المجلة الأردنية في العلوم التربوية، مجلد 6، عدد 3، 2010، 191 - 202

تطوير أسلوب قائم على الأهداف السلوكية لتحديد درجة القطع: دراسة مقارنة مع أسلوب أنجوف

أحمد عوده*، أحمد الشريم**

تاريخ تسلم البحث 2009/6/1 تاريخ قبوله 2010/7/20

ملخص: هدفت هذه الدراسة إلى تطوير أسلوب قائم على الأهداف لتقدير درجة القطع ، وتقييم هذا الأسلوب بمقارنته مع أسلوب أنجوف القائم على فقرات الاختبار. وقد تم تطوير قائمة بالأهداف السلوكية التفصيلية لوحدة تحليل المقادير الجبرية من منهاج الرياضيات، ويتبع كل هدف مجموعة من الفقرات التي تقيس هذا الهدف، وقد تم عرض قائمة الأهداف وقائمة الفقرات على المحكمين لتقدير درجة القطع لكل هدف، كما تم تطويرصورتين لاختبار محكي المرجع في الرياضيات يتكون كل منهما من ثلاثين فقرة من نوع الاختيار من متعدد تغطي القائمتين حسب درجة انسجام الفقرة بالهدف. وتكونت عينة الدراسة من ثلاثين محكماً، و(171) طالباً، وتم تقسيم المحكمين إلى ثلاث مجموعات وفق تصميم المجموعات المتساوية باختلاف المؤهل العلمي وطبيعة العمل، قامت المجموعات الثلاث بتقدير درجات القطع للأهداف، ثم قامت مجموعتان بتقدير درجة القطع للاختبار حسب طريقة أنجوف، وبعد مرور أسبوعين تم إعادة التحكيم مرة أخرى. أما عينة الطلبة فقد تم تطبيق الاختبار عليهم بعد إكمالهم دراسة الوحدة مباشرة لحساب المؤشرات الإحصائية للطلبة والفقرات (الصعوبة والتمييز). وأشارت نتائج الدراسة إلى أن درجة القطع للأهداف كانت (0.56) في جولتي التحكيم، بينما كانت للصورة الولى من الاختبار حسب اسلوب أنجوف (0.67) في الجولة الأولى و(0.66) في الجولة الثانية، وللصورة الثانية (0.63) في الجولتين. وقد كانت جميع معاملات الارتباط بين تقديرات المحكمين في جولتي التحكيم مرتفعة ودالة إحصائياً عند مستوى دلالة إحصائية 0.01، وبينت النتائج أنه لا يوجد فرق دال إحصائياً بين معاملي الثبات للنموذجين. وفيما يتعلق بثبات التصنيف للطلبة باستخدام درجتي القطع من نموذج الأهداف ونموذج أنجوف فقد بينت النتائج أن معدّلي الصواب للنموذجين كان مرتفعاً. وفيما يتعلق بالصدق التنبؤي للنموذجين فقد بينت النتائج أن قيمة معامل كابا لنموذج الأهداف كان الأعلى عند محك الجودة (0.55) حيث بلغ (0.95) بينما كان لنموذج أنجوف عند هذا المحك (0.65)، وكانت قيمة معامل كابا تتغير للنموذجين بتغير محك الجودة. وخلصت الدراسة إلى أن النموذج القائم على الأهداف يعد مناسباً لتقدير درجة القطع، وللتنبؤ بالتحصيل المدرسي للطلبة، لغايات التشخيص في مدى تحقيق الأهداف.(الكلمات المفتاحية: درجة القطع، نموذج الأهداف، نموذج أنجوف، اختبار محكي المرجع،التوافق بين الهدف والسؤال).

خلفية الدراسة: لقيت مسألة تحديد درجات القطع في الاختبارات بغرض تصنيف الطلبة حسب مستوى التمكن من الأهداف التعليمية اهتماماً كبيراً من المهتمين بالقياس والتقويم التربوي والنفسي، فهناك العديد من الحالات أو المواقف التي تتطلب استخدام درجات القطع لتقسيم الأفراد في مجموعتين أو أكثر وفقاً لمتغير معين أو توزيع معين لتحديد من سينجح في الاختبار ومن سيفشل، من سيحصل على إجازة أو ترخيص ومن لا يُرخّص، من سَيُقبل ومن

Developing an Approach Based on Behavioral Objectives to Estimate the Cut-off Score: A Comparison with Angoff's Method

Ahmad Audeh, Faculty of education, Yarmouk university, Irbid, jordan.
Ahmad Shraim, college of education, king saud university, riyadh, kingdom of Saudi Arabia.

Abstract: This study aimed at developing an approach based on behavioral objectives to estimate the cut-off score compared with Angoff's method, which determines the cut-off score based on test-items.In order to achieve this aim a content domain was selected, well defined, and analyzed by translation this content into a set of behavioral objectives. Every objective in the set was followed by a group of multiple-choice items, which measure this objective to enable the raters to estimate the cut-off score for every objective. Two test forms were constructed using two procedures of selecting items for each form based on item-objective congruence for one form, and random selection for the other one.A total group of 30 raters estimated the cut-score of each objective and item in the two test forms. The raters were classified according to their educational qualifications and work status into three equal groups. Two of the three groups estimated the cut-off scores using Angoff's method, two times or rounds over a period of two weeks. A sample of 171 ninth grade students were tested immediately after they had completed studying the unit in order to calculate the statistical indices of the students and the items, for validating the cut-off scores and for comparison purposes.The results of the study indicated that the cut-off score based on objectives was (0.56) in the two raters' rounds, while it was (0.67) for the first form using Angoff's method in the first round, (0.66) in the second round, and (0.63) for the second form in the two rounds. Furthermore, all correlation coefficients between the raters' estimations in the two rounds were high (0.93; 0.96) and were statistically significant at (α = 0.01). The results also indicated that there was no statistically significant difference between the reliability coefficients of the two methods (objective-based and test-item-based) of estimating cut-off scores.The results of the study also indicated that there was no statistically significant difference between the estimated cut-off scores by the raters in the three groups based on the behavioral objectives. The two methods were high and equivalent statistically of two reliability indices (Hit rate, and Kappa coefficient); but they were significantly different in their ability to predict a student's school achievement in favor of the objective-based method.
*The study concludes that the objective-based method is superior in estimating cut-off scores compared with the most frequently used test-item-based method; it is more efficient in predictive validity, more powerful and universal due to the strong-grounded early steps starting specifying the content domain, and the population of behavioral objectives representing this domain. (**Keywords**: cut–off score, objectives-based, Angoff's Method, test-item-based, criterion referenced test, item-objective congruence).*

* كلية التربية، جامعة اليرموك، اربد، الأردن.
** كلية التربية، جامعة الملك سعود، الرياض، السعودية.

سَيُرفض في وظيفة معينة، ومن يُسمى متمكناً ومن هو غير المتمكن. ويشير جودوين (Goodwin, 1996) إلى أنه قد تستخدم درجات القطع لتصنيف المجموعات والأفراد في عمليات التشخيص والتجارب الطبية طبقاً لموقعهم على متغير أو مجموعة متغيرات، وفي عيادات الطب النفسي بشكل خاص لتصنيف الأفراد إلى أيٍ الفئات التي ينتمون اليها، من أجل تخصيص برامج علاجية لهم بناءً على تصنيفهم ذلك، وهناك استخدامات أخرى لدرجات القطع مثل سوق العمل كالترشيح للوظائف، وفي معايير ضبط الجودة للصناعات والخدمات وغيرها.

ويخضع تحديد درجات القطع العالية أو المنخفضة في الاختبارات لاعتبارات عملية عديدة قد يكون أهمها النتائج المترتبة على أداء الخريج في سوق العمل، فقد يكون الحد الأدنى للمعدّل التراكمي للطالب في جامعة معينة حتى يعدّ ناجحاً في المعدّل العام مختلفاً باختلاف المرحلة كأن يكون 60% مثلاً لمرحلة البكالوريوس، و 75% لمرحلة الماجستير، و 80% لمرحلة الدكتوراه، وقد يكون هذا الاختلاف مرتبطاً بطبيعة العمل لكل خريج من هذه المراحل، فالطالب في مرحلة البكالوريوس لا يُفترض به أن يصل إلى مستوى الإتقان بتعريف إجرائي معين كما في مرحلة الدكتوراه، ومع ذلك فقد يتم بين الحين والآخر إعادة النظر في درجات القطع وفق أسس واعتبارات تحددها تلك الجامعة. ومن الأمثلة كذلك على درجات القطع الصفر الجامعي (اي الحد الأدنى للعلامة في المقرر) وارتباطه بالحد الأدنى للمعدل التراكمي للطالب، وكذلك مستوى النجاح لامتحان الكفاءة الجامعية لمن أنهى مرحلة البكالوريوس، والقضايا المتعلقة بتفسير نتائجه، إلاّ أن ما يستحق الإشارة إليه هنا هو أن القاعدة التي ترتكز عليها درجة القطع هي الأهداف بوصفها الخطوة الأساس في خطوات بناء الاختبارات والمقاييس، بصرف النظر عن درجة تحديد مجال(مجتمع) الفقرات أو الأسئلة وما يرتبط به من تمثيل لعينة الأسئلة المأخوذة من هذا المجال.

ويربط المختصون في القياس (Kane,1987; Shepard, 1984; Jager,1989; Goodwin, 1996) بين درجات القطع(cut-off scores) والمستويات(standards) عند الحديث عن المعايير، ونجد ذلك واضحاً في الحديث عن الاختبارات محكية المرجع مقابل الاختبارات معيارية المرجع وتحديد ما يتعلق بالتأكيد المتزايد على الحد الأدنى من الكفاية، ويتضح ذلك أيضاً في التطور الذي حصل على اختبارات التصنيف المحوسبة واختبارات التمكن والاختبارات المكيفة لأغراض الشهادات والتراخيص والتصنيف. حيث تعدّ درجة القطع مكوناً أساسياً من مكونات إعداد هذه الاختبارات، ويطلق على مستويات الأداء في الاختبارات محكية المرجع تسميات مختلفة مثل: درجات القطع Cutting Scores، ودرجات التمكن Mastery Scores، ودرجات النجاح Passing Scores (Cizek, 1996 ; Glass,1977).

