

دقة تقدير معالم الفقرات عند استخدام اربعة نماذج لوجستية في إطار نظرية الاستجابة للفقرة

معين سلمان نصرأوين *

ملخص

هدفت الدراسة إلى تحري دقة تقدير معالم الفقرات عند استخدام أربعة نماذج في إطار نظرية الاستجابة للفقرة باختلاف متغير طول الاختبار (25 فقرة و 50 فقرة) وحجم عينة المفحوصين (500 مفحوص و 1500 مفحوص)، هي: النموذج الثلاثي المعلمة، والنموذج الثنائي المعدل بإضافة قيمة ثابتة (0.20)، ونموذج عتبة التخمين، والنموذج القائم على الصعوبة. استخدمت الدراسة أسلوب المحاكاة (Simulated Data) بحيث تكون مطابقة لافتراضات ومتطلبات النموذج الثلاثي المعلمة، وذلك وفقاً للخصائص الآتية: معلمة الصعوبة توزعت توزيعاً طبيعياً بمتوسط (صفر) وانحراف معياري (واحد)، كذلك توزعت معلمة التمييز توزيعاً منتظماً بقيمة دنيا تساوي (0.40) لوجت وقيمة عظمى مساوية لـ (1.7) لوجت، أما معلمة التخمين فقد وُلدت بقيمة دنيا تساوي (0.05) وقيمة عظمى مساوية لـ (0.40). أظهرت نتائج الدراسة أن الفروق بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ والتحيز في حالة كل من معلمة التمييز، ومعلمة الصعوبة، ومعلمة التخمين كانت دالة إحصائياً عند مستوى الدلالة ($\alpha=0.05$) في حالة النموذج الثلاثي المعلمة مقارنة مع النماذج الأخرى، يليه النموذج الثنائي المعدل، أما فيما يتعلق بمؤشر الكفاءة النسبية فقد أظهرت النتائج عموماً أفضلية لكل من النموذج الثلاثي المعلمة، والنموذج الثنائي المعدل مقارنة مع النماذج الأخرى المستخدمة، كما أظهرت نتائج الدراسة أن الجذر التربيعي لمتوسط مربعات الخطأ، ومؤشر التحيز جميعها تقلّ بزيادة حجم العينة من (500) مفحوص إلى (1500) مفحوص، وبزيادة طول الاختبار من (25) فقرة إلى (50) فقرة.

الكلمات الدالة: دقة التقدير، النماذج المعدلة، النماذج المختلطة.

المقدمة

تفترض نظرية الاستجابة للفقرة (Item Response Theory) وجود سمات أو خصائص معينة يشترك فيها جميع الأفراد، ولكنهم يختلفون في مقدار أو نوع ما يمتلكون من هذه السمات، ويمكن الاستدلال على مقدار ما يمتلك الفرد من هذه السمات من خلال السلوك الملاحظ للفرد المرتبط بالسمة موضوع القياس، ويتمثل في استجابته على فقرات مقياس السمة؛ ولذلك تسمى هذه السمات بالسمات الكامنة (Embretson and Reise, 2000)، وتُعد هذه النظرية من التطورات الحديثة في مجال القياس النفسي والتربوي؛ لما قدمته من طرق سيكومترية ذات فعالية كبيرة في بناء المقاييس النفسية والتربوية، بالإضافة إلى طريقة تفسير درجات الأفراد على هذه المقاييس (علام، 2005). إذ تقوم هذه النظرية على افتراض وجود متصل للسمة، يمكن من خلاله تقدير احتمال الإجابة الصحيحة للفرد عن فقرة ما إذا عُلم موقعه (θ) على هذا المتصل، كما تفترض أن العلاقة بين أداء الفرد على فقرة ما وقدرته تتحدد من خلال منحنى خاصية الفقرة (Item Characteristic Curve)، وأن احتمال أن يجيب الفرد إجابة صحيحة على فقرة ما يزداد بزيادة قدرة الفرد (Hambleton and Jones, 2004).

وتُعد استقلالية القياس إحدى أهم المزايا المرتبطة بنظرية الاستجابة للفقرة، وتعني أن تقدير معالم الفقرات (الصعوبة، والتمييز، والتخمين)، يكون مستقلاً عن خصائص الأفراد الذين استخدموا في تقدير هذه المعالم (Sample-Free) وأن تقدير قدرة الفرد يكون مستقلاً عن عينة الفقرات التي تُطبق عليه (Item-Free)، وتحقق هذه الخاصية يصبح بالإمكان القيام ببعض التطبيقات مثل: معادلة الاختبارات، وبناء بنوك الأسئلة، والقياس التكيفي (Reckase, 2009)، ولتحقيق استقلالية القياس؛ اهتم البحث السيكمترى لنظرية الاستجابة للفقرة بتطوير نماذج احتمالية تقوم على تحديد العلاقة بين أداء الفرد على فقرات الاختبار وبين السمات أو القدرات الكامنة التي تقف وراء هذا الأداء، وتصنف هذه النماذج في فئتين، الفئة الأولى: النماذج ثنائية التدرج (Dichotomous Models)، الفئة الثانية: النماذج متعددة التدرج (Polytomous Models) (Hambleton and Jones, 2004) وتُعد الفئة الأولى من أشهر النماذج استخداماً في بناء الفقرات والمقاييس، كما توصل البحث السيكمترى المتعلق بنظرية

