل الاختصاص في الرياضيات والعلوم التربوية. وطلب منهم تقدير درجة مطابقة الفقرة للهدف السلوكي الذي تقيسه على تدريج خماسي، و مدى مراعاة الفقرات لقواعد وأسس كتابة الفقرات من نوع الاختيار من متعدد. وقد تم استرجاع 24 نسخة من نسخ التحكيم وتم استخراج الأوساط الحسابية والانحرافات المعيارية للتقديرات. وتبين أن قيم الأوساط الحسابية لتقديرات المحكمين لدرجة مطابقة الفقرة للهدف الذي تقيسه، تراوحت بين 4.33 و 4.71 وهي قيم عالية، وتراوحت قيم الانحرافات المعيارية بين 0.46 و 0.82 وهي قيم منخفضة. مما يعني اتفاق المحكمين على مطابقة كل فقرة للهدف الذي تقيسه.

أما عن مدى مراعاة الفقرات لمعايير كتابة الفقرات من نوع الاختيار من متعدد، فقد تبين أن هناك نسبة اتفاق عالية بين المحكمين على كل من مدى دقة الصياغة اللغوية للفقرات ووضوحها ومدى ملائمة البدائل للفقرات، حيث زادت النسبة المئوية للاتفاق على (88%).وقد تكون الاختبار بصورته الأولية من 30 فقرة قدر الباحثان والخبراء المختصون زمن إجابتها بساعة.

وبهدف تحديد الزمن الفعلي لتطبيق الاختبار، ودراسة مدى وضوح فقراته وتعليماته، ورصد الاستفسارات الواردة من الطلبة، تم تطبيق الاختبار على اثني عشر طالباً هم جميع طلبة الصف العاشر الأساسي في مدرسة اختيرت عشوائياً من خارج عينة الدراسة. وقد تبين أن تعليمات الاختبار وفقراته واضحة لدى معظم الطلبة الذين أنهى عشرة (حوالي 83%) منهم الاختبار في 50 دقيقة. وهذا يدل على مناسبة الزمن الذي تم تخصيصه للاختبار.

وبهدف دراسة فاعلية مموهات الفقرات وتقليل عدد بدائلها ليصبح أربعة بدائل، تم تجريبهاعلى 120 طالباً وطالبة من طلبة الصف العاشر الأساسي ممن انهوا دراسة وحدة الهندسة التحليلية. وتم حساب معاملات التمييز لجميع المموهات، وقد كانت غالبية المموهات جذابة؛ إذ كانت معاملات التمييز للموهات (البدائل الخطأ) سالبة في 85% من الحالات.وللتحقق من أن الاختبار يميز بين المجموعات المتمايزة وأن الفقرات حساسة للتدريس تم تطبيقه على 125 طالب وطالبة اختيروا عشوائياً من ثلاث مدارس، من خارج عينة الدراسة قبل بدء تدريس وحدة الهندسة التحليلية لهم، وبعد انتهاء تعلمهم لها. ويقصد بمؤشر حساسية الفقرة للتدريس الفرق بين معاملي الصعوبة لها قبل التدريس وبعده. وقد استخدم اختبار مكنمار (McNemar) (Hays, 1980) للفرق بين النسب المرتبطة في جدول رباعي (2 X 2) لاختبار دلالة الفرق بين معاملات الصعوبة للفقرات لعينة الطلبة قبل التدريس وبعده. وتوزيع هذا الاختبار هو توزيع مربع كاي بدرجات حرية تساوي 1. وقد تبين بأن هناك فروقاً ذات دلالة إحصائية لجميع الفقرات؛ إذ بلغت قيمة الاختبار 4.05 للفقرة 26 وهي أعلى من القيمة الحرجة 3.84 عند مستوى الدلالة الإحصائية (α=0.05) بينما كانت أقل قيمة له للفقرات الأخرى 12.96 وهي أعلى من القيمة الحرجة 6.635 عند مستوى الدلالة الإحصائية (α=0.01). وتراوحت قيم مؤشر حساسية الفقرة للتدريس بين 0.08 للفقرة 26 و0.62 وجميعها قيم موجبة تدل على حساسية الفقرات لعملية التدريس وتأثرها بها والذي يعد مؤشراً من مؤشرات صدق الفقرة في الاختبار مرجعي المحك. وقد ميز الاختبار بصورته الكلية بين أداء الطلبة قبل تدريس وحدة الهندسة التحليلية وبعد تدريسها حيث كان الفرق بين الوسطين 13.7 وكانت قيمة الاختبار t للبيانات المرتبطة 22.25 وهي دالة إحصائيا عند مستوى دلالة يقل عن 0.01.

إجراءات جمع البيانات

تم تطبيق الاختبار بصورته النهائية التي تحتوي 30 فقرة لكل منها أربعة بدائل بعد حذف أضعف المموهات على عينة الدراسة والبالغ عددها 110 طالباً وطالبة بعد أن انهوا دراسة وحدة الهندسة التحليلية. وبعد جمع إجابات الطلبة على فقرات الاختبار، صححت الإجابات بإعطاء درجة واحدة للإجابة الصحيحة وصفر للإجابة الخاطئة. كما تم إعادة تطبيق الاختبار على عينة الدراسة مرة أخرى بعد مرور أسبوعين على التطبيق الأول. وقد تم استخراج معاملات الصعوبة الفعلية للفقرات بالاعتماد على البيانات

من التطبيق الأول وتبين أنها تقع بين 0.36 و0.71 بوسط حسابي 0.54. وستوفر علامات المفحوصين الأساس لتقدير معامل دقة القرار في تصنيفهم إلى فئتين في ضوء علامة القطع المقدرة من المحكمين في مرتي التطبيق.