وقد عرّف كين(Kane,1998) درجة القطع بأنها نقطة على مقياس العلامة الملاحظة Observed-Score Scale، بينما يعرف المستوى Standard بأنه حدٌ تصوري على مقياس العلامة الحقيقية True-Score Scaleبين الأداء المقبول وغير المقبول. وتعرفها شيبارد (Shepard,1984) بأنها الدرجة التي تميز بين المتمكن وغير المتمكن في القياس محكي المرجع، أو الدرجة التي تميز بين الجيد وغير الجيد وفق تعريف إجرائي معين وأسس منطقية تحددها الجهات المعنية. بينما يعرفها هامبيلتون (Hambleton, 1982) بأنها درجة على متصل السمة المقيسة، والتي تستخدم لتصنيف المفحوصين إلى فئتين تعكس مستويات الكفاءة بالنسبة لمجموعة من الأهداف أو المهارات التي يقيسها الاختبار، بحيث يمكن من خلالها الحكم على مستوى أداء المفحوص بأنه متمكن أو غير متمكن.

وباستعراض مختلف التعريفات لدرجة القطع أو مستويات الأداء نجدها متقاربة ومتسقة فيما بينها من حيث أن جميعها تتفق على تقسيم المفحوصين إلى قسمين على الأقل: قسم متمكن أو يمتلك المهارات والمعارف والموضوعات التي يقيسها الاختبار، والقسم الآخر غير متمكن ولا يمتلك تلك المهارات والمعارف التي يقيسها الاختبار، الذي يفترض بأنه يقيس بشكل غير مباشر عينة من الأهداف التي تعمل كمؤشرsign، أو ينطبق عليها خصائص العينة الممثلة sample في ضوء درجة تحديد مجال تلك الأهداف. ونظراً لإدراك المختصين في التربية لأهمية تحديد درجات القطع؛ اجتهد عدد كبير منهم في تطوير أساليب ونماذج لتقديرها يمكن حصرها في فئتين:

- الفئة الأولى المتمركزة حول الاختبار Test - Centered Methods مثل: طريقة نيدلسكي، وطريقة أنجوف، وطريقة إيبل، وطريقة جيجر.
- والفئة الثانية المتمركزة حول المفحوصين - Examinees Centered Methods مثل: طريقة المجموعة الحديّة Group Borderline؛ طريقة المجموعات المحكّية Criterion Group Method لبيرك؛ وطريقة المجموعات المتضادة Contrasting Groups؛ وطريقة هوفستي، (Sizmur,1997؛ Jager, 1989). وهناك طرق تجمع بين النوعين أو الفئتين تعتمد على أسس رياضية وإحصائية بحتة مثل النموذج ذي الحدين(Shepard, 1984).

وترتكز أغلب الطرق السابقة في تحديد درجة القطع على تقديرات المحكمين؛ لذا فلا بد أن يمتاز هؤلاء المحكمون حسب جيجر (Jager, 1989) (بالخبرة والمعرفة في مجال الاختبار، وأن يتفهم الخبراء أنماطاً واسعة من الممارسات في المجال المقصود، وتؤكد شيبارد (Shepard, 1984) أن المعلمين هم أفضل هؤلاء الخبراء في المنهاج والمحتوى التعليمي لأنهم أقرب الناس للمنهاج وأكثرهم تعاملاً معه، والقدرة على التحليل والمراقبة الذاتية، وإدراك التنوع في الممارسات ضمن المجال المقصود.

ويقترح كين (Kane,1987 ؛1998) مجموعة من الإجراءات التي تؤثر بشكل كبير في تقديرات درجة القطع أهمها: تحديد الغرض من عملية تقدير مستوى الأداء، واختيار المحكمين

المؤهلين وتدريبهم بشكل مناسب وخاصة على معنى مفهوم مستوى الأداء، ومفهوم الحد الأدنى من الكفاية عند الفرد المستهدف، وتوفير المعلومات والبيانات الكافية عن مجموعات المفحوصين والمجتمع المقصود. كما يشير كين (Kane,1987) إلى قضية أخرى متعلقة بصدق معايير الأداء حيث يمكن التحقق منه بطريقتين أساسيتين الأولى : بمقارنة تقديرات المحكمين لدرجة القطع أو مستويات الأداء مع النسب الحقيقية المستمدة من التجريب الواقعي لعمليات القياس والاختبارات، أما الثانية فيتم باختيار مجموعتين من المفحوصين المجموعة الأولى نجحوا بصعوبة والمجموعة الثانية فشلوا بصعوبة، ويتم مقارنة هاتين المجموعتين مع نسبة المفحوصين الذين أجابوا عن الفقرة إجابة صحيحة، فإذا كانت نسبة الذين أجابوا عن الفقرة من المجموعة الأولى إجابة صحيحة أكبر منها في المجموعة الثانية فإن هذا يعدّ دليلاً على صدق علامة القطع المقدرة من المحكمين.

وفي الوقت الذي تزداد فيه الحاجة لاستخدام درجات القطع في مختلف المواقف اليومية التي تهم مختلف القطاعات التربوية والمهنية وغيرها، وعلى وجه الخصوص اختبارات التمكن أو الاختبارات التصنيفية، يتزايد الحديث عن مدى صحة القرارات التي يتم اتخاذها في ضوء هذه الدرجات التي يعبر عنها عادة بنوعين من الخطأ هما الخطأ من النوع الأول الذي يتم فيه قبول أفراد غير متمكنين أو لا يمتلكون الحد الأدنى من الكفاية ويشار إليه بالخطأ الإيجابي، والخطأ من النوع الثاني وهو عندما يتم رفض قبول أفراد متمكنين أو يمتلكون الحد الأدنى من الكفاية ويشار إليه بالخطأ السلبي، والمؤسسات التعليمية المختلفة كالجامعات والمعاهد والمدارس المتخصصة والمؤسسات المهنية والصناعية من جهة، والأفراد الذين يترشحون للمنافسة للالتحاق بهذه المؤسسات من جهة أخرى يعانون من الآثار الاقتصادية والاجتماعية والنفسية الناجمة عن أخطاء التصنيف هذه. وما زلنا نلاحظ في كثير من الأوقات بعض الذين يوجهون اللوم إلى مثل هذه الاختبارات، وكذلك إلى المعايير التي تم استخدامها في تصنيف الفئات من المفحوصين، عند تطبيقها ميدانيا سواءً في المواقف التربوية أو القطاعات المهنية المختلفة، بالإضافة إلى عدد من القضايا والأسئلة التي ما زالت تبحث عن حلول وإجابات منها على سبيل المثال : مشكلة الطلبة الناجحين ولكنهم لا يستطيعون أداء العمل كما هو متوقع(Hambleton,1982).

وتشير بعض الدراسات إلى أن هناك العديد من القضايا المتعلقة بتحديد درجة القطع التي ما زالت بحاجة للبحث والدراسة في ضوء المحددات والافتراضات التي ترافق تطبيق الأساليب التي تعتمد فقرات الاختبار أو التي تعتمد على خصائص المفحوصين (Haertel, 2000)، وما يتعلق بقدرات المحكمين ومستوى تدريبهم وعددهم وخصائصهم الشخصية وجدوى تزويدهم ببيانات عن الفقرات والمفحوصين، وانعكاس ذلك كله على مؤشرات الصدق والثبات لتقديرات المحكمين في تحديد درجة القطع (Hambleton,1982). وبهذا الصدد فقد أشارت بعض الدراسات (Diane, et al., 2005؛Kaufman, et al., 2000) إلى أن طريقة أنجوف في الغالب تعطي تقديرات أعلى عند مقارنتها بغيرها من الطرق الأخرى.

وبينت دراسة جورج وسيد وأويبود (George, Sayeed & Oyebode, 2006) عند مقارنة طريقتين لتقدير درجة القطع إحداها تعتمد على تقديرات المحكمين (طريقة أنجوف)، والأخرى تعتمد على الأساليب الإحصائية لخريجي كلية الطب أن طبيعة الاختبار قد أثرت على آراء المحكمين (كونه اختباراً لخريجين من كلية الطب اعتقد المحكمون أن الطبيب يجب أن يصل إلى مستوى إتقان قريب من 100%) الأمر الذي جعلهم يقدّرون درجة قطع مرتفعة حسب طريقة أنجوف.

وتبين بورسيكوت (Boursicot, 2006) التي قامت بدراسة لمتابعة الطلبة الخريجين من كلية الطب في الميدان، والذين تم اعتبارهم مؤهلين بالحد الأدنى المقبول لممارسة مهنة الطب بناءً على تقييمٍ أساسهُ الأداء في مستوى التخرج من الكلية؛ أن هناك اتساقاً مقبولاً في المعرفة والمهارات العيادية، بينما هناك اتساق أقل بكثير في استخدام الوصفات ومهارات الاتصال والمهارات الشخصية، وتعزو بورسيكوت السبب في ذلك إلى أن الاختبار الذي تعده الكلية لم يكن يشتمل على المهارات الأخيرة (الاتصال والمهارات الشخصية واستخدام الوصفات)، وتؤكد في دراستها أهميةٍ اشتقاق من المهام المتوقعة للخريج؛ مما يعني أن علامة القطع أو علامة النجاح في الامتحان مضللة طالما أن الاختبار لا يغطي المعرفة والمهارات المتوقعة.