* وزارة التربية والتعليم، الأردن. تاريخ استلام البحث 2016/6/30، وتاريخ قبوله 2016/8/15.

الاستجابة للفقرة إلى إيجاد طرق تقدير عددية بهدف التوصل إلى أفضل تقدير لمعالم الفقرات ولمعلمة القدرة للأفراد (Embretson and Reise, 2000).

ويرى لندن وهامبلتون (Linden and Hambleton, 1997) أن وجود أخطاء في تقدير معالم الفقرات ينعكس بشكل سلبي على دقة تقدير تلك المعالم، كما يؤثر على دالة معلومات الاختبار، وعموماً تُعتمد عملية تقدير معالم الفقرات وتقديرات معلمة القدرة للمفحوصين على عوامل عدة، أهمها:

أ- اختيار الطريقة المناسبة للتقدير، فقد أشار بعض الباحثين مثل هامبلتون، ولورد إلى أن طريقة بيزز تعمل بشكل أفضل من طريقة الأرجحية العظمى مع الاختبارات القصيرة وبحجم عينة أقل (Grujter and Kamp, 2002).

ب- تزداد الدقة في تقدير القدرة كلما قلّ الخطأ المعياري للتقدير (Standard Error of Estimate (SEE وفي هذا السياق أشار كل من هامبلتون وسواميناثان وروجرز (Hambleton, Swaminathan and Rogers, 1991) إلى أن الخطأ المعياري للتقدير يعتمد على عوامل هي:

1- عدد فقرات الاختبار: عند زيادة عدد فقرات الاختبار تقل قيمة الخطأ المعياري في التقدير.
2- القدرة التمييزية لفقرات الاختبار: فكلما زادت معلمة التمييز للفقرة، قلّ الخطأ المعياري للتقدير؛ لأن إجابة الفقرة من خلال التخمين العشوائي تُصبح صعبة.

3- درجة المواءمة بين صعوبة الاختبار وقدرة المفحوصين: فكلما اقتربت صعوبة الفقرة من قدرة المفحوص قلّ الخطأ المعياري للتقدير. ومن هنا فإن الاختبار السهل جداً أو الصعب جداً يرتبط بخطأ معياري كبير.

كما أشارت ستوكينغ (Stocking, 1990) إلى أن نجاح تطبيقات نظرية الاستجابة للفقرة كالمقياس التكيفي، وبنوك الأسئلة وغيرها من التطبيقات يعتمد بشكل كبير على دقة التقديرات لكل من معالم الفقرات ومعلمة القدرة للمفحوصين، وفي هذا الصدد أشارت (ستوكينغ) إلى إمكانية الحصول على تقديرات لمعالم الفقرات بشكل أدق، إذا كان التخطيط لاختبار عينة المعايرة يأخذ بعين الاعتبار دقة تقديرات معالم الفقرة، وقد قامت بإيضاح العلاقة التي تربط بين معلمة قدرة المفحوص (θ) ودقة تقدير معالم الفقرات، من خلال استخدام توقع المعلومات، حيث استخدمت إجراءات الاشتقاق البسيط والاشتقاق الجزئي لدالة معلومات الفقرة بالنسبة لكل معلم من معالم الفقرة بهدف تحديد مستويات القدرة التي تقابل أقصى معلومات، عند تقدير معالم الفقرات للحصول على دقة أكبر في تقدير تلك المعالم.