كما تم جمع البيانات من المحكمين لغايات تحديد درجة القطع وفق المراحل التالية:-

- تم الالتقاء بعينة المحكمين الذين طلب منهم تحديد درجة القطع للاختبار،بعد تقديم معلومات تفصيلية لهم عن هدف الدراسة، وتوضيح الأساس المنطقي لطريقة أنجوف التي سيستخدمونها في تحديد الحد الأدنى لمستوى الأداء المقبول، كما تمّ تزويدهم بنشرة تفصيلية عن خطوات تحديد مستوى الأداء وفق أسلوب أنجوف مشفوعة بالأمثلة التطبيقية.
- في الجولة الأولى من جولات التحكيم، طلب من كل محكم تقدير الحد الأدنى لمستوى الأداء باستخدام أسلوب أنجوف، وفق نموذج خاص يشتمل على التعليمات وفقرات الاختبار. وأمام كل فقرة طلب من كل محكم إجابة السؤال "من بين 100 طالب ترى أنهم يملكون الحد الأدنى من الكفاية (مستوى الأداء المقبول من وجهة نظرك) في محتوى الهندسة التحليلية في مبحث الرياضيات للصف العاشر الأساسي، قدر عدد الذين سيجيبون عن الفقرة إجابة صحيحة؟". وبعد استرداد نماذج تحديد مستويات الأداء من المحكمين، تم تحديد مستوى الأداء المقدر لكل فقرة من فقرات الاختبار استنادا إلى البيانات التي تم الحصول عليها. من خلال إيجاد الأوساط الحسابية لتقديرات المحكمين لكل فقرة من فقرات الاختبار، الذي يعبر عنه بمعامل الصعوبة المقدر للفقرة من قبل المحكمين، وحساب درجة القطع للاختبار من خلال إيجاد الوسط الحسابي لجميع معاملات الصعوبة المقدرة للفقرات.
- في الجولة الثانية، تم توزيع نموذج خاص لمستوى الأداء وفق أسلوب أنجوف يشتمل على التعليمات والفقرات والوسط الحسابي لمعامل الصعوبة المقدر من المحكمين لكل فقرة من فقرات الاختبار الذي تم إيجاده في الجولة الأولى. طلب من المحكمين تقدير عدد الطلاب من بين 100 طالب يملكون الحد الأدنى من الكفاية في محتوى الهندسة التحليلية، الذين يستطيعون الإجابة عن الفقرة إجابة صحيحة في ضوء معرفتهم بالوسط الحسابي لتقديراتهم لكل فقرة من فقرات الاختبار في الجولة الأولى. وبعد استرداد نماذج التحكيم تم تحديد مستوى الأداء المقدر لكل فقرة من فقرات الاختبار بنفس أسلوب الجولة الأولى.
- في الجولة الثالثة، تم تزويد المحكمين بمعاملات الصعوبة الفعلية لفقرات الاختبار. وطُلب منهم إعادة النظر في تقديراتهم في ضوء خصائص الفقرة الفعلية. وتم تحديد مستوى الأداء ودرجة القطع بالأسلوب السابق نفسه.

النتائج ومناقشتها

أولاً: النتائج المتعلقة بتقديرات المحكمين لدرجات القطع للاختبار

لمعرفة تقديرات المحكمين لدرجات القطع للاختبار في كل جولة من جولات التحكيم تم إيجاد الأوساط الحسابية والانحرافات المعيارية لتقديرات كل محكم لنسبة الذين سيجيبون عن فقرات الاختبار إجابة صحيحة من بين الذين هم في مستوى الكفاية الدنيا. ويمثل الوسط الحسابي لتقديرات كل محكم علامة القطع للاختبار المقدرة من ذلك المحكم. ويمكن النظر لعلامة القطع هنا باعتبارها النسبة المئوية لفقرات الاختبارالتي يجب إجابتها إجابة صحيحة من قبل المفحوص ليصنف في أدنى مستوى للأداء المقبول في المجال السلوكي الذي يقيسه الاختبار. ويبين الجدول (1) الأوساط الحسابية والانحرافات المعيارية لتقديرات المحكمين لدرجات القطع في الجولات التحكيمية الثلاث.

الجدول (1): الأوساط الحسابية والانحرافات المعيارية لتقديرات المحكمين لفقرات الاختبار باستخدام أسلوب انجوف في الجولات الثلاث

رقم المحكم	الجولة الأولى		الجولة الثانية		الجولة الثالثة	
	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
1	70.13	9.27	63.23	7.60	63.77	8.03
2	58.13	7.65	51.77	9.31	48.77	11.34
3	66.03	7.05	63.73	8.88	60.10	4.84
4	64.57	10.44	54.90	10.75	69.70	6.04
5	73.83	8.42	65.83	9.25	57.17	5.38
6	57.43	10.51	54.77	11.53	55.53	12.00
7	72.33	6.94	68.77	9.20	54.63	10.13
8	66.93	10.56	65.13	11.87	58.67	11.61
9	69.77	9.11	66.03	9.00	51.07	10.48
10	73.63	7.59	68.17	9.42	58.60	6.91
11	74.73	9.01	66.03	10.32	52.13	11.68
12	72.10	9.20	67.03	7.62	59.80	8.27
13	58.17	8.25	54.30	7.70	56.47	9.97
14	75.33	6.81	75.83	4.75	53.57	9.12
15	45.03	9.00	55.53	14.85	52.07	13.12
الوسط الحسابي	66.54	8.65	62.74	9.47	56.80	9.26

تظهر النتائج الواردة في الجدول (1) أن قيم الأوساط الحسابية لدرجات القطع المقدرة لفقرات الاختبار في الجولة الأولى تتراوح بين 45.03 و74.73 بوسط حسابي 66.54، وبذلك تصبح علامة النجاح على الاختبار 67%، أي أن الطالب يجب أن يجيب عن 20 فقرة إجابة صحيحة حتى يعد متمكناً من المجال السلوكي الذي يقيسه الاختبار. وفي الجولة الثانية تراوحت قيم الأوساط

الحسابية لتقديرات المحكمين لدرجات القطع بين 54.30 و75.83 بوسط حسابي 64.07 وتعّد هذه القيمة بمثابة درجة القطع للاختبار، أي أن الطالب يجب أن يجيب عن 19 فقرة إجابة صحيحة حتى يعد متمكناً. وأظهرت تقديرات المحكمين في الجولة الثالثة أنّ قيم الأوساط الحسابية تتراوح بين 48.77 و69.77 بوسط حسابي 56.80، وبذلك تكون درجة القطع 57%، أي أن الطالب يجب أن يجيب عن 17 فقرة إجابة صحيحة حتى يعد متمكناً.