وفي دراسة قام بها شانج (Chang, 2000) لمعرفة هل المعيار أو درجة القطع تعكس الحد الأدنى لكفاية المفحوصين أم كفاية المحكمين ؟ فقد استنتج أن المحكمين يميلون لأن يعطوا تقديراً أعلى للفقرات التي يجيبون عنها إجابة صحيحة من الفقرات التي يجيبون عنها إجابة خاطئة، وأشار إلى أن هناك ثلاثة عوامل تؤثر على كفاية الأحكام بشكل عام هي : خلفية المحكمين، والفقرات وطريقة بنائها، والإجراءات المستخدمة لتحديد درجات القطع.

ويرى هيرتز وشين (Hertz & Chinn,2002) أن التشاور بين أعضاء اللجان يمكن أن يؤدي إلى تقارب وجهات النظر، والوصول إلى إجماع في الرأي، وقد يزيل المعيقات الخاصة بقدرة بعض المحكمين، الأمر الذي سينعكس إيجاباً على الثقة بدرجة القطع الناتجة مما يعني اختلاف درجة القطع باختلاف إجراءات التحكيم، وأن هذا التشاور يمكن أن يقلل من مصادر الأخطاء أو يخفف من الاعتماد على ما يقيسه الاختبار والانتقال إلى ما يجب أن يقيسه.

في ضوء المحددات التي أشارت إليها الدراسات السابقة بصورة ضمنية أو صريحة حول مقصود الاختبار وتأجيل تحديد درجة القطع إلى مرحلة متأخرة فإن هذه الدراسة تأتي للخروج من إطار الاعتماد كلياً على الاختبار وفقراته وعلى خصائص المفحوصين، وذلك باقتراح أسلوب لتقدير درجة القطع قائم على الأهداف السلوكية التي تمثل المجال والمحتوى المقصود لأنها

أي(الأهداف) تعد الأصل في أي عملية قياس للمعارف والمهارات، وبالتالي يتوقع أن يخفف هذا الأسلوب من سلبيات الاعتماد على فقرات الاختبار، أو مجموعات المفحوصين، وقد ظهرت بعض ملامح التخلص من هذه السلبيات في الدراسة التي قام بها هيرتل (Haertel, 2000) للتقليل من الارتكاز على فقرات الاختبار في تحديد درجة القطع فقد اقترح طريقة لتقدير الحد الأدنى من الكفاية لاعتبار الطالب متمكناً بالاعتماد على تقديرات المحكمين في تحديد المهمات التي على الطلبة أداؤها وإكمالها بنجاح لاعتبارهم متمكنين، ومن هنا أطلق على هذه الطريقة "طريقة المهمة القياسية" Benchmark Task Procedure، وتتلخص هذه الطريقة في قيام مجموعة من المختصين التربويين وأصحاب العلاقة وأرباب العمل بتحديد جميع المهام الرئيسة التي لو أخذت مجتمعة يمكن أن تعدّ ممثلة للمجال المقصود، ثم إعداد إرشادات لتقدير الأداء المقبول (المقنع) مقابل الأداء غير المقبول على كل مهمة من هذه المهام، وأخيراً يتم تقدير نسبة المهمات التي يجب أن تكمل بنجاح، واحتمال أكمال كل مهمة لوحدها من قبل الطالب الذي يمتلك الحد الأدنى من الكفاية.

ومن الخصائص المميزة لهذا الإجراء المقترح أنه يحوّل أو ينقل أحكام المحكمين من النجاح على فقرات الاختبار متعددة الاختيارات إلى النجاح في الأداء الملموس بمعنى النجاح في إكمال مهمات محددة بعناية تمثل المجال المقصود. وهذه إشارة ضمنية إلى استبدال المشكلة الأكبر في تعريف الإتقان فيما يتعلق بمجال النتيجة الذي يكتنفه الغموض في الغالب في العديد من القضايا الأصغر والتي يسهل تعريفها، وهي توضّح تعريف المجال باختيار مجموعة المهام القياسية Benchmark Tasks، ثم تعرف النجاح على كل مهمة قياسية، وأخيراً تقرير نسبة المهام التي يجب أن تكمل بنجاح، وهذا يتطلب من المحكمين استحضار كل معرفتهم العملية وخبراتهم أثناء العمل. ويرافق هذه العملية اختبار الحد الأدنى للكفاية(MCT) Minimum Competency Testing للطلبة، وتقديرات المحكمين فيما يتعلق بالنسبة لأداء كل مهمة لإظهار العلاقة التجريبية بين علامات الطلبة على الإختبار واحتمال إكمال المهمة بنجاح.

ومن الواضح ان تحديد مستوى الأداء يستلزم القدرة المطلوبة على التمييز بين السلوك المعرفي والمهارة في كتابة الفقرات في المستويات المعرفية العليا خصوصاً التطبيق والتحليل (Haswell, 2003). ففي كثير من الأحيان نجد أن الاختبارات تُكتَب فقراتها في المستويات الدنيا، والتي تسمى مستوى الاستظهار Memorization Level، وعلى النقيض من ذلك إن الأغلبية الواسعة من الوظائف تتطلب أداءً فوق مستوى الاستظهار، أي مهارات التحليل والتركيب والتقويم في التصنيف السداسي التقليدي لبلوم في المجال المعرفي، وهذا يفصل بين ممارسة الاختبار والأداء (ممارسة العمل) الذي يدعو الإدارات في أغلب الأحيان إلى التساؤل حول قيمة التدريب، وربما تحوّل الاختبار إلى مؤشر مضلّل للأداء الحقيقي او الواقعي، ومثال ذلك السؤال الذي يُطرح كثيراً: كيف ينجح المتدربون في مستوى معين؛ وبالمقابل لا يستطيعون إنجاز العمل عند مباشرتهم مهام واقعية ؟(Coscarelli & Shrock, 2006) .

في ضوء هذا التصور يقترح الباحثان حلاً لهذه المشكلة يتلخص في العودة إلى الجذر أو الأساس في العملية التربوية وهي الإنطلاق من الأهداف التدريسية، حيث تشكل الأهداف محور النشاط في هذه العملية، ويبذل التربويون جهداً كبيراً ووقتاً طويلاً في بلورة هذه الأهداف وصياغتها؛ ومن هنا فإنه عند تصميم الاختبار لا بدّ أن يؤخذ بالاعتبار الأهداف التي نريد تحقيقها كخطوة اولى، ثم تحديد المستوى المقبول للنجاح المحدد بدرجة القطع بناءً على هذه الأهداف، ومن ثم يتم تقييم التعلّم من خلال اختبار طُوّر بشكل أساسي في ضوء هذه الأهداف، أي أن الاختبار يجب أن يمثل الأهداف تمثيلاً كافياً، عند ذلك يمكن الوثوق بهذه الدرجة لأنها تعكس مستوى أداء حقيقي مشتق من الأهداف التي طُوّر المنهاج على أساسها، كما قد تعكس صورة مسبقة عن مستوى الأداء على الأهداف في مرحلة لاحقة عند قبوله على اعتباره متمكناً منها.

مشكلة الدراسة:إن المتأمل لمختلف الطرق الشائعة الاستخدام في تحديد درجة القطع، المشار اليها في أدبيات القياس، والدراسات التي تناولت معظم هذه الطرق لتقييمها وتجريبها ومقارنتها، يجد أنها تنقسم إلى قسمين رئيسين: الأول يعتمد فقرات الاختبار، وهو الأكثر استخداماً، والثاني يعتمد مجموعات المفحوصين، وقد لاحظ الباحثان وجود عددٍ من القضايا المهمة التي تؤثر على فاعلية استخدام هذه الطرق، وعلى طبيعة القرارات الناتجة عن استخدامها، منها ما يلي:

اولا: الطرق التي ترتكز على فقرات الاختبار: هناك مجموعة من القضايا المهمة التي تؤثر على قرارات التصنيف (تمكّن / عدم تمكّن)، وعلى كفاءة الاختبارات في عمليات التقييم للبرامج والعمليات التعليمية والتدريبية، وكذلك على قدرة الاختبارات التنبؤية أهمها ما يلي : إن هذه الطرق جميعها تعتمد على تقديرات وأحكام المحكمين، وأن الشيء الأساسي الذي يوضع بين يدي المحكّم هو فقرات الاختبار، وبالتالي فإن صياغة فقرة الاختبار تلعب دوراً أساسيا في طبيعة القرار الذي يتخذه المحكم؛ فدرجة القطع ترتبط بسهولة الفقرات أو صعوبتها بالنسبة للفرد الذي يمتلك الحد الأدنى من الكفاية في موضوع الاختبار، وبالتالي فإن درجة القطع تعكس مواصفات الاختبار نفسه؛ ولذلك فان الأهداف هي الأصل او المرجعية في تحديد درجة القطع.

ثانيا: الطرق التي تعتمد مجموعات المفحوصين: إن هذه الطرق تعتمد بشكل أساسي على المعلومات القبلية عن توزيعات مستويات المجموعات (المجموعات التي تلقت التعليم، والمجموعات التي لم تتلق التعليم)، وهذا قد يتعذّر في كثير من الأحيان، ويتعارض في بعض الأحيان مع طبيعة الاختبارات وخصائصها مثل: سرية الاختبار، وإمكانية تطبيقه على مجموعات تجريبية، كما أن هذه المجموعات متغيرة في خصائصها، وبالتالي

فان كل موقف اختباري يحتاج إلى تقديرات جديدة خاصة به تتناسب مع خصائص المجموعة لذلك الموقف، وبالتالي يتوقع تغير أخطاء التصنيف تغيراً جوهرياً ومتناسباً مع التغير في خصائص هذه المجموعات.