دالة معلومات الفقرة:

تُعدّ دالة المعلومات من المفاهيم الأساسية في نظرية الاستجابة للفقرة، سواء كانت الفقرة ثنائية التدرج أو متعددة التدرج؛ حيث يُنظر لدالة معلومات الفقرة على إنها دالة للعلاقة بين قدرة المفحوص والمعلومات التي تقدمها الفقرة، إذ تُمثل هذه الدالة كمية المعلومات التي يُمكن أن توفرها فقرات الاختبار عند مستوى قدرة محدد، بالإضافة لإمكانية تحديد النقطة على متصل القدرة التي يمكن من خلالها الحصول على أقصى معلومات عند مستوى قدرة (θ) معين، ويُشير كل من هامبلتون وسواميناثان وروجرز (Hambleton, Swaminathan and Rogers, 1991) إلى أن دالة المعلومات تُعد وسيلة لوصف كل من الفقرات والاختبار، ومن التطبيقات التي تساهم بها دالة المعلومات ما يأتي:

أ- بناء الاختبارات واختيار فقرات الاختبار بناءً على كمية المعلومات التي تقدمها كل فقرة وبما يتلاءم مع طبيعة وهدف الاختبار.
ب- تقييم مدى دقة أدوات القياس، مما يتيح فرصة المقارنة بينها، والمقارنة بين طرق التصحيح المستخدمة فيها. إن ما يميز دالة المعلومات أنها يمكن أن تُعرف عند كل مستوى من مستويات معلمة القدرة (θ) على متصل السمة (De Ayala, 2009).

أثر معالم الفقرة على دالة المعلومات:

يمكن تحديد الأثر الذي تتركه معالم الفقرات في إطار النموذج ثلاثي المعلمة على دالة المعلومات من خلال المعادلة الآتية:

$$I(\theta, U_i)_{\max} = \frac{D^2 a_i^2}{8(1 - c_i^2)} [1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{3/2}] \dots\dots\dots [1]$$

(Hambleton, Swaminathan, 1985: 105)

حيث إن:

$I(\theta, U_i)_{\max}$: القيمة القصوى لدالة المعلومات، c_i : تشير لمعلمة التخمين.

a_i : معلمة التمييز، D : ثابت ويساوي (1.77).

إذ يُلاحظ من المعادلة أعلاه أن هناك أثرًا لكل من معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين على دالة المعلومات؛ فالمعلومات تزداد كلما كانت قيمة معلمة الصعوبة (b) قريبة من مستوى القدرة (θ) للمفحوص، كما تزداد قيمة دالة المعلومات باقتراب معلمة التخمين (c) من الصفر، حيث يمكن الحصول على أعلى قيمة لدالة المعلومات حينما تصبح قيمة معلمة التخمين مساوية للصفر، وعندها فإن القيمة القصوى لدالة المعلومات (θ_{max}) تساوي معلمة الصعوبة (b)، وعندما تكون قيمة معلمة التخمين أكبر من صفر ($c > 0$) فإن قيمة (θ_{max}) تُصبح إلى يمين معلمة الصعوبة (b) وبازدياد قيمة معلمة التمييز (a) يزداد ميل منحني الفقرة، ويقل التباين وترتفع قيمة المعلومات (Hambleton, Swaminathan and Rogers, 1991).

وقد هدفت هذه الدراسة إلى اختبار فعالية نماذج وُضعت من قبل كل من كاو وستوكس (Cao & Stokes, 2008) تستند إلى طريقة التقدير وفق نظرية بيزن (Baysian). وهذه النماذج هي: نموذج عتبة التخمين والنموذج القائم على الصعوبة، كما تم استخدام نموذجين آخرين هما: النموذج الثلاثي المعلمة، والنموذج الثنائي المعلمة المعدل إلى ثلاثي من خلال تثبيت معلمة التخمين عند القيمة (0.20)، وفيما يأتي استعراض لهذه النماذج المُستخدمة في هذه الدراسة:

نماذج نظرية استجابة الفقرة المعدلة Modified IRT Models

تُقسّم النماذج المعدلة لنماذج نظرية الاستجابة للفقرة إلى قسمين رئيسيين، هما:

1- النموذج الأحادي المعلمة المعدل:

تم تعديل النموذج أحادي المعلمة بتحويله إلى ما يُعادل النموذج ثنائي المعلمة؛ وذلك بافتراض أن معاملات تمييز الفقرات تأخذ قيمة واحدة يتم تقديرها، بإضافة ثابت غير مساوٍ للصفر عوضًا عن معلمة التخمين (Barnes & Wise, 1991). Baldwin, 2006)

2- النموذج الثنائي المعلمة المعدل:

يتمثل التعديل الذي أُدخل على هذا النموذج بافتراض أن قيمة معلمة التخمين ثابتة وغير مساوية للصفر، وتساوي في حدّها الأعلى واحدًا على عدد البدائل، ومن هنا ينتج عن استخدام هذا النموذج مطابقة أفضل للبيانات، ودقة في تقدير كل من معالم الفقرات ومعلمة القدرة، وأما حجم العينة المطلوب من المفحوصين لإعطاء تقديرات دقيقة فهو مُماثل لحجم العينة المطلوب في النموذج الثنائي المعلمة التقليدي، الأمر الذي يجعل التطبيق العملي لهذه النماذج أقل صعوبة، خاصة عند وجود أعداد قليلة من المفحوصين.