وللكشف عن تأثير اختلاف الجولات التحكيمية في تقديرات المحكمين لدرجات القطع تم استخدام تحليل التباين للقياسات المتكررة (Repeated Measurement Design) لنتائج تقديرات المحكمين لدرجات القطع باستخدم أسلوب أنجوف في الجولات الثلاث. ويقوم هذا التحليل على مجموعة من الإفتراضات هي: الاختيار العشوائي للمحكمين والتوزيع الطبيعي لكل من أخطاء التقدير وأثر التفاعل بين المحكم وجولة التحكيم وتجانس التباين لتلك الأخطاء والتفاعلات. ولا يخفى أن المحكمين في الدراسة الحالية يشكلون عينة عشوائية وأن الأوساط الحسابية للتقديرات تتوزع بصورة طبيعية، ونظراً لأن الفروق بين معاملات الارتباط الثنائية بين الجولات (0.92 بين الاولى والثانية، 0.70 بين الاولى والثالثة، 0.68 بين الثانية والثالثة) ليست دالة إحصائياً (α=0.01) _ إذ بلغت قيمة إختبار t للفرق بين أعلى معامل ارتباط وأدنى معامل ارتباط 2.76 وهي أقل من القيمة الحرجة 3.054 عند درجات حرية تساوي 12 _ فإن ذلك يعد دليلاً على تحقق تجانس التباين (Feldt, 1988). ويلخص الجدول (2) نتائج تحليل التباين لهذه التقديرات.

الجدول (2): نتائج تحليل التباين للقياسات المتكررة لدرجات القطع المقدرة للاختبار في الجولات الثلاث

مصدر التباين	درجات الحرية	مجموع المربعات	وسط المربعات	ف المحسوبة	قيمة الاحتمال
جولات التحكيم	2	768.866	384.43	11.743	0.000
المحكمين	14	966.664	69.047	2.109	0.045
التفاعل	28	916.649	32.737		

تظهر النتائج في الجدول (2)، بأنه توجد فروق ذات دلالة إحصائية (α=0.01) بين جولات التحكيم في تقديردرجات القطع للاختبار. وقد كانت الفروق وفق طريقة توكي بين الجولة الأولى وكل من الجولة الثانية والثالثة دالة إحصائيا (α=0.01). وقد أخذت تقديرات المحكمين لدرجات القطع تتناقص مع الجولات التحكيمية. وربما يعود السبب في ذلك لاطلاع المحكمين على تقديراتهم الأولية في الجولة الثانية، وعلى معاملات الصعوبة الفعلية في الجولة الثالثة. وقد كانت درجات القطع المقدرة في الجولتين الثانية والثالثة اقل من درجة القطع المقدرة في الجولة الأولى. واتفقت هذه النتيجة مع نتائج العديد من الدراسات السابقة (الشريم، 2003؛ Bukendahi, Smith, Impara & Plake, 2000; Chinn & Hertz, 2002; Croos, Impara, Frary, & Jager, 1984). وقد يكون من أهم الأسباب التي جعلت تقديرات المحكمين لصعوبة فقرات الاختبار تختلف باختلاف جولة التحكيم، هي تأثر المحكمين بالمعلومات المعطاة لهم أثناء عملية التقدير، الأمر الذي جعلهم يقومون بمراجعة تقديراتهم الأولية بالاعتماد على تلك المعلومات سواء كانت تتعلق بقيمة الوسط الحسابي لتقديراتهم في الجولة الأولى أو بالصعوبة الفعلية للفقرات. وهذا ما جعل تقديراتهم في الجولتين الثانية والثالثة تميل نحو الوسط الحسابي لتقديراتهم في الجولة الأولى أو نحو الوسط الحسابي (0.54) لصعوبة الفقرات الفعلية. ويظهر جلياً أن تقديرات المحكمين لدرجات القطع في الجولة الثانية كانت قريبة من تقديراتهم في الجولة الأولى، لاعتمادهم على تقديراتهم في الجولة الأولى والتي تم توفيرها لهم أثناء عملية التقدير في الجولة الثانية، والذي أبرز الميل نحو النزعة المركزية في تقديراتهم في الجولة الثانية، حيث مال معظم المحكمين في الجولة الثانية من ذوي التقديرات العالية في الجولة الأولى نحو الوسط الحسابي لتقديراتهم في الجولة الأولى. في حين انخفض مدى التقديرات في الجولة الثالثة، واقترب الوسط الحسابي لها بصورة أوضح من الوسط الحسابي لصعوبة الفقرات الفعلية، مما يدعم تأثر تقديرات المحكميين بمعاملات الصعوبة الفعلية لفقرات الاختبار. وقد وفّرت قيم معاملات الارتباط بين تقديرات المحكمين لصعوبة الفقرات وبين معاملات الصعوبة الفعلية لها دليلا إضافياعلى تأثر تقديرات المحكمين بمعاملات الصعوبة الفعلية لفقرات الاختبار؛ إذ تبين أن قيم معامل ارتباط بيرسون في الجولتين الأولى والثانية (0.51، 0.49) كانت منخفضة مقارنة مع الجولة الثالثة (0.95) على الرغم من دلالتها الإحصائية. وقد جاءت هذه النتيجة متفقة مع نتائج الدراسات السابقة (الشريم، 2003؛ Croos, Impara, Frary, & Jager, 1984; Plake & Impara, 2001). إن ظهور مثل هذه النتيجة، يؤكد أنّ تزويد المحكمين بمعاملات الصعوبة يؤثر بشكل مباشر على تقديرات المحكمين لدرجات القطع، مما يجعل تقديراتهم تقترب من معاملات الصعوبة الفعلية.