إن صدق المعيار أو تقديرات المحكمين لدرجة القطع مرتبط بصدق محتوى الاختبار، وصدق الاختبار هو دائماً موضع تساؤل، إذ أن مؤشرات الصدق للاختبار هي مؤشرات غير مباشرة،فهي على الأغلب مشتقة من آراء محكمين يعرض عليهم الاختبار مع جدول المواصفات ويطلب منهم تقدير صدق المحتوى للاختبار، وبالتالي سيكون صدق المعيار هو موضع تساؤل أيضاً. كما يعتمد رأي المحكم على خصائص البدائل في فقرة الاختبار من نوع الاختيار من متعدد، وكلما كانت هذه البدائل(المموهات) جذابة كانت الفقرة بنظر المحكم أصعب، ولو افترضنا أن نفس الفقرة عُرضت على المحكم ولكن البدائل لها تغيرت وكانت أقل جاذبية للمفحوص فسوف يعتبرها المحكم فقرة سهلة، وبالتالي فان قراره سيختلف في الحالتين مع أن محتوى الفقرة لم يتغير (الشيء الذي تقيسه الفقرة نفسه في الحالتين).

إن معظم الطرق في هذه الفئة لا يُناسب إلا الفقرات من نوع الاختيار من متعدد، ومع ما تتميز به هذه الفقرات من سهولة في التطبيق والتصحيح وموضوعية وثبات؛ إلاّ أنها لا تصلح في كثير من الأحيان لقياس بعض الأهداف التعليمية من المستويات العليا مثل التركيب والتحليل والتقويم والإبداع، وهناك مواقف تربوية كثيرة يتطلب أن تكون فقرات من أنواع أخرى من الفقرات مما يترتب عليه صعوبة تطبيق أغلب تلك الطرق لتحديد درجة القطع.

قضية أخرى - قد تكون هي الأهم - إن فقرات الاختبار تتغير في عددها وخصائصها من موقف لآخر ومن مجموعة لأخرى، وفي هذه الحالة يتوقع أن تتغير جوهرياً تبعاً لذلك درجة القطع للاختبار، فعندما يرتبط تحديد درجة القطع باختبار معين؛ فإن أي تغيير في درجة الاختبار يتطلب العودة إلى كل الإجراءات لتحديدها من جديد، وهذا يحتاج إلى كثير من الجهد والوقت وخاصة فيما يتعلق بإجراءات التحكيم وتحديد قائمة المحكمين. وتزداد هذه المشكلة تعقيداً عندما يتم حوسبة الاختبارات وما يرافقه من تفاعل بين خصائص المفحوص وخصائص الفقرات، مما يعني حوسبة درجة القطع أيضاً، أي عرض العدد الكبير من الفقرات في بنك الفقرات على المحكمين، وهذه الفقرات ستطبق على مجموعات قد لا تكون معروفة مسبقاً للمحكمين، بالإضافة إلى عدد الفقرات المتجدد والمتغير في بنك الفقرات وكل ذك يشكل صعوبة بالغة.

إن كل ما تقدم دعا الباحثين إلى التفكير بأسلوب جديد لتحديد درجة القطع لمحتوى تعليمي معين يعالج مختلف القضايا السابقة وهو (تحديد درجة القطع اعتماداً على الأهداف). ويرتكز هذا الأسلوب على تقديم قائمة الأهداف التي تغطي المجال السلوكي إلى المحكمين ويُطلب من المحكم تحديد نسبة الطلبة الذين يتوقع أن يكونوا قد حققوا الأهداف ذات الصلة ببرنامج معين أو بمهام معينة وفقاً لطبيعة الغرض الذي يحكم درجة القطع، والقرارات المترتبة عليها كان يكون الغرض هو الحكم على مدى نجاح برنامج تدريبي معين، وفي هذه الحالة تكون الأهداف مشتقة من المنهاج التعليمي، ومُمثِّلة له تمثيلاً كافياً، أو يُطلب من المحكم تقدير نسبة الطلبة الذين يتوقع أن يحققوا هذه الأهداف إذا كان الغرض من عملية القياس هو التنبؤ، كالقبول بتخصصٍ ما، أو برنامجٍ مهني معين، أو لممارسة مهنه بعد التخرج في سوق العمل، وتكون الأهداف في هذه الحالة مشتقة من طبيعة المرحلة القادمة. ولذلك تكمن أهمية هذه الدراسة باقتراح أسلوب جديد لتحديد درجة القطع يقوم بشكل أساسي على تقديرات المحكمين ولكن للأهداف بدلاً من الفقرات، وإلى أي مدى يمكن تحقيقها، وبالتالي مهما تغيرت فقرات الاختبار فإن المعيار للأداء سيبقى ثابتاً، لأنه مرتبط بقاعدة ثابتة ولا تتغير مع تغير الفقرات والمفحوصين وهي الأهداف.

أسئلة الدراسة: على الرغم من ارتباط الأهداف بالقياسات محكية المرجع وبالتالي ارتباطها بتحديد مستويات الأداء (درجات القطع)، حتى أنه يُطلق على الاختبارات محكية المرجع بالاختبارات هدفية المرجع Objective Referenced Tests والتي تستخدم في قياس مجموعة من الأهداف التعليمية المصاغة بطريقة إجرائية قابلة للقياس، وعلى الرغم من تعدد الأساليب في تحديد درجات القطع؛ إلاّ أنه لا يوجد من بين تلك الأساليب أسلوب يُحدّد درجة القطع بالاعتماد على الأهداف التي تمثل المحتوى التعليمي أو المجال السلوكي لموضوع القياس، ومن ثم إجراء تقييم شامل لهذا الأسلوب من خلال تطبيقه وتجريبه بشكل واقعي، وتقدير مؤشرات الثبات والصدق والدقة له، وكذلك تقييم سهولة الإجراءات ووضوح الخطوات، وذلك من خلال مقارنته مع أحد الأساليب التي ترتكز على فقرات الاختبار في تحديد درجة القطع وشائع الاستخدام وهو أسلوب أنجوف، بالإضافة إلى الكشف عن أثر بعض المتغيرات التي يمكن أن تؤثر على دقة التقديرات مثل: معرفة توزيعات الطلبة الواقعية، وكذلك خصائص المحكمين.

وبالتحديد تسعى الدراسة للإجابة عن الأسئلة البحثية التالية:

1. هل يختلف معامل ثبات تقديرات المحكمين لدرجات القطع باستخدام النموذج القائم على الأهداف عن معامل الثبات لنموذج أنجوف القائم على فقرات الاختبار اختلافا احصائيا على مستوى 05. ؟
2. هل تؤثر خصائص المحكمين على تقديراتهم لدرجة القطع بالارتكاز على الأهداف السلوكية ؟
3. ما قيمة معامل ثبات التصنيف للمفحوصين إلى متمكنين وغير متمكنين باستخدام درجة القطع الناتجة من النموذج القائم على الأهداف، ودرجة القطع الناتجة من نموذج أنجوف القائمة على الإختبار؟
4. ما قيمة معامل كابا للنموذج القائم على الأهداف ونموذج أنجوف للتنبؤ بمستويات التحصيل للطلبة في الرياضيات عند أكثر من محك للجودة ؟

عينات الدراسة: يتضح من الأسئلة أن هناك محكمين ومفحوصين ولذلك تكونت عينات الدراسة من قسمين : عينة المحكمين، وعينة الطلبة. وتكونت عينة المحكمين من (30) محكماً ومحكمة تم تقسيمهم إلى ثلاث مجموعات على النحو التالي:

1. مجموعة التأليف، مكونة من عشرة محكمين ممن شاركوا في تأليف منهاج الرياضيات في وزارة التربية والتعليم.
2. مجموعة الخبراء، مكونة من عشرة محكمين ممن يحملون مؤهل دكتوراه أو ماجستير في أساليب تدريس الرياضيات ولديهم خبرة في تدريس الصف التاسع الأساسي.
3. مجموعة المعلمين، وهم المعلمون الذين يحملون درجة البكالوريوس في الرياضيات ويدرّسون الصف التاسع لمدة لا تقل عن سنتين. وقد تم اختيار المجموعتين الثانية والثالثة بعناية من قبل الباحثين وبالتعاون مع المشرف التربوي لمبحث الرياضيات وعدد من المعلمين الذين عُرف عنهم التميّز في الأداء والتعاون.

أما عينة الطلبة فقد تكونت من (171) طالباً تم اختيارهم من المدارس التابعة لإحدى مديريات التربية والتعليم في الأردن لتنفيذ إجراءات تحديد درجة القطع وفق اسلوب انجوف القائم على الإختبار .

أدوات الدراسة:

أولاً: قائمة الأهداف السلوكية: يشكل تحديد النطاق السلوكي للأهداف نقطة الأساس في تحديد درجة القطع بهذا الأسلوب ولذلك تم اختيار وحدة تحليل المقادير الجبرية من منهاج الرياضيات للصف التاسع الأساسي، حيث تم تحليلها تحليلاً دقيقاً ومفصلاً، ثم صياغة الأهداف العامة للوحدة بالاستعانة بدليل المعلم والكتاب المدرسي ومعلمي المبحث من أصحاب الخبرة، كما تم اشتقاق جميع الأهداف السلوكية الممكنة التي تشتمل عليها الوحدة. وللتأكد من صياغة الأهداف، وتحديدها، وشمولها للوحدة الدراسية، وتمثيلها للمستويات المعرفية المختلفة (التذكر، والاستيعاب، والتطبيق، والتحليل، والتقويم، والإبداع) حسب تصنيف بلوم المعدّل للأهداف (Krathwohl, 2002) ، قام الباحثان بعرض قائمة الأهداف على مجموعة من المتخصّصين وأصحاب الخبرة في مجال أساليب تدريس الرياضيات والمعلمين في العينات المشار اليها لمراجعتها وإبداء الرأي فيها، وقد تم مناقشة ملاحظاتهم بالتفصيل وإجراء التعديلات المناسبة بناءً عليها، حيث تم الوصول إلى صيغة نهائية لقائمة مكونة من (30) هدفاً سلوكياً محدداً تحديداً دقيقاً بحيث يمثل كل هدف من هذه الأهداف جزءاً محدداً من الوحدة الدراسية، وتغطّي هذه الأهداف المجالَ السلوكي للوحدة الدراسية. وبعد إقرار جميع الأهداف السلوكية للوحدة قام الباحثان بالتعاون مع ثلاثة معلمين من أصحاب الخبرة في تدريس الرياضيات للصف التاسع بكتابة جميع الفقرات الممكنة التي تقيس كل هدف بشكل منفصل، حيث تم كتابة قائمة أولية كبيرة من الفقرات، وفي جلسة مشتركة بين الباحثين والمعلمين تم مراجعتها بعناية واستبعاد الفقرات غير المناسبة والمكررة والمتشابهة والاتفاق على قائمة نهائية مكونة من (108) فقرات موزعة على الأهداف الثلاثين وقد تراوح عدد الفقرات المحتملة لكل هدف بين ثلاث إلى سبع فقرات، وقد رُتّبت الأهداف في قائمة بحيث يقابل كل هدف مجموعة الفقرات التي تقيسه، وعُرضت على مجموعة من المعلمين والمشرفين التربويين من أصحاب الخبرة لمراجعتها من جديد لابداء الملاحظات على الفقرات من حيث مناسبتها للهدف، وصياغتها، والبدائل الصحيحة، والمموهات لكل فقرة، وأية ملاحظات أخرى يرونها مناسبة وفق الإرشادات الواردة في أدبيات القياس الخاصة بصياغة فقرات الاختبار من نوع الاختيار من متعدد، وقد قام الباحثان بدراسة ومناقشة جميع ملاحظات المحكمين وإجراء التعديلات المناسبة، ومن ثم التوصّل إلى الصيغة النهائية لقائمة الأهداف السلوكية ومجموعة الفقرات التي تقيس كل هدف، والتي تُكوّن أداة الدراسة الأولى. ويرى الباحثان أن مكونات أداة الدراسة الأولى على هذا النحو بحيث يُكتب الهدف بشكل مميّز ويتبعه مباشرة مجموعة الفقرات المحتملة التي تقيس هذا الهدف بالتحديد - فيه فائدة كبيرة للمحكّم وتسهّل عليه عملية التقدير لاحتمالات إتقان الهدف، على اعتبار أن مجموعة الفقرات التي تلي الهدف تشكل عينة من مجتمع الفقرات التي تقيس الهدف، أو المجال للهدف، وبالتالي يكون قد تَوفّر للمحكم مدى أوسع من المعلومات عن الهدف، الأمر الذي يتوقع أن ينعكس إيجابياً على دقة التقديرات وموضوعيتها؛ لأن مجموعة الفقرات المقابلة للهدف تشكل تعريفا اجرائيا لذلك الدف عند تقدير درجة القطع المرتكز على الأهداف، مما يعزز وضوح التعريف الإجرائي للهدف القائم على مدى قدرة المحكم على تحديد المستوى العقلي للهدف، ونوع العملية المعرفية من خلال ناتج التعلم والفعل الدال على السلوك الوارد في صياغة العبارة الهدفية.

ثانياً : الاختبار(الأداة الثانية في الدراسة): تطلبت إجراءات هذه الدراسة وجودصورتين للاختبار محكي المرجع يمثل الوحدة الدراسية المذكورة، ولذلك تم اعداد صورتين للاختبار بطريقتين مختلفتين :

الصورة الأولى: تتكون من (30) فقرة من نوع الاختيار من متعدد، ولكل فقرة أربعة بدائل، وبالرجوع إلى أداة الدراسة الأولى (قائمة الأهداف والفقرات التي تقيسها) تم اختيار الفقرات من هذه القائمة بشكل قصدي، وذلك باختيار الفقرة التي تعتبر أفضل فقرة تقيس الهدف Item-objective congruence بشكل مباشر (فقرة واحدة لكل هدف).

الصورة الثانية: تتكون كذلك من (30) فقرة من نوع الاختيار من متعدد، ولكل فقرة أربعة بدائل، وقد تم اختيار هذه الفقرات من هذه قائمة الفقرات – الأهداف بالطريقة العشوائية على مستوى كل هدف، ويعتبر هذا الأسلوب للاختبار أكثر واقعية في الظروف العادية لإعداد الاختبارات. حيث تم اختيار فقرة واحدة عشوائياً من بين مجموعة الفقرات التي تقيس كل هدف بشكل منفصل لتشكل مجموعة الفقرات الثلاثين صورة

الاختبار الثاني. وبذلك يكون هناك ثلاثة مواقف عملية لتحديد درجة القطع.

إجراءات جمع البيانات

أولاً : البيانات الخاصة بالطلبة: تطلّب تصميم هذه الدراسة أن يتم تطبيق أحد صورتي الاختبار على عينة من الطلبة، وقد تم اختيار الصورة الأولى لتطبيقها على عينة الدراسة من الطلبة المكونة من (171) طالباً، وقد تم إبلاغ معلمي الرياضيات في المدارس التي وقع عليها الاختيار لعينة الدراسة ليتم تطبيق الاختبار على الطلبة في موعد التقويم المدرسي الأول الذي يصادف في بداية الشهر الثاني من الفصل الدراسي الأول أي بعد الانتهاء من تدريس وحدة تحليل المقادير الجبرية مباشرة بحيث يتم اعتماد درجات الاختبار للتقويم المدرسي ويكون الطلبة قد استعدوا للاختبار بشكل جيد، وقد لاقت هذه الخطوة قبولاً جيداً لدى المعلمين، وبالتالي يمكن القول أن النتائج التي تم الحصول عليها تعبّر عن المستويات الواقعية للطلبة. وقد تم تحليل البيانات على مستوى الفقرة لتحديد معاملات الصعوبة والتمييز، وكذلك نسب النجاح على اعتبار أن هذه النسب هي قيم واقعية تعبر عن مستويات الطلبة الفعلية ليتم مقارنة نسب النجاح الفعلية للطلبة على الفقرات مع النسب المقدرة من مجموعات المحكمين لتقدير درجات القطع للاختبارحسب انجوف. وبعد مرور أسبوعين على التطبيق الأول تم إعادة تطبيق الاختبار نفسه على العينة نفسها من أجل حساب دقة ثبات التصنيف للمفحوصين في التطبيقين لدرجات القطع الناتجة عن تقديرات المحكمين

ثانياً: البيانات الخاصة بالمحكمين: تطلبت إجراءات الدراسة أن يكون هناك ثلاث مجموعات مستقلة من المحكمين، حيث تكونت عينة الدراسة من المحكمين الذين سيقومون بتقدير درجات القطع لأدوات الدراسة من (30) محكماً موزعين على ثلاث مجموعات متساوية في العدد بالصورة التي وردت تحت عنوان عينات الدراسة، وقد تم دعوة مجموعتين (الثانية والثالثة) إلى جلسة مشتركة، وتقديم شرح تفصيلي لما سيقومون به من تقدير لدرجات القطع لكل من الأهداف وصورتي الإختبار، وفي نهاية الجلسة تم مناقشة النسب التي تم تقديرها وسُمح لهم بمراجعتها والتعديل عليها لمن يرغب في ذلك. وقد تم بالفعل إجراء بعض التعديلات على القيم المتطرفة من بعض المحكمين. وبعد مرور أسبوعين على الجولة الأولى تم زيارة المحكمين بشكل فردي وطلب منهم إعادة التقدير لدرجات القطع. ويكون كل محكم من عينة المحكمين العشرين قد قام بتقدير درجة القطع لقائمة الأهداف وأحد نموذجي الاختبار.

أما فيما يتعلق بالمجموعة الأولى من المحكمين وهم أعضاء فريق التأليف لمبحث الرياضيات في وزارة التربية والتعليم فقد تم زيارتهم بشكل فردي لتقدير درجة القطع لقائمة الأهداف حيث قام الباحثان بتوضيح طبيعة عملية التقدير، وقد تمت العملية بسهولة ويسر، وقد تم إعادة الجولة الثانية بعد أسبوعين على انتهاء الجولة الأولى لعملية التقدير وذلك لتقدير الثبات في تقديراتهم لتحديد درجة القطع.

النتائج: يبين الجدول (1) المتوسطات الحسابية والانحرافات المعيارية لتقديرات المحكمين لدرجات القطع لقائمة الأهداف لكل محكم بشكل منفصل، ويعتبر المتوسط الحسابي لتقديرات المحكم هو درجة القطع ، ومتوسط تقديرات جميع المحكمين هو بمثابة درجة القطع التي تمثلها قائمة الأهداف السلوكية ، حيث يلاحظ أن قيم متوسطات التقدير لدرجات القطع للأهداف في الجولة الأولى قد تراوحت بين (0.45 و 0.65)، وأن متوسط جميع التقديرات لجميع المحكمين يساوي (0.56)،. وقد بلغت درجة القطع في الجولة الثانية من التقدير (0.56) وهي مساوية لدرجة القطع في الجولة الأولى . كما يلاحظ أن قيم الفرق المطلقة بين متوسطات التقديرات في الجولتين قد تراوحت بين (0.00 و 0.08) ، وبلغ معامل ارتباط بيرسون بين التقديرات في جولتي التحكيم (0.93)، وهذه مؤشرات على ارتفاع ثبات الاستقرار لتقدير درجة القطع وفق النموذج القائم على الأهداف.