(Baldwin, 2006; Barnes & Wise, 1991; Sireci, 1992).

النماذج المختلطة (Mixture Model)

قامت ستوك وكاو (Cao and Stokes, 2008) بتطوير نموذجين يستندان إلى نماذج نظرية بيزن وهما: نموذج عتبة التخمين، ونموذج التخمين القائم على الصعوبة؛ وقد تم في هذه الدراسة استخدام كل من هذين النموذجين بالإضافة إلى كل من النموذج الثلاثي المعلمة والنموذج الثنائي المعلمة المعدل إلى ثلاثي وفيما يلي استعراض للنموذجين:

1- نموذج عتبة التخمين: (The IRT Threshold guessing Model)

يفترض نموذج عتبة التخمين أن بعض المفحوصين يُمكن أن يجيبوا عن فقرات الاختبار على أساس معرفتهم إلى أن يصلوا إلى حدٍّ معين، فيقومون بالتخمين العشوائي بعد ذلك، ويقترح هذا النموذج تحديد عتبة التخمين للمفحوص لتحديد أي الفقرات التي ستُقدّم له، وبناءً عليها يتم تقدير معلمة القدرة له، كما يتضمن معلمة يُطلق عليها موقع الفقرة، وتُحدد العتبة (الحد الفاصل) الذي يبدأ عنده المفحوص باللجوء إلى التخمين للإجابة عن فقرات الاختبار.

وتتمثل الصيغة الرياضية لهذا النموذج، التي تبين احتمالية إجابة المفحوص (i) على الفقرة (j) إجابة صحيحة بالآتي:

$$P(x_{ij} = 1 | \theta_i, \delta_j, \gamma_j, c_j, \alpha_i) = \frac{\exp[\gamma_j(\theta_i - \delta_j) - I(j > \alpha_i)(\gamma_j(\theta_i - \delta_j) - c_j)]}{1 + \exp[\gamma_j(\theta_i - \delta_j) - I(j > \alpha_i)(\gamma_j(\theta_i - \delta_j) - c_j)]} \quad \dots\dots\dots [2]$$

(Cao and Stokes, 2008: p.5)

حيث إن:

(α_i): تشير إلى موقع معلمة الصعوبة للفقرة لكل مفحوص (Examinee's Item Location Threshold) Parameter.

(c_i): معلمة التخمين للفقرة.

γ_j : معلمة التمييز للفقرة.

θ_i : معلمة القدرة للمفحوص.

δ_j : معلمة صعوبة الفقرة.

أما الاقتران ($I(j > \alpha_i)$) فيقارن بين احتمالية الإجابة على الفقرة (j) والحدّ الفاصل (α_i)، ويُعطى هذا الاقتران القيمة (1) في حالة كانت صعوبة الفقرة أكبر من العتبة، مما يعني حصول عملية التخمين، ويُعطى القيمة (صفر) في حال كانت قيمة العتبة أكبر من موقع الفقرة (j) أو تساويها، وبهذا يصبح الجزء التالي من النموذج مساوياً للصفر، ويتحوّل النموذج المختلط إلى النموذج الثنائي المعلمة، وفيما يأتي الصيغة الرياضية لهذا الاقتران:

$$I(j > \alpha_i) = \begin{cases} 1, & j > \alpha_i \\ 0, & j \leq \alpha_i. \end{cases}$$

2- نموذج التخمين القائم على الصعوبة: The IRT Difficulty-Based Guessing Model

يفترض هذا النموذج أن كل المفحوصين سواء الذين يستخدمون سلوك التخمين أو الذين لا يستخدمونه سيجيبون عن فقرات الاختبار، ولكن الطلبة الذين يستخدمون سلوك التخمين يجيبون عن فقرات الاختبار السهلة، ثم يستخدمون التخمين للإجابة عن بقية فقرات الاختبار، وبحسب الصيغة الرياضية لهذا النموذج، فإن عتبة الصعوبة أو فاصل الصعوبة ترتبط مع معلمة القدرة للمفحوص.

إذ تتمثل الصيغة الرياضية لنموذج التخمين القائم على الصعوبة التي تبين احتمالية إجابة المفحوص (i) على الفقرة (j) إجابة صحيحة تعطى وفق العلاقة الآتية:

$$p(x_{ij} = 1 | \theta_i, a_j, b_j, c_j, \delta_i) = \frac{\exp[a_j(\theta_i - b_j) - \eta_i I(b_j - (\theta_i + \delta_i)) (a_j(\theta_i - b_j) - c_j)]}{1 + \exp[a_j(\theta_i - b_j) - \eta_i I(b_j - (\theta_i + \delta_i)) (a_j(\theta_i - b_j) - c_j)]}$$

حيث إن:

a_j : معلمة تمييز الفقرة (j) (Cao and Stokes, 2008:p6)

c_j : معلمة تخمين الفقرة، $b_j(j)$: معلمة الصعوبة للفقرة (j).

e : الأساس اللوغاريتمي الطبيعي ويساوي (2.718)، θ : معلمة القدرة للمفحوص (i)

η : معامل رياضي يساوي (1) إذا لجأ المفحوص للتخمين وتساوي (صفر) إذا لم يلجأ المفحوص للتخمين.