كذلك فان اكتساب المحكم للخبرة في تحديد درجة القطع وتغير فهمه لمفهوم الطالب الذي يملك الحد الأدنى من الكفاية من خلال معرفته للصعوبة الفعلية للفقرات، قد تكون من العوامل التي تؤدي إلى تغيير تقديراته في الجولات اللاحقة. وتجدر الإشارة أيضا أنه عند استخدام طريقة أنجوف في تقدير صعوبة الفقرات، فان المحكم قد يركز انتباهه على عدد الذين يستطيعون الإجابة على الفقرة إجابة صحيحة ممن يملكون الحد الأدنى من الكفاية من الفئة المستهدفة بالاختبار، وهذه القضية تجعل المحكم يتأثر بنوعية الفئة المستهدفة بالاختبار أثناء عملية التقدير، وبفهمه لمفهوم الحد الأدنى من الكفاية. لذلك ظهرت فروق دالة إحصائيا (α=0.05) بين المحكمين؛ إذ قد تعزى هذه الفروق إلى اختلاف خصائص طلبة الصف العاشر الأساسي من مدرسة إلى أخرى واختلاف درجة ألفة المحكم بطريقة أنجوف المستخدمة في تقدير درجة القطع.

كما وتجدر الإشارة، إلى أن جميع درجات القطع الناتجة عن تقديرات المحكمين جاءت أعلى من 50% على مدار الجولات الثلاث. والذي يؤكد النتيجة التي توصل إليها شانج (Change, 1999) بأن أسلوب أنجوف ينتج درجات قطع عالية وغالباً ما تكون أعلى من 50%. وقد برر ذلك بأن معظم المحكمين غالباً ما يفكرون بالطالب في مستوى الوسط (أداءه على الاختبار مقارنة بغيره يكون في وسط التوزيع) أو فوقه، ولا يفكرّ الا القليل منهم بالطالب الذي يمتلك الحد الأدنى من الكفاية (الذي وصل علامة النجاح)، كما أنهم يتأثرون بنوعية وقدرات الطلبة الذين يدرسونهم وتوقعاتهم العالية لهم، مما يجعل تقديراتهم لدرجات القطع تميل نحو التشدد.

ثانيا: النتائج المتعلقة بفحص جودة تقديرات المحكمين باستخدام نموذج المحاولات ذات الحدين

استخدمت الدراسة الحالية نموذج المحاولات ذات الحدين في فحص تقديرات المحكمين لدرجات القطع، حيث استخدمت محكات المطابقة مع النموذج في غربلة التقديرات. وقد اُعتمد معيار ثبات التصنيف للافراد باستخدام درجات القطع المقدرة من المحكمين إلى متمكنين أو غير متمكنين للحكم على جودة هذا النموذج في عملية انتقاء التقديرات. و لتقدير ثبات التصنيف باستخدام درجات القطع المقدرة من المحكمين تم حساب قيم كل من معامل كابا ومعامل دقة القرار (معامل الصواب hit rate) قبل استخدام نموذج المحاولات ذات الحدين وبعد استخدامه. ويشير معامل دقة القرار إلى نسبة الاتساق في قرارات تصنيف الطلبة، أما معامل كابا فيشير إلى نسبة الاتساق في قرارات التصنيف بعد أخذ العشوائية بالاعتبار. ويبين الجدول (3) تصنيفات الطلبة (N= 110) في ضوء درجات القطع المقدرة للاختبار في مرّتي التطبيق إلى متمكنين وغير متمكنين من المجال السلوكي المستهدف وقيم معاملات كابا ودقة القرار قبل استخدام نموذج المحاولات ذات الحدين في انتقاء التقديرات.

الجدول (3): تصنيفات الطلبة المختلفة في التطبيقين الأول والثاني للاختبار، ودرجات القطع المقدرة ومعاملات كابا ودقة القرار قبل استخدام نموذج المحاولات ذات الحدين

الجولة التحكيمية	درجة القطع %	عدد المتمكنين في التطبيقين	متمكن في الأول وغير متمكن في الثاني	متمكن في الثاني وغير متمكن في الأول	غير متمكن في التطبيقين	معامل كابا	معامل دقة القرار
الأولى	67	30	7	6	67	0.73	0.88
الثانية	64	42	5	3	60	0.85	0.93
الثالثة	57	55	2	2	51	0.93	0.96

يلاحظ من الجدول (3) أن درجة القطع في الجولة الأولى هي إجابة 67 % من فقرات الاختبار بصورة صحيحة (20 فقرة) وقد توزع الطلبة (N=110) في ضوء هذه الدرجة إلى متمكنين وغير متمكنين في مرّتي تطبيق الاختبارعليهم على الفئات الأربع الممكنة على النحو الآتي: 30 صنفوا إلى متمكنين في التطبيق الأول ومتمكنين في التطبيق الثاني، 7 صنفوا إلى متمكنين في التطبيق الأول وغير متمكنين في التطبيق الثاني، 6 صنفوا إلى غير متمكنين في التطبيق الأول ومتمكنين في التطبيق الثاني، و67 صنفوا إلى غير متمكنين في التطبيق الأول وغير متمكنين في التطبيق الثاني. ويلاحظ أن درجات القطع قبل فحص التقديرات باستخدام نموذج المحاولات ذات الحدين تتناقص من جولة إلى أخرى، ويبلغ أكبر فرق بينها 10 درجات، وأن قيم معامل كابا ومعامل دقة القرار، تتزايد من جولة إلى أخرى. وقد بلغ أعلى فرق في معامل دقة القرار 0.08 وفي معامل كابا 0.20. وقد يعزى التحسن في ثبات التصنيف إلى الخبرة المكتسبة من الجولات السابقة من ناحية وإلى المعلومات الاضافية التي تم توفيرها للمحكمين سواء كانت تتعلق بالوسط الحسابي لتقديراتهم السابقة أو الصعوبة الفعلية للفقرات. ويلاحظ أن قيم معامل كابا كانت اقل من قيم معامل الصواب في الجولات الثلاث لأن معامل كابا يأخذ بالاعتبار أخطاء التصنيف.