جدول (1): المتوسطات الحسابية والانحرافات المعيارية لتقديرات المحكمين لدرجات القطع لقائمة الأهداف لوحدة تحليل المقادير الجبرية في جولتي التحكيم

	الجولة الأولى			الجولة الثانية		
رقم المحكم	المتوسط الحسابي	الانحراف المعياري	الفرق المطلق	المتوسط الحسابي	الانحراف المعياري	الفرق المطلق
1	0.56	0.16	0.00	0.60	0.12	0.04
2	0.61	0.14	0.01	0.64	0.10	0.03
3	0.62	0.12	0.01	0.62	0.18	0.00
4	0.65	0.13	0.02	0.66	0.18	0.01
5	0.64	0.12	0.01	0.56	0.16	0.08
6	0.58	0.16	0.01	0.60	0.13	0.02
7	0.59	0.16	0.03	0.61	0.14	0.02
8	0.65	0.10	0.00	0.63	0.09	0.02
9	0.64	0.10	0.02	0.63	0.10	0.01
10	0.55	0.15	0.02	0.54	0.14	0.01
11	0.62	0.14	0.03	0.62	0.15	0.00
12	0.52	0.15	0.03	0.52	0.15	0.00
13	0.48	0.11	0.00	0.49	0.12	0.01
14	0.64	0.14	0.02	0.63	0.15	0.01
15	0.46	0.15	0.01	0.47	0.16	0.01
16	0.45	0.18	0.01	0.48	0.16	0.03
17	0.54	0.16	0.01	0.54	0.16	0.00
18	0.54	0.13	0.03	0.54	0.14	0.00
19	0.56	0.13	0.01	0.54	0.13	0.02
20	0.49	0.15	0.01	0.49	0.14	0.00
21	0.51	0.13	0.01	0.50	0.12	0.01
22	0.50	0.13	0.02	0.49	0.13	0.01
23	0.58	0.16	0.02	0.56	0.16	0.02
24	0.55	0.16	0.02	0.54	0.16	0.01
25	0.49	0.12	0.02	0.50	0.12	0.01
26	0.48	0.17	0.05	0.48	0.18	0.00
27	0.55	0.10	0.03	0.57	0.11	0.02
28	0.52	0.19	0.04	0.52	0.17	0.00
29	0.53	0.18	0.01	0.52	0.16	0.01
30	0.58	0.13	0.01	0.58	0.14	0.00
جميع المحكمين	0.56			0.56		

وبالمقابل، يبين الجدول 2 المتوسطات الحسابية والانحرافات المعيارية لتقديرات المحكمين لدرجات القطع للصورة الأولى والثانية من الاختبار في جولتين منفصلتين من التحكيم حسب إجراءات نموذج أنجوف لتقدير درجة القطع الذي يعتمد بشكل أساسي على فقرات الاختبار.

جدول (2): المتوسطات الحسابية والانحرافات المعيارية لتقديرات المحكمين (10 محكمين) لفقرات الصورة الأولى والصورة الثانية من الاختبار محكي المرجع باستخدام نموذج أنجوف لتحديد درجة القطع.

	الصورة الأولى الجولة الأولى		الصورةالأولى الجولة الثانية		الصورة الثانية الجولة الأولى		الصورة الثانية الجولة الثانية		
رقم المحكم	المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري	الفرق المطلق	المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري	الفرق المطلق
1	0.71	0.12	0.70	0.11	0.01	0.62	0.17	0.63	0.16	0.01
2	0.68	0.08	0.67	0.09	0.01	0.58	0.17	0.58	0.15	0.00
3	0.77	0.13	0.76	0.11	0.01	0.62	0.14	0.62	0.14	0.00
4	0.63	0.17	0.64	0.15	0.01	0.70	0.17	0.68	0.08	0.02
5	0.69	0.10	0.67	0.11	0.02	0.56	0.15	0.58	0.16	0.02
6	0.61	0.13	0.62	0.11	0.01	0.59	0.15	0.60	0.13	0.01
7	0.67	0.13	0.65	0.14	0.02	0.63	0.19	0.64	0.17	0.01
8	0.57	0.13	0.57	0.11	0.00	0.62	0.12	0.62	0.12	0.00
9	0.65	0.15	0.64	0.12	0.01	0.72	0.13	0.71	0.12	0.01
10	0.70	0.15	0.68	0.16	0.02	0.62	0.20	0.59	0.18	0.03
جميع المحكمين	0.67		0.66			0.63		0.63		

وللمقارنة بين معامل الثبات لتقديرات المحكمين لدرجات القطع باستخدام النموذج القائم على الأهداف ومعامل الثبات لنموذج أنجوف القائم على فقرات الاختبار، فقد تم استخدام اختبار (Z) للعلامات الفشرية، وقد بلغت قيمة Z المحسوبة (0.064) وهي أقل من القيمة الحرجة لـ Z (1.96) وبالتالي لا يوجد فرق دال إحصائياً بين معاملي الثبات للنموذجين .

ولفحص تأثير خصائص المحكمين حسب وظيفتهم على تقديراتهم لدرجة القطع بالارتكاز على الأهداف السلوكية ، فقد تم استخدام تحليل التباين الأحادي لتقديرات مجموعات المحكمين الثلاث، حيث يبين الجدول (3) نتائج تحليل التباين لأوساط مجموعات المحكمين.

جدول (3): نتائج تحليل التباين الأحادي لدرجات القطع المقدرة من مجموعات الدراسة الثلاث للنموذج القائم على الأهداف

مصدر التباين	درجات الحرية	مجموع لمربعات	وسط لمربعات	ف المحسوبة	قيمة الاحتمال
بين المجموعات	2	13.031	4.344	1.471	0.246
داخل المجموعات	27	76.780	2.953		
المجمـــوع	29	89.811			

ويبين الجدول (3) أنه لا يوجد فرق ذو دلالة إحصائية بين تقديرات المحكمين من المجموعات الثلاث التي قامت بتقدير درجة القطع لقائمة الأهداف السلوكية.

وللمقارنة بين قيمة معامل ثبات التصنيف للمفحوصين إلى متمكنين وغير متمكنين باستخدام درجة القطع الناتجة من النموذج القائم على الأهداف، ودرجة القطع الناتجة من نموذج أنجوف، فقد تم تقدير ثبات التصنيف للمفحوصين في مرتي التطبيق على أساس درجة القطع المقدرة من المحكمين للأهداف (0.56) بالمقارنة مع درجة القطع الناتجة من تقديرات المحكمين حسب إجراءات نموذج أنجوف لفقرات الصورة الأولى من الاختبار والتي تساوي (0.67). ويبين الجدول 4 تصنيفات الطلبة إلى متمكنين وغير متمكنين في ضوء درجة القطع لكل صورة في مرتي تطبيق الاختبار عليهم.

جدول (4): درجات القطع للنموذجين، وتصنيفات الطلبة المختلفة ومعدلات الصواب ومعاملات كابا لها

النموذج	درجة القطع	عدد المتمكنين في التطبيقين	متمكن في الأول وغير متمكن في الثاني	متمكن في الثاني وغير متمكن في الأول	غير متمكن في الاثنين	معدل الصواب	معامل كابا
النموذج القائم على الأهداف	0.56	70	0	6	95	0.96	0.93
نموذج أنجوف القائم على الاختبار	0.67	46	0	4	121	0.97	0.94

يتبين من الجدول 4 أن معدل الصواب - وهو يعبر عن نسبة اتفاق التصنيف، أي مجموع نسب أعداد الأفراد الذين يتم تصنيفهم في المجموعة نفسها في مرتي التطبيق– فقد بلغت قيمته بالنسبة لدرجة القطع المحسوبة باستخدام النموذج القائم على الأهداف (0.96) وهي قيمة مرتفعة، وهي تقريباً مساوية لقيمة معدل الصواب بالنسبة لدرجة القطع المحسوبة باستخدام نموذج أنجوف، وقد تم حساب معامل كابا Kappa Coefficient وهو أسلوب إحصائي يأخذ أخطاء التصنيف الموجبة والسالبة بعين الاعتبار (علام، 2000)، حيث ان هذا المؤشر يحدد نسبة اتساق التصنيف وفقاً لدرجة قطع معينة وتصحيح هذه النسب من أخطاء التصنيف التي تعزى للصدفة، وكانت قيمتا هذا المعامل متساوية تقريباً كذلك بالنسبة للنموذجين وتساوي (0.93) و(0.94) وهي أيضاً قيم مرتفعة، ولتفسير هذه القيم تجدر الإشارة إلى أن قيم معامل كابا تتراوح بين (-1 و +1) وكلما اقتربت القيمة من (+1) دل ذلك على أن النسب الهامشية تقترب من بعضها إلى أن تتساوى في مرتي التطبيق، أما إذا كانت القيمة (-1) فتكون في حالة الاختلاف الكبير في تصنيف الأفراد في مرتي التطبيق. ونظراً لأن القيمة التي حصلنا عليها في هذه الدراسة تقترب من (+1) فهي تعتبر قيمة مرتفعة وتدل على ارتفاع نسبة اتساق التصنيف؛ مما يدل على ارتفاع قيمة معامل ثبات تصنيف الطلبة إلى متمكنين وغير متمكنين باستخدام درجة القطع الناتجة من النموذج القائم على الأهداف وكذلك الحال في نموذج أنجوف.

وللإجابة عن سؤال الدراسة الرابع المتعلق بتقدير الصدق التنبؤي للنموذج القائم على الأهداف بالمقارنة بنموذج أنجوف للتنبؤ بمستويات تحصيل الطلبة المدرسية (علامة نهاية الفصل الدراسي) في الرياضيات عند أكثر من محك للجودة، فقد تم الحصول على علامات الطلبة (عينة الدراسة) من الجداول النهائية للعلامات المدرسية، ومن ثم حساب معدل الصواب ومعامل كابا لدرجتي القطع لنموذج الأهداف ونموذج أنجوف باستخدام ثلاث محكات لجودة التحصيل هي (0.55، 0.60، 0.70). ويبين الجدول 5 قيم معامل كابا للنموذجين مع العلامات المدرسية النهائية للطلبة.