δ_i : معلمة تقيس عتبة الصعوبة التي يُخمن عندها المفحوص.

والاقتران ($I(b_j - (\theta + \delta_i))$) تكون قيمته مساوية (واحد) إذا كانت معلمة الصعوبة للفقرة (j) أكبر من قدرة المفحوص (i)، وتساوي (صفر) إذا كانت أقل وبذلك فإن قيمة المعلمة (δ_i) ستختلف من مفحوص لآخر.

المحكّات المستخدمة للحكم على دقة تقدير معالم الفقرة ومعلمة القدرة:

التحيز (Bias): ويُعبّر عن عدم التحيز بالدرجة التي تكون فيها القيم المقدرة لمعالم الفقرة وفق الطريقة المستخدمة قريبة من القيم الحقيقية لتلك المعالم، وكلما اقتربت قيمة هذا المؤشر من الصفر دلّ ذلك على دقة أعلى، إذ يتم إيجاد مؤشر عدم التحيز عن طريق حساب الفرق بين القيم المقدرة والقيم الحقيقية.

الجذر التربيعي لمتوسط مربعات الخطأ (RMSE):

يُقاس هذا المؤشر من خلال إيجاد الجذر التربيعي لمربع الفرق بين القيمة المقدرة والقيمة الحقيقية، لكل من معالم الفقرات أو معلمة القدرة للمفحوصين، وكلما اقتربت قيمة هذا المؤشر من الصفر دلّ ذلك على دقة التقدير.

الكفاءة النسبية (Relative Efficiency)

أشار كل من ين وفتزباترك (Yen and Fitzpatrick, 2006) ان لدالة المعلومات مكانية المقارنة بين الاختبارات المختلفة

التي تقيس القدرة نفسها، أو تحري فعالية نماذج مختلفة استخدمت لتقدير معالم الفقرات أو معلمة القدرة للمفحوصين، أو طرق مختلفة لتصحيح الاختبار ذاته، أو مقارنة استراتيجيات مختلفة استخدمت لبناء الاختبارات مثل المقارنة بين الاختبارات التقليدية والاختبارات التكيفية.

وفيما يلي عرضاً للدراسات التي تناولت دقة تقدير معالم الفقرات من خلال المقارنة بين أكثر من نموذج لوجستي ضمن إطار نظرية الاستجابة للفقرة.

أجرى بيلتون (Pelton, 2002) دراسة هدفت إلى مقارنة الدقة والاستقرار في تقدير معلمة الصعوبة، ومعلمة القدرة باستخدام النظرية الكلاسيكية في القياس، والنماذج في نظرية الاستجابة للفقرة، واستخدم الباحث برمجيتين: (Winsteps) و (Bigsteps)، أظهرت نتائج الدراسة أنه بوجود حجم عينة معتدل (999) مفحوصاً وعدد فقرات ملانم (33) فقرة، فإن النموذج الثنائي المعلمة يُقدم تقديرات أكثر دقة لمعلمة الصعوبة من نموذج راش والنموذج الثلاثي المعلمة والنظرية الكلاسيكية في القياس وذلك في ظل وجود تخمين قليل في الإجابة على فقرات الاختبار.

وأجرى بالدين (Baldwin, 2006) دراسة هدفت إلى المقارنة بين كل من النموذج أحادي المعلمة، والنموذج ثنائي المعلمة من جهة والنموذج الثنائي المعلمة المعدل من جهة أخرى، استخدم الباحث نوعين من البيانات: مولدة وحقيقية، وبأحجام مختلفة بلغ مجموعها (135,000) مفحوص في اختبارات في مبحث الرياضيات، وتألفت الاختبارات من (25 فقرة و 40 فقرة و 50 فقرة) على التوالي، واستخدم برمجية (Multilog) لتحليل البيانات، وأظهرت نتائج الدراسة أن النموذج الثنائي المعدل يُقلل من أخطاء القياس مقارنة بالنموذج الأحادي المعلمة والنموذج الثنائي المعلمة، وبالنسبة لدقة تقديرات معالم الصعوبة فقد أظهرت النتائج أن النموذج الثنائي المعدل أدق في التقدير من النموذج الثنائي والأحادي في كل الحالات عدا حالة واحدة عندما حجم العينة (100) مفحوص، وطول الاختبار (40) فقرة.