ولمعرفة درجة مطابقة تقديرات المحكمين مع نموذج المحاولات ذات الحدين تم استخدام برمجية FACETS، المصممة لتحليل البيانات، وفقاً لنموذج راش متعدد الأوجه (Many Facet (Rash Model (Linacer, 2003). وقد تم تشكيل مصفوفة البيانات المكونة من وجهين، الوجه الأول الفقرات (Item facet)، والوجه الثاني المحكمين (Judge facet) لكل جولة من جولات التحكيم الثلاث. وبتطبيق نموذج المحاولات ذات الحدين، تم إيجاد تقدير كل محكم للحد الأدنى من الكفاية (β_n) بوحدات الترجيح اللوغاريتمي لكل جولة تحكيمية، بالإضافة إلى إحصائي المطابقة ZSTD (The Standardized Information fit Statistics for judge) الذي تتوزع قيمته بصورة طبيعية بمتوسط حسابي يساوي صفر وانحراف معياري يساوي 1 عندما تتوافق تقديرات المحكم للحد الأدنى من الكفاية مع القيم التي يتوقعها النموذج، وإحصائي متوسط مربعات الأخطاء (MSE_n (Mean Square Error الذي يتوقع أن تكون قيمته واحد صحيح عندما تتوافق تقديرات المحكم لصعوبة الفقرات تماماً مع قيم الصعوبة المتوقعة من نموذج

المحاولات ذات الحدين. وللحكم على درجة مطابقة كل من تقديرات المحكمين للحد الأدنى من الكفاية (درجات القطع) وتقديراتهم لصعوبة الفقرات اعتمدت المحكات بأن تكون قيمة الإحصائي ZSTD للمحكم محصورة بين ± 2 وقيمة MSE_n للمحكم محصورة بين 0.6 و1.5 (Linacer, 2003). ويلخص الجدول (4) تقديرات المحكمين للحد الأدنى من الكفاية لفقرات الاختبار في كل جولة تحكيمية بالإضافة إلى قيمة وسط مربعات الأخطاء MSE_n والإحصائي ZSTD.

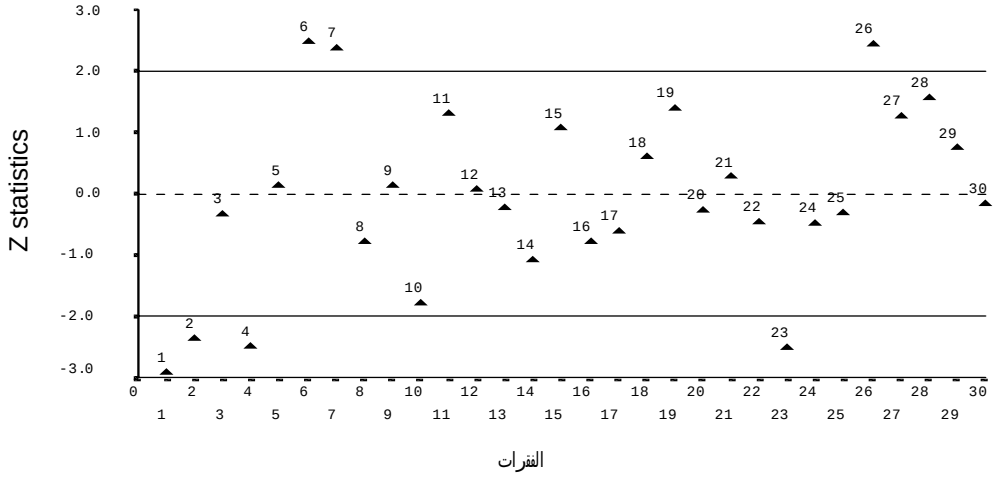
الجدول (4): درجات القطع المقدرة من المحكمين لفقرات الاختبار وقيم MSE_n و ZSTD في الجولات الثلاث

رقم المحكم	الجولة الأولى			الجولة الثانية			الجولة الثالثة		
	درجة القطع β_n	وسط مربعات الأخطاء MSE_n	قيمة الإحصائي ZSTD	درجة القطع β_n	وسط مربعات الأخطاء MSE_n	قيمة الإحصائي ZSTD	درجة القطع β_n	وسط مربعات الأخطاء MSE_n	قيمة الإحصائي ZSTD
1	0.87	1.65	2.1	0.56	0.34	3.30-	0.58	1.43	1.50
2	0.34	1.71	2.3	0.19	1.49	1.70	0.05-	1.36	1.30
3	0.68	1.25	0.90	0.58	0.65	1.40-	0.42	0.92	0.20-
4	0.61	1.57	1.90	0.20	1.29	1.10	0.85	1.99	0.30
5	1.06	2.44	4.10	0.68	0.71	1.10-	0.30	0.73	1.00-
6	0.31	2.17	3.50	0.20	0.93	0.20-	0.23	1.29	1.10
7	0.98	2.85	4.80	0.81	2.96	5.10	0.19	0.66	1.40-
8	0.72	2.33	3.80	0.64	1.67	2.20	0.36	1.59	2.00
9	0.85	1.61	2.00	0.69	0.78	0.80-	0.04	1.02	0.10
10	1.05	1.26	1.00	0.36	1.26	1.00	0.36	1.26	1.00
11	1.11	2.11	3.30	0.69	0.61	1.70-	0.09	0.74	1.00-
12	0.97	1.53	1.80	0.73	1.14	0.50	0.41	0.59	1.70-
13	0.34	1.66	2.20	0.18	1.81	2.60	0.27	0.56	1.90-
14	1.14	1.59	2.00	1.18	2.51	4.20	0.15	0.40	2.90-
15	0.2-	1.27	1.00	0.23	2.84	4.90	0.09	2.13	3.40
الوسط الحسابي	0.72	1.80	2.50	0.56	1.40	1.00	0.29	1.11	0.20