جدول (5): قيم معامل كابا لتصنيف الطلبة في الرياضيات حسب درجات القطع للأسلوبين والعلامات المدرسية

	قيم معامل كابا			
الأسلوب	درجة القطع للنموذج	مستوى 0.55	مستوى 0.60	مستوى 0.70
الأهداف	0.56	0.95	0.89	0.48
أنجوف	0.67	0.65	0.80	0.82

ويتبين من الجدول 5 أن أعلى قيمة لمعامل كابا للتنبؤ بمستوى تحصيل الطلبة حسب درجة القطع من نموذج الأهداف هي عند محك الجودة (0.55) ، وأن أقل قيمة لمعامل كابا هي عند المحك(0.70) وقد بلغت (0.48) وهي قيمة متدنية مقارنة بأعلى قيمة(95. عند محك الجودة 55. 0) ويرجع السبب في تدني هذه القيمة إلى الفارق الكبير بين درجة القطع المقدرة من النموذج (0.56) وقيمة محك الجودة كمعيار للتمكن (0.70)؛ ولذلك تعتبر درجة القطع الناتجة من تقديرات المحكمين بالارتكاز على الأهداف مناسبة جداً عند محكات الجودة (0.55، 0.60) بينما هي غير مناسبة عند المحك (0.70) وفق مؤشرات الصدق والثبات المشار اليها.

مناقشة النتائج والتوصيات: لقد لوحظ في كثير من الدراسات السابقة مثل : (Hambleton & Plake, 1995؛ Livingeston & Zieky, 1989؛ Diane, et al. 2005) أن درجة القطع المقدرة بطريقة أنجوف إذا ما قورنت مع غيرها من الطرق تكون غالباً أعلى من غيرها. وقد وجه شانج (Chang,2000) انتقاداً لأسلوب أنجوف كونه غالباً يعطي درجات قطع أعلى من غيره من خلال تحليله لأربعين دراسة قارنت بين أسلوب أنجوف وأساليب أخرى. والارتفاع الملحوظ في درجة القطع حسب أنجوف أثار اهتمام الكثير من الباحثين في مجال القياس النفسي، (Angoff, 1971)، الأمر الذي دعاهم إلى التفكير بإجراء تعديل على الأسلوب يؤدي إلى ضبط هذا الارتفاع في قيم التقديرات من المحكمين لمستويات الصعوبة للفقرات، وبالفعل فقد تم التوصل إلى طريقة أطلق عليها اسم طريقة أنجوف المعدلة، وتتلخص هذه الطريقة بأن يكون مدى التقديرات للمحكم لكل فقرة محصوراً بين قيمتين، في محاولة لتجنب القيم التي تقع خارج هذا المدى. وتلتها محاولات أخرى للتعديل على طريقة أنجوف منها تزويد المحكمين بمعلومات عن مستويات الصعوبة الفعلية للفقرات بعد أن يتم تجريبها على عينة من المفحوصين، وقد اشارت نتائج بعض الدراسات (Chinn &

Hertz, 2002؛ Plake, 2000)الى أن قيم التقديرات انخفضت واقتربت من المستويات الفعلية للمفحوصين، لكن هذا الإجراء أيضاً لاقى انتقاداً من بعض الباحثين (Bowers & Shendoll,1989) وهو أن أحد الافتراضات الهامة في تحديد درجة القطع يتمثل في ان تقدير هذه الدرجة يتم على أساس محكي المرجع، على اعتبار أنها تمثل الحد الأدنى من الكفاية في المحتوى المقصود بالقياس دون الاهتمام بخصائص مجموعة معينة، ولكن في حال إطلاع المحكمين على مستويات الصعوبة للفقرات الواقعية المأخوذة من عينة ما من المفحوصين وأخذ المحكمين هذه المعلومات بعين الاعتبار أثناء تقديراتهم فسيؤثر ذلك على تقديراتهم، وبالتالي سيدخل عامل معياري المرجع ويؤثر على التقديرات. وبالإضافة إلى ذلك في كثير من حالات الاختبارات لأغراض التحصيل أو لأغراض القبول والمنافسة يكون من الصعب تجريب الاختبار على عينات مماثلة للمفحوصين لأن ذلك يعرض افتراض سرية الاختبارات للانتهاك.

إن من أهم الأسباب التي دعت للتفكير بأسلوب جديد في هذه الدراسة لتقدير درجة القطع بالارتكاز على الأهداف بدلاً من الارتكاز بشكل مباشر على الفقرات، هو أن درجة القطع في أسلوب أنجوف تتأثر بشكل كبير ومباشر بطبيعة الفقرات ومستويات صعوبتها وصياغتها، ولا يوجد أمام المحكّم إلا متن الفقرة وبدائلها، وبناءً على هذا المتوفر سيقرر مستوى صعوبة الفقرة، ولذلك فإن تقديراته ستتغير بتغير تلك الفقرة حتى ولو كانت تقيس الجزء نفسه من المعرفة، وما يؤكد ذلك أنه في حال بناء أكثر من صورة للاختبار مأخوذة من نفس المجال او المحتوى الدراسي، فإن درجات القطع المقدرة من المحكمين تختلف باختلاف الفقرات (بني عطا،2005).

إن مجموعة الفقرات التي تشكل الاختبار حسب أسلوب أنجوف هي عينة تمثل مجال السلوك المراد قياسه، وبالتالي فان درجة القطع لا بد وان تتأثر بالخطأ العيني (خطأ المعاينة Sampling Error، بالإضافة إلى خطأ التقدير Estimating Error، بمعنى أن اسلوب أنجوف يتعرض لهذين النوعين من الخطأ، أما بالنسبة للأسلوب المقترح فهو متحرر نسبياً من الخطأ العيني، لأنه يعرض على المحكم الأهداف السلوكية ضمن كل مجال من المجالات الفرعية المكونة للمجال الكلي للسلوك، ومع كل هدف مجموعة الفقرات التي تقيسه وتقع في مجاله، مما قد يؤدي إلى التقليل إلى حدٍ كبير من الخطأ العيني أثناء عملية التقدير. فعندما يُعرض على المحكم الهدف ويليه مجموعة من الفقرات التي تقيس هذا الهدف يتوقع أن تجعل المحكم أكثر تبصّراً بالهدف ومستواه المعرفي ومدى تعقيد العمليات العقلية المعرفية التي يتطلبها هذا الهدف. كما أن الأهداف لأي محتوى دراسي يتوقع أن تكون مستقرة ومتفق عليها ضمناً عند تطوير المنهاج التعليمي؛ ولذلك يمكن أن نتصور توزيعاً واحداً للأهداف يقابل توزيع مجتمع Population distribution يشتمل على عدة عينات غير محدودة من الفقرات بحيث تشكل كل عينة من هذه الفقرات اختباراً، وتوزيع أوساط العينات هو التوزيع العيني للأهداف Sampling distribution. وأن درجة القطع بالنسبة للنموذج القائم على الأهداف يتوقع أن تمثل متوسط متوسطات درجات القطع لعينات من الأسئلة مسحوبة من المجال السلوكي المقصود، لأن المحكم ينظر إلى الطيف الكامل من الأسئلة المحتملة للهدف الواحد، ويحكم على علامة القطع من خلال فهم واضح لطبيعة الهدف والأسئلة التي يمكن أن توضع لقياسه، وبالتالي يتوقع أن يقدم درجة قطع أكثر ثباتاً(More precise) وأكثرصدقا وواقعية(More accurate).اما الميزة التي تتوفر في اسلوب انجوف مقارنا بالأساليب الأخرى حسب بعض الدراسات(Livingeston & Zieky, 1989؛ Diane, et al. 2005؛ Kaufman,2000؛ George & Sayeed & Oyebode, 2006) والمتمثلة بارتفاع معامل ثبات تقدير المحكمين(0.96) فلم تعد ميزة ينفرد بها عند مقارنته بألأسلوب القائم على الأهداف(0.93).

وقد أظهرت النتائج المتعلقة بأثر خصائص المحكمين على تقديراتهم لدرجة القطع للأهداف، ومن خلال تحليل التباين لتقديرات المجموعات الثلاث في الجدول (3) أنه لا يوجد فرق ذو دلالة إحصائية بين تلك التقديرات للمجموعات المختلفة، ويمكن أن تفسر هذه النتيجة لصالح النموذج ايضا، فهي تعبر عن وضوح المهمة التي يتطلبها المحكم لإعطاء تقديراته لدرجات القطع، بمعنى أن الثبات عبر المحكمين يعمل هنا كمؤشر صدق على ما يتم تقديره لأن الأهداف تشكل إطاراً مرجعياً لجميع فئات المحكمين.كما أظهرت النتائج المتعلقة بثبات تصنيف الطلبة إلى متمكنين وغير متمكنين في مرتي التطبيق، ومعدلات الصواب باستخدام درجة القطع المقدرة من قبل المحكمين للنموذج القائم على الأهداف، ولنموذج أنجوف بأن قيم معاملات كابا كانت عالية للنموذجين حيث كانت (0.93) و (0.94) على التوالي، كما انّ معدلات الصواب (Hit-rate) أيضاً مرتفعة للنموذجين وقد بلغت (0.96) لنموذج الأهداف و(0.97) لنموذج أنجوف، وهذا يعدّ مؤشراً جيداً على دقة التصنيف للنموذج القائم على الأهداف كما هو الحال بالنسبة لنموذج أنجوف، بمعنى أن هذه الخاصية الايجابية أو الميزة التي يتمتع بها نموذج أنجوف (ملاءمته لأغراض التصنيف) يتمتع بها أيضاً النموذج المقترح.

وفيما يتعلق بالنتائج المتعلقة بالصدق التنبؤي للنموذجين؛ فقد بينت النتائج للتنبؤ بمستويات تحصيل الطلبة المدرسية في الرياضيات أن أعلى قيمة لمعامل كابا كان لنموذج الأهداف عند استخدام محك الجودة (0.55) وقد بلغت قيمته (0.95)، ويعود سبب ارتفاع قيمة هذا المؤشر هو أن قيمة المحك مساوية تقريباً لدرجة القطع الناتجة من تقديرات المحكمين للأهداف، بينما كانت قيمة معامل كابا لنموذج أنجوف عند المحك نفسه أقل بكثير منها بالنسبة للأهداف وقد بلغت (0.65)، ويعود السبب في انخفاض هذه القيمة بالنسبة لنموذج أنجوف الى الفرق الكبير بين درجة القطع لنموذج أنجوف والمحك، فكلما اقتربت درجة القطع من محك الجودة زادت قيمة معامل كابا. أما أقل قيمة لمعامل كابا لنموذج

الأهداف فكانت عند محك الجودة (0.70) وقد بلغت قيمته (0.48) ويعود انخفاض هذه القيمة يعود إلى نفس السبب السابق وهو ابتعاد درجة القطع عن المحك، مما يعني تميّز النموذج المقترح على نموذج أنجوف وفقاً لهذا المؤشر الهام.