كما أجرى كل من وايز وديمارز (Wise and DeMars, 2006) دراسة هدفت إلى مقارنة دقة تقديرات معالم الفقرات ودالة المعلومات، حيث تم استخدام: النموذج المعدل المعتمد على الزمن والنموذج الثنائي المعلمة، وجرى استخدام بيانات مولدة بنسب تخمين (10% و 30% و 50%) واستخدمت ثلاثة أحجام عينات بواقع (500 مفحوص) لكل عينة، واستخدم الباحث برمجيات حاسوبية لاجراء الاختبارات (CBT)، وحُلَّت البيانات حسب النموذج الثلاثي المعلمة وحسب النموذج المعدل المعتمد على الزمن، باستخدام برنامج (BILOG-MG)، أظهرت نتائج الدراسة أن التحيز في تقدير معالم الفقرات ومعلمة القدرة أقل في حالة استخدام النموذج المعدل المعتمد على الزمن، وكلما ارتفعت نسبة التخمين ارتفعت نسبة التحيز في النموذج الثلاثي المعلمة.

كما أجرى عابنة (2004) دراسة هدفت لتفحص أثر حجم العينة وطريقة انتقائها وعدد الفقرات وطريقة انتقائها على دقة تقدير معالم الفقرة ومعلمة القدرة لاختبار قدرة عقلية باستخدام نظرية الاستجابة للفقرة، حيث تم إعداد اختبار قدرة عقلية مكون من (71) فقرة، وتكونت عينة الدراسة من (1000) طالب وطالبة، واستخدم الباحث لتحليل البيانات كل من برنامج Bilog-Mg و Spss وقد توصلت الدراسة إلى أن الدقة في تقدير معالم الفقرة تزداد بزيادة حجم عينة المفحوصين، كما تزداد الدقة في تقدير معلمة الصعوبة ومعلمة القدرة عندما يكون مدى القدرة للمفحوصين متوافقاً مع صعوبة الفقرات، وكذلك تزداد الدقة في تقدير معلمة التخمين عند استخدام عينة من ذوي القدرة المتدنية في معايرة الفقرات.

وفي دراسة قام بها سوبادي (Subedi, 2009) بهدف تحري العلاقة بين التخمين الناجح والسمات الكامنة في نماذج استجابة الفقرة الثنائية، استخدم الباحث في دراسته النماذج الآتية: النموذج ثنائي المعلمة، والنموذج الثلاثي المعلمة، ونموذج (MixIRT-R) (نموذج استجابة الفقرة المختلط للتخمين العشوائي) الذي يفترض أن سلوك التخمين الذي يقوم به الطلبة هو تخمين عشوائي، ونموذج (MixIRT-A) (نموذج استجابة الفقرة المختلط للتخمين المبني على القدرة)، واستخدم الباحث برنامج (Winbug) لتحليل البيانات، وولدت البيانات بأحجام مختلفة، وخلصت هذه الدراسة إلى أن النماذج المختلطة المستخدمة أعطت تقديرات أدق لمعالم الفقرة مقارنة مع كل من النموذج ثنائي المعلمة والنموذج الثلاثي المعلمة.

وأجرى فو (Fu, 2010) بدراسة سعت إلى مقارنة الدقة في تقدير معلمة القدرة، ومعلمة الصعوبة باستخدام نماذج نظرية الاستجابة للفقرة، في ظل الظروف التالية: وجود التخمين على فقرات اختبار الاختيار من متعدد، وعدم وجود توافق بين قدرة المفحوصين وصعوبة الفقرة وباختلاف كل من حجم العينة وطول الاختبار. واستخدمت النماذج: الثلاثي المعلمة وراش، وثلاثة نماذج معدلة من نموذج راش الأول هي: نموذج راش بالإضافة لمعلمة تخمين ثابتة ($c=0.20$)، والثاني: نموذج راش مع نقطة قطع لوغريتمية ($Cutlo=1$)، والثالث مع نقطة قطع لوغريتمية ($cutlo=2$)، وتم توليد البيانات لثلاث عينات مختلفة الحجم

($n=200$, $n=50$, $n=1000$) وبأطوال اختبارات مختلفة ($k=20$, $k=40$)، استخدم الباحث برمجيتين هما: (Parascale) و (Winstep) أظهرت نتائج الدراسة أنه في حالة النموذج الثلاثي المعلمة ونموذج تثبيت معلمة التخم ($c=0$) تكون قيم الجذر التربيعي لمتوسط مربع الانحرافات لتقدير معلمة القدرة أقل منها للنماذج الأخرى، وتبين أنه بناءً على معاملات الارتباط بين الجذر التربيعي لمتوسط مربع الانحرافات ومتوسط مربع الانحرافات (RMSD) و (MSD) فإن النموذج الثلاثي المعلمة ونموذج تثبيت معلمة التخم ($c=0$) هما الخيار الأفضل.