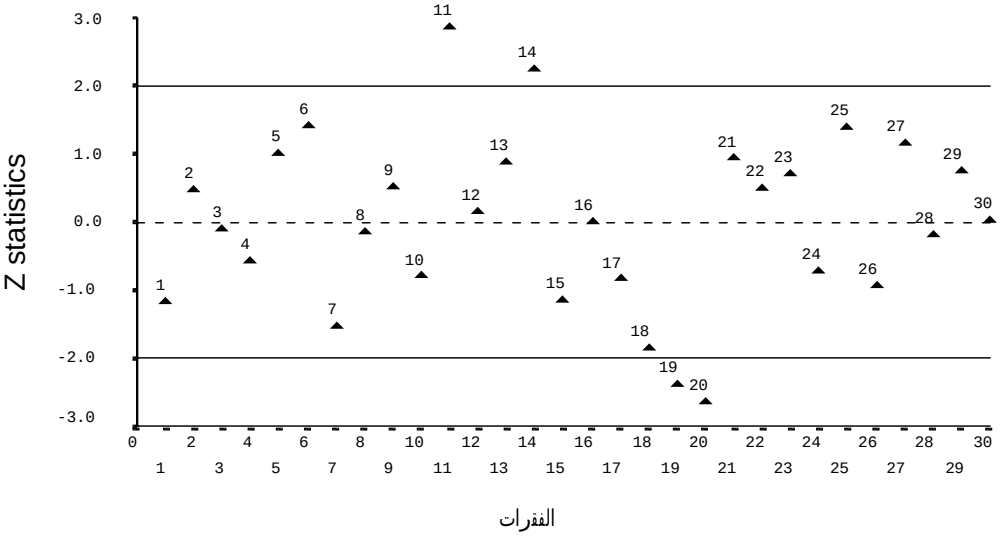
يتضح من النتائج الواردة في الجدول (4)، بان قيمة الوسط الحسابي لاوساط مربعات الأخطاء (MSE_n) قد ابتعدت عن الواحد الصحيح في الجولتين الأولى والثانية حيث بلغت قيمته للجولتين على التوالي 1.8، 1.4، في حين اقتربت قيمته في الجولة الثالثة من الواحد حيث بلغت 1.11. أي ابتعدت تقديرات المحكمين عن توقعات نموذج المحاولات ذات الحدين في الجولتين الأولى والثانية، في حين اقتربت من الوضع المثالي في الجولة الثالثة كما يتوقعه النموذج.

وبتفحص قيم إحصائيي المطابقة يتضح أن تقديرات 12 محكما للحد الأدنى من الكفاية تبتعد عن توقعات نموذج المحاولات ذات الحدين في الجولة الأولى، حيث كانت قيم MSE_n وZSTD لهم خارج الحدود المقبولة. وفي كل من الجولتين الثانية والثالثة ابتعدت تقديرات 6 محكمين عن توقعات النموذج. أي ان عدد المحكمين الذين تطابقت تقديراتهم مع توقعات النموذج في الجولتين الثانية والثالثة ارتفع نسبياً، وذلك لأنه أتيح لهم فرص إعادة النظر بتقديراتهم الأولية التي تمت في الجولة الأولى.

ولعمل مخططات نوعية لكل محكم، تم تعيين قيمة الإحصائي ZSTD المناظرة لكل فقرة من فقرات الاختبار، فإذا كانت قيمة ZSTD اكبر من صفر، تكون نظرة المحكم لهذه الفقرة أنها فقرة صعبة، أما إذا كانت اقل من صفر فينظر إلى الفقرة على أنها سهلة. فعلى سبيل المثال يظهر الشكلان (1)، (2) رسماً توضيحياً لمحكمين هما: المحكم 4 والمحكم 8 أحدهما لم تتطابق تقديراته مع توقعات النموذج، والاخر تطابقت تقديراته مع توقعات النموذج.



الشكل (1): رسم توضيحي لقيم ZSTD المناظرة لتقديرات المحكم 4 في الجولة الثالثة (MSE=1.99)



الشكل (2): رسم توضيحي لقيم ZSTD المناظرة لتقديرات المحكم 8 في الجولة الثالثة (MSE=1.02)

يتضح من الشكل (1)، أنّ تقديرات المحكم 4 لصعوبة الفقرات 1، 2، 4، 6، 7، 23، 26 لم تتطابق مع توقعات نموذج المحاولات ذات الحدين، إذ كانت قيمة الإحصائي ZSTD المناظرة لهذه التقديرات خارج فترة الثقة ± 2. وبالنظر إلى تقديرات المحكم 8 من الشكل (2)، نجد أن تقديراته للفقرات 11، 14، 19، 20 لم تتطابق مع توقعات النموذج، في حين جاءت جميع تقديراته لبقية فقرات الاختبار متطابقة مع توقعات النموذج مما جعل تقديره للحد الأدنى من الكفاية متطابقاً مع النموذج.

وقد أعيد تقدير درجات القطع لنماذج الاختبار، بعد استبعاد تقديرات المحكمين التي لم تف بمحكات المطابقة لنموذج المحاولات ذات الحدين (BTM). كما تم إعادة حساب ثبات تصنيف الطلبة باستخدام درجات القطع المقدرة من المحكمين الذين تطابقت تقديراتهم لدرجات القطع مع توقعات نموذج المحاولات ذات الحدين.ويلخص الجدول (5) تصنيفات الطلبة في ضوء درجات القطع المقدرة لكل نموذج بعد حذف المحكمين غير المطابقين لنموذج المحاولات ذات الحدين في مرّتي التطبيق ومعاملات الثبات للتصنيف.