من خلال النتائج التي أسفرت عنها هذه الدراسة التي هدفت بشكل أساسي إلى تطوير أسلوب مناسب لتقدير درجة القطع لمجال سلوكي محدد، ومن ثم تجريب هذا الأسلوب بشكل عملي وتقديم مؤشرات الثبات ومؤشرات الصدق المختلفة ذات الصلة. وقدرته على التنبؤ بمستويات المفحوصين، ودقة قرارات التصنيف لهم، ومقارنته مع أحد أكثر الأساليب شيوعاً واستخداماً في تقدير درجة القطع؛ يمكن استخلاص أهم الاستنتاجات التالية:

- إن درجة القطع الناتجة من تقديرات المحكمين بالارتكاز على الأهداف تعدّ أقل من درجة القطع الناتجة من أسلوب أنجوف، وهي قريبة جداً من المستويات الفعلية للطلبة، ومن هنا فهي تعتبر أكثر واقعية ودقة من تلك المقدرة بأسلوب أنجوف القائم على فقرات الاختبار.
- قد تتأثر درجات القطع الناتجة من استخدام إجراءات نموذج الأهداف بتغير فقرات الاختبار، أو بمجموعة المفحوصين، وهي متحررة نسبياً من الفقرات والمفحوصين، وبالتالي يتوقع عدم التغيرInvariance لعلامات القطع. ويمكن ان يكون اجراء المزيد من الدراسات مبررا في هذا الإطار.
- للنموذج القائم على الأهداف قيمة تشخيصية، إذ يمكن عن طريقه تحديد جوانب القوة والضعف في تحصيل كل طالب وإتقانه للمهارات الأساسية بالنسبة لكل هدف، مما يساعد على تعديل مسار عملية تعلم هذه المهارات، خاصة في المتعلقة بالأهداف التي لم تتحقق. فمن ناحية عملية يمكن الاستفادة من نتائج القياس في حال كانت درجة القطع مرتكزة على الأهداف بمقارنة نسبة الذين حققوا الهدف بشكل فعلي مع النسبة المقدرة للهدف من المحكمين، إذ يمكن وضع خطط لمعالجة الضعف في مستويات الطلبة على الأهداف التي لم تتحقق.

وبناءً على النتائج والاستنتاجات السابقة يمكن التقدم بالتوصيات التالية :

- ترجيح اعتماد النموذج القائم على الأهداف، نظراً للمزايا التي يتمتع بها هذا النموذج مقارناً بأكثر النماذج شيوعاً لأنه يعطي تقديرات تمتاز بالدقة والصدق والثبات، وهي متحررة من خصائص الفقرات والمفحوصين.
- قد يدفع اعتماد هذا النموذج مطوري المناهج إلى تزويد كل وحدة في المنهاج بقائمة الأهداف السلوكية التفصيلية (Mini - objectives) التي تعرف اجرائيا الهدف التدريسي العام نسبيا، ومستوى الأداء المتوقع لكل طالب على الهدف، وبالتالي مستوى الأداء لأهداف الوحدة بشكل عام، وإذا ما تم ذلك لجميع وحدات الكتاب وجميع المباحث سيصبح لدى المعلم ما يمكن أن نطلق عليه بنك الأهداف (Objectives bank) ليطلّع عليه الطالب وولي الأمر ومدير المدرسة، مما يؤدي إلى تحسين عملية التعليم ونوعيته، وخاصة التعلم الفردي، من خلال توجيه الطالب في التحضير والمراجعة للمادة الدراسية والاستعداد للامتحانات، كما يتوقع أن يساعد المعلم نفسه في عملية التخطيط والتعليم، ومساعدة أولياء الأمور الذين يحاولون توجيه أبناءهم في عمليات التحضير والمراجعة، والتقويم الذاتي.
- دعوة الباحثين الى إخضاع النموذج القائم على الأهداف لدراسات أخرى في مجالات هدفية Domain of objectives، أو محتويات تعليمية مختلفة حتى في المجالات المفتوحة نسبياً مثل امتحان مستوى اللغة الانجليزية في الجامعات، وامتحان الرخصة الدولية لقيادة الحاسوب، ، فقد لا تسمح الكثير من المواقف الانتظار طويلاً لتجريب نماذج متكررة من الاختبار قبل الوصول إلى درجات قطع شبه مستقرة كما هو الحال اختبار التو فل TOEFL.

المصادر والمراجع:

بني عطا، زايد صالح إبراهيم. (2005). استخدام نموذج ذي الحدين لفحص تقديرات المحكمين لدرجة القطع لاختبار بدلالة عدد بدائل الفقرة. رسالة دكتوراه غير منشورة، جامعة اليرموك : الأردن.

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

Angoff, W. (1971). *Scales, norms and equivalent scores.* In Thorndike R.L. (Ed.). Educational Measurement. (2nd ed.) Washington, D.C.: American Council In Education.

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

Boursicot, K. (2006). Setting Standards in a Professional Higher Education. *Medical School Higher Education Quarterly 60 (1), 74–90.*

Chang, L. (2000). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12, 151-165.*

Chinn, R. N., & Hertz, N. R. (2002). Alternative Approaches to Standard Setting for Licensing and Certification Examinations. *Applied Measurement in Education, 15, 1-14.*

Cizek, G. j. (1996). Setting Passing Scores. *Journal of Educational Measurement, 15, 20-30.*

Cizek, G. J. (Ed.) (2001). *Setting Performance Standards: Concepts Methods, and Perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Coscarelli, W. & Shrock, S. (2006). *The Two Most Important Things You Can Do to Improve Testing in Your Organization.* Retrieved 12th February 2007, From: www. coscarel@siu.edu

Diane W. ; Fudala, M.; Butter, J.; Siddall, V.; Feinglass, J.; Wade, L. & McGaghie, W. (2005). *Comparison of Two Standard-setting Methods for Advanced Cardiac Life Support Training.* Academic

Medical Education. Retrieved 12th, February 2007, from: http://www.nlm.nih.govLaird.

Krathwhl, D. (2002). A Revision of Bloom's Taxonomy: an Overview. *Theory into Practice*, 41(4), 212-218.

Livingston, S. & Zieky, M. (1989). A Comparative Study of Standard Setting Methods. *Applied Measurement in Education,2, 121-141.*

Plake, B.; Impara, J., & Irwin, P. (2000). Consistency of Angoff- Based Predictions of Item Performance: Evidence of Technical Quality of Results from the Angoff Standard Setting Method. *Journal of Educational Measurement, 37, 437-355*

Shepard, L. . (1984). *Setting performance standards,* In Berk, R. (ed.), A guide to Criterion – referenced Test Construction. Hopkins University Press. Texas A&M University, Corpus Christi.

Sizmur, S. (1997). Look Back in Angoff : A Cautionary Tale. *British Educational Research Journal : 23, 3-11*. Retrieved: 20th, Feb.2007 from: EBSCOhost Research database.

Medicine. Retrieved: 10th February 2007, from: http:// academicmedicine.org/pt/

George, S.؛ Sayeed H. & Oyebode, F. (2006*). Standard Setting: Comparison of Two Methods.* Retrieved 15th July, 2007, From http://www.biomedcentral.com/1472-6920/6/46.

Glass, G. (1977). *Standards and Criteria*. Retrieved 2nd July, 2007, from: http://www.wmich.edu/evalctr/pubs/ops/ops10.html

Goodwin, L. (1996). Focus on Quantitative Methods Determining Cut- Off Scores. *Research in Nursing & Health, 19, 249-256*

Haertel, E. (1985). Construct Validity and Criterion-referenced Testing. *Review of Educational Research, 55(1), 23-46*

Haertel, E. & Lorie, W. (2000). *Validating Standards-based Test Score Interpretations*. Retrieved 5th July, 2007, From: www-stat.stanford.edu/~rag/ed351/Std-Setting.pdf

Hambleton, R. (1982). *Test Score Validity and Standards-setting Mmethods.* In Berk, R. A. (ed.), Criterion-referenced Measurement: the State of Art, 2ed. London: The John Hopkins Press Ltd.

Hambleton, R. & Plake, B. (1995). Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education, 8, 41-55.*

Haswell, R. (2003). *Raising the Cut-Off Score on the College Board Advanced Placement Examinations in Composition and Literature: Justification for a First-Year Writing Curriculum*. Retrieved 12th February 2007, from: http://www.collegeboard.com/press/senior99/html/990831b.html.

Hertz, N. & Chinn, R. (2002). *The Role of Deliberation Style in Standard Setting For Licensing and Certification Examinations.Paper* Presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana, Retrieved 2nd July, 2007, From www.ncme.org/repository/incoming/91.pdf.

Impara, J., & Plake, B. (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement, 35, 69-81*

Jager, R. (1989). Certification of Student Competence. In Robert L. Linn (ed.). *Educational Measurement*. Collier Macmillan Publishers London.

Kane, M. (1987). On the Use of IRT Models with Judgmental Standard Setting Procedures. *Journal of Educational Measurement, 24, 333-345.*

Kane, M. (1998). Choosing Between Examinee-Centered and Test- Centered Standard-Setting Mmethods. *Educational Assessment, 5,129-145.*

Kaufman, D. ; Manny K. ; Muijt, A. & Van Der V. (2000). *A Comparison of Standard-Setting Procedures for an OSCE in Undergraduate*