كما قام جاو (Gao, 2011) بدراسة هدفت لمقارنة دقة التقدير للنموذج ثلاثي المعلمة ونموذج (راش للتخمين)، والنموذج (الثلاثي للتخمين) اللذان يربطان بين احتمالية إجابة الفقرة إجابة صائبة مع معلمة القدرة للمفحوص ومعلم الفقرة، واستخدم الباحث برمجية (SAS)، أظهر النموذج الثلاثي المعلمة تقديرًا أفضل لمعلمة الصعوبة من النماذج الأخرى المستخدمة، كما أظهرت النتائج أنه في العينات الصغيرة الحجم من المفحوصين فإن دقة التقدير للنماذج غير المعدلة أفضل من تلك المعدلة، وبزيادة حجم العينة تزداد الدقة في تقدير معالم الفقرات، وأن النموذج الثلاثي المعلمة أفضل من حيث مطابقة البيانات ودقة تقدير معالم الفقرات. وفي دراسة أجراها الحاج (2011) هدفت إلى معرفة أثر التحكم بمعلمة التخمين في نماذج استجابة الفقرة على تقدير معالم الفقرات ومعلمة القدرة، باستخدام النموذج الأحادي المعلمة والنموذج ثنائي المعلمة والنموذج الثلاثي، استخدم الباحث الرزمة الإحصائية (Spss) وجرى استخدام النماذج المعدلة أحادي المعلمة والنموذج ثنائي المعلمة بافتراض ثابت غير صفري، تكون مجتمع الدراسة من طلبة الصف التاسع الأساسي في محافظة إربد، تكونت عينة الدراسة من (747) طالباً، وحللت الاستجابات باستخدام النماذج الآتية: ثلاثة نماذج غير معدلة وستة نماذج معدلة، أشارت النتائج أن النموذج ثلاثي المعلمة كان أكثر توافقاً مع البيانات مقارنة بالنماذج أحادية المعلمة المعدلة والنماذج ثنائية المعلمة المعدلة.

كما قام الشرفين (2012) بدراسة هدفت إلى الكشف عن أثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة، والخصائص السيكمترية للاختبار، في ضوء تغير حجم العينة، ولتحقيق هدف الدراسة بُني اختبار تحصيلي في الفيزياء من نوع الاختيار من متعدد بأربعة بدائل، تكون بصورته النهائية من (33) فقرة، وطُبق الاختبار على عينة الدراسة المكونة من (1000) طالب وطالبة، واستخدم الباحث كل من برنامج Bilog-MG و Spss وبينت النتائج وجود فروق ذات دلالة إحصائية عند مستوى الدلالة ($\alpha=0.05$) في متوسطات الأخطاء المعيارية لتقديرات معالم الفقرات تُعزى للتفاعل بين طريقة التقدير وحجم العينة، في حين لم تظهر فروق ذات دلالة إحصائية تُعزى لمتغير حجم العينة أو طريقة التقدير. يُمكن استخلاص بعض الاتجاهات في النتائج التي أمكن التوصل إليها والمتعلقة بمعالجة التخمين ودقة تقدير معالم الفقرات، في الاختبارات الموضوعية/الاختيار من متعدد كما يأتي:

إن دقة تقدير معالم الفقرات والأفراد تتغير بشكل واضح تبعاً للطريقة المستخدمة في التقدير والبرنامج الحاسوبي المستخدم أيضاً، بعض الدراسات السابقة استخدمت أكثر من برنامج حاسوبي مع كل نموذج، مما يعني وجود اختلافات تُعزى لاختلاف البرامج الحاسوبية المستخدمة لتقدير المعالم في الدراسة الواحدة، بينما استخدمت الدراسة الحالية برنامجاً حاسوبياً واحداً هو برنامج (Winbug).

أن النماذج التي تناولتها هذه الدراسة لم تتناولها أية دراسة أخرى مجتمعة، وهذا ما ميّز هذه الدراسة عن غيرها من الدراسات السابقة.

ويمكن القول إن هذه الدراسة تتفق مع الدراسات السابقة في سعيها لتحري دقة تقدير معالم الفقرة وفي بعض النماذج المستخدمة، إلا أنها تُجري مقارنة أوسع مما هو موجود في الدراسات السابقة، سواء من ناحية عدد النماذج المشمولة بالدراسة، أو تنوع ظروف التجربة من حيث عدد الفقرات وحجم العينة، كما تختلف خصائص البيانات المولدة في هذه الدراسة عن تلك المولدة في الدراسات السابقة.