الجدول (5): تصنيفات الطلبة المختلفة ودرجات القطع للاختبار ومعامل كابا ومعامل دقة القرار بعد حذف تقديرات المحكمين غير المطابقة لنموذج المحاولات ذات الحدين

الجولة التحكيمية	عدد المحكمين المطابقين للنموذج	درجة القطع %	عدد المتمكنين في التطبيقين	متمكن في الأول وغير متمكن في الثاني	متمكن في الثاني وغير متمكن في الأول	غير متمكن في التطبيقين	معامل كابا	معامل دقة القرار
الأولى	3	62	50	3	2	55	0.91	0.96
الثانية	9	62	44	4	3	59	0.88	0.94
الثالثة	9	66	55	2	2	51	0.93	0.96

يتضح من الجدول (5)، أن ثبات التصنيف ممثلا بمعاملي كابا ودقة القرار مرتفع نسبيا وهو أفضل مما كان عليه قبل فحص التقديرات باستخدام نموذج المحاولات ذات الحدين؛ إذ تراوحت قيم معامل دقة القرار بين 0.94 و0.96 وتراوحت قيم معامل كابا بين 0.88 و0.93 على مدار الجولات الثلاث. أي أن قيم معامل كابا ومعامل دقة القرار للجولات الثلاث متقاربة؛ إذ لم يتعد أعلى فرق في معامل دقة القرار 0.02 وفي معامل كابا 0.05. كما أن الفروق بين درجات القطع للجولات الثلاث قد تناقصت عما كانت عليه قبل فحص التقديرات؛ إذ يبلغ أكبر فرق بينها 6 درجات. وقد يُعزى التحسن في ثبات التصنيف واستقراره في الجولات الثلاث بعد أن تم حذف المحكمين الذين خالفت تقديراتهم بشكل جوهري التقديرات المتوقعة لنموذج المحاولات ذات الحدين، إلى فاعلية النموذج في فحص تقديرات المحكمين لدرجات القطع واستبعاد التقديرات الشاذة من بينها. وقد تعود التقديرات الشاذة لدى بعض المحكمين إما لعدم وضوح مفهوم الطالب الذي يملك الحد الأدنى من الكفاية في ذهن المحكم، أو لعدم كفايته في التعامل مع أسلوب أنجوف المستخدم في تحديد مستوى الأداء، أو ضعف خبرته ومعرفته بالمحتوى الذي يقيسه الاختبار. ويمكن كذلك أن تبرز أخطاء في التقديرات نتيجة أثر الهالة والانحدار الاحصائي والميل نحو التساهل أو التشدد في التقديرات، وكذلك العوامل الذاتية والشخصية للمحكم من مثل اهتمامه وجديته أثناء قيامه بعملية التقدير التي تنعكس بالتالي على تقديراته.

وقد أشارت نتائج الدراسات السابقة (Berk, 1996; Buckendahl, Smith, Impara & Plake, 2000; Chinn & Hertz, 2002) إلى أن تقديرات المحكمين لدرجات القطع تميل إلى الاتساق في الجولتين الثانية أو الثالثة، خاصة عندما يزود المحكمون بمعاملات الصعوبة الفعلية لفقرات الاختبار المراد تقدير درجة القطع له، والذي يشير بالتالي إلى أثر التغذية الراجعة أو المعلومات على تقديرات المحكمين. وبالفعل فقد كانت تقديراتهم لصعوبة فقرات الاختبار أكثر استقراراً واتساقاً في الجولة الثالثة لاطلاعهم على الصعوبة الفعلية لفقرات الاختبار في هذه الجولة.

وبالرجوع إلى النتائج المتعلقة بفحص جودة تقديرات المحكمين للحد الأدنى من الكفاية باستخدام نموذج المحاولات ذات الحدين لفقرات الاختبار، أشارت النتائج إلى أن عدد المحكمين الذين استبعدت تقديراتهم في الجولة الأولى، بلغ اثني عشر محكما في حين استبعدت تقديرات ستة محكمين في كل من الجولتين الثانية والثالثة. وجاءت هذه النتيجة متسقة مع ما هو متوقع من تزايد عدد المحكمين الذين تطابقت تقديراتهم الملاحظة مع توقعات النموذج مع مرور الجولات التحكيمية. وبينت النتائج أن تقديرات المحكمين لدرجة القطع كانت اكثر اتساقاً في الجولة الثالثة عندما زود المحكمون بمعاملات الصعوبة الفعلية لفقرات النموذج وهي بذلك تتفق مع نتيجة دراسة انجلهارد واندرسون (Engelhard & Anderson, 1998). وقد يعود السبب في زيادة عدد المحكمين الذين تطابقت تقديراتهم مع توقعات النموذج في الجولتين الثانية والثالثة، إلى أن المحكم كانت لديه الفرصة في إعادة النظر في تقديراته الأولية في ضوء المعلومات التي زود بها أثناء القيام بعملية التقدير في الجولتين الثانية والثالثة.

إن ارتفاع مؤشرات ثبات التصنيف بعد حذف تقديرات المحكمين الذين لم تتطابق تقديراتهم الملاحظة لتوقعات النموذج، يدعم فاعلية نموذج المحاولات ذات الحدين في فحص مدى ملاءمة تقديرات درجات القطع خاصة وأنه يعتمد على نماذج النظرية الحديثة في القياس. فقد أكد كين (Kane, 1987) أهمية استخدام النظرية الحديثة في تحديد درجات القطع. ويرى الباحثان كذلك أن من الأسباب والعوامل التي أدت إلى ارتفاع مؤشرات ثبات التصنيف أيضا هي انخفاض درجات القطع الناتجة عن استخدام أسلوب أنجوف بعد استبعاد تقديرات المحكمين، التي لم تف بمحكات المطابقة للنموذج واقترابها من الوسط الحسابي للعلامات، وقد يكون من أهم الأسباب التي تفسر تحسن الثبات هو انخفاض درجة القطع ونزوعها نحو الوسط الحسابي، خاصة أن ثبات التصنيف يتأثر بدرجة القطع، فكلما اقتربت درجة القطع من الوسط الحسابي ازداد ثبات التصنيف.

وبالرغم من اقتصار الدراسة على عينة عشوائية من طلبة الصف العاشر في ست مدارس وعدد محدود من معلمي الرياضيات للصف العاشر في محافظة عجلون عملوا كمحكمين واقتصار الدراسة على اختبار واحد في موضوع الهندسة التحليلية، يبدو في ضوء النتائج التي توصلت إليها الدراسة والمتمثلة بتقارب درجات القطع وثبات التصنيف في الجولات الثلاث، أنه يمكن اختصارعدد الجولات التحكيمية التي اقترح هامبيلتون (Hamblton cited in

زيادتها من أجل الوصول إلى (Keller & Zaanetti, 2000 تقديرات دقيقة، والاكتفاء بجولة واحدة شريطة اخضاع التقديرات الناتجة منها للفحص باستخدام نموذج المحاولات ذات الحدين، وعندها تحسب درجة القطع للاختبار من خلال التقديرات المتوافقة مع توقعات النموذج. وفي ذلك توفير لوقت المحكمين، ولوقت الباحث حيث لا يحتاج إلى توفير بيانات سابقة للمحكمين عن فقرات الاختبار. لكنه في هذا السياق ينبغي زيادة عدد المحكمين نظرا لأن نسبة عالية من التقديرات في الجولة الأولى قد لا تتطابق مع نموذج المحاولات ذات الحدين.

المصادر والمراجع

الشريم، احمد. (2003). **دراسة مقارنة لنموذج "أنجوف" ونموذج "نيدلسكي" في تحديد درجة القطع لاختبار محكي المرجع في الرياضيات**. رسالة ماجستير غير منشورة، جامعة اليرموك، الأردن.

علام، صلاح الدين. (2001). **الاختبارات التشخيصية مرجعية المحك في المجالات التربوية والنفسية والتدريبية** (ط.2). القاهره: دار الفكر العربي.

عودة، احمد. (1993). **القياس والتقويم في العملية التدريسية** (ط.2). اربد، الأردن: دار الأمل للنشر والتوزيع.

Behuniak, P., Archambault, F. & Gable, R. (1982). Angoff and Nedelsky standard setting procedures: Implications for validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42,* 247-256.

Berk, R. (1996). Standard Setting: The Next Generation (Where Few Psychometricians Have Gone Before !). *Applied Measurement in Education,9*, 215-235.

Brandon, P. (2002). Two versions of the contrasting-groups standard-setting method: A Review. *Measurement and Evaluation in Counseling and Development,35,* 167-181.

Buckendahl, C., Smith, R., Impara, J. & Plake, B. (2000). *A Comparison of the Angoff and Bookmark Standard Setting Methods.* Paper presented at the annual meeting of the Mid-Westen Educational Resarch Association in Chicago, IL.

Change, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education,12,* 151-166.

Chinn, R. & Hertz, N. (2002) . alternative approaches to standard setting for Licensing and Certification Examinations. *Applied Measurement in Education,15,* 1-14.

Cizek, G. (1993).Reconsidering standards and criteria. *Jounal of Educational Measurement, 30*, 93-106.

Cross, L., Impara, G., Frary, R. & Jager, R. (1984). A Comparison of three methods for establishing minimum standards on the National Teaching Examinations. *Journal of Educational Measurement, 21*, 113-129.

Feldt, L. S. (1988). *Design and Analysis of Experiments in the Behavioral Sciences.*Iowa City, Iowa:Iowa Testing Programs.

Engelhard, G. & Anderson, D. (1998). A Binomial Trials Model forexamining the ratings of standard- setting judges. *Applied Measurement in Education. 11,* 209-230.

Haladyna, T., Downing, S. & Rodriguze, M. (2002). A review of multiple-choice item- writing guidelines for classroom assessment. *Applied Measurement in Education,15,* 309-334.

Hays, W. L. (1980). *Statistics* (3rd edition). New York: Holt, Renehart and Winston.

Kane, M. (1987). On the use IRT models with judgmental standard setting procedures. *Journal of Educational Measurement, 24*, 333-345.

Keller, Lisa. & Zanetti, Marey. (2000). *Validity issues in standard setting.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Lee-Sing, A. (2000). Performance standard determination in the health professions: A comparison of judgmental-based methods and cluster-analytic techniques. *Dissertation Abstract International,60*, 3615-B.

Linacre, J. (2003). *A user's guide to FACETS Rasch-Model computer programs.*

Maurer, T. & Alexander, R. (1991). Methodlogical and psychometric issues in setting cutoff scores using the Angoff Method. *Personal Psychology,44,* 235-262.

Mills, C., Melican, G. & Alluwalia, N. (1991). Defining minimal competence.. *Educational Measurement: Issues And Practice. 10*, 7-10.

Plake, Barbara, & Impara, J. (2001). Ability of panelists to estimate item perfomance for target group of candidates: An issue in judgmental standard setting. *Educational Assessment,7,* 87-97.

Plake, Barbara, Impara, J. & Irwin, P. (1999). *Validation of Angoff-based predictions of item performance.* Paper presented at the 1999 annual meetin of the American Educational Research Asociation,Montreal, Canada.

Plake, Barbara, Melican, G. & Mills, C. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice. 10,* 15-17.

Popham, W. (2000). *Modern Educational Measurement: Practical guidelines for educational leaders* (3rd ed.). Boston: Allyn and Bacon.

Thorndike, R. (1997). *Measurement and evaluation in psychology and education* (6th ed.). New Jersey: An imprint of Prentice Hall.