مشكلة الدراسة

تُعد دقة تقدير معالم الفقرات من القضايا الهامة في نظرية الاستجابة للفقرة، حيث تعتمد هذه الدقة على كثير من العوامل التي اهتمّ البحث السيكمترى بدراستها، إذ تباينت الآراء حول العوامل المؤثرة على دقة التقدير. فمنهم من يرى أن لحجم عينة المفحوصين أثراً أكبر من طول الاختبار، بالإضافة إلى أن دقة التقديرات تختلف من نموذج لآخر، وبالتالي تكمن مشكلة الدراسة في الكشف عن مدى فعالية أربعة نماذج لوجستية في إطار نظرية الاستجابة للفقرة في إعطاء تقديرات دقيقة لمعلم الفقرات

(الصعوبة والتمييز والتخمين)، باختلاف متغير طول الاختبار (25 فقرة، و 50 فقرة) وعدد المفحوصين (500 مفحوص و 1500 مفحوص). وبشكل أكثر تحديداً هدفت هذه الدراسة إلى الإجابة عن الأسئلة الآتية:

السؤال الأول

هل تختلف دقة تقدير معالم الفقرات باختلاف النموذج المستخدم في إطار النموذج الثلاثي المعلمة تبعاً لمتغيري: طول الاختبار (25 فقرة و 50 فقرة)، وعدد المفحوصين (500 مفحوص و 1500 مفحوص)؟

السؤال الثاني

هل يختلف مؤشر الكفاءة النسبية لتقدير معالم الفقرات باختلاف متغيرات الدراسة (النموذج المستخدم وعدد الفقرات وحجم العينة)؟

أهمية الدراسة

تكمن أهمية هذه الدراسة في الجانبين النظري والتطبيقي كما يأتي:

الأهمية النظرية: محاولتها توفير معلومات حول فاعلية أربعة نماذج في دقة تقدير معالم الفقرات، وخاصة النماذج المختلطة (Mixture Model)، مما سيسهم في استقطاب اهتمام الباحثين لإجراء مزيد من الدراسات حول هذا الموضوع، وبالتالي إغناء البحث العلمي في هذا المجال.

الأهمية العملية: توظيف الطريقة البييزية (Bayesian) في تقدير معالم الفقرات ومعلمة القدرة للمفحوصين التي تتميز عن طريقة الأرجحية العظمى؛ بعدم إعطاء أخطاء عندما تكون إجابة المفحوص جميعها خطأ أو جميعها صحيحة (Hambelton, 1989). وتوجيه اهتمام الباحثين إلى برنامج (Winbug) المتاح مجاناً عبر الشبكة العنكبوتية والمزايا البحثية التي يوفرها، والنماذج المختلطة، الأمر الذي سيتيح للباحثين إمكانية التعامل مع هذا البرنامج والتعرف إلى إمكانياته ومزاياه البحثية، توجيه اهتمام الباحثين إلى النموذج الذي يُعطي دقة تقدير أفضل لمعالم الفقرات لاستخدامه في الاختبارات النفسية والتربوية.

مصطلحات الدراسة وتعريفاتها الإجرائية

سيكون للمفاهيم والمتغيرات المعاني الآتية في إطار هذه الدراسة:

- **معالم الفقرات (Item Parameter):**

تم استخدام مفهوم معالم الفقرات في إطار هذه الدراسة لتشير إلى كل من: معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين.

- **معلمة صعوبة الفقرة:**

هي النقطة على متصل القدرة ويكون احتمال إجابة الفقرة عندها إجابة صحيحة مساوياً لـ $(C_i + 1)$ مقسوماً على 2 ، حيث (C_i) هي معلمة التخمين (Baker, 2004).

معلمة تمييز الفقرة:

هي ميل منحني خصائص الفقرة الذي يقابل النقطة التي تكون عندها معلمة القدرة تساوي معلمة صعوبة الفقرة (Hambleton, 1991). Swaminthan., and Rogers,

معلمة التخمين الفقرة:

هي احتمال حصول المفحوص ذي القدرة المتدنية على الإجابة الصحيحة بالصدفة (Baker, 2004)

- **النموذج ثلاثي المعلمة (3PL) Three-Parameter Logistic Model:**

هو النموذج الذي يتحدد به الأداء على الفقرة بناءً على جميع معالم الفقرة: معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين (Embretson, and Reise, 2000).

- **نموذج عتبة التخمين IRT Threshold Guessing Model:**

أحد أشكال النماذج المختلطة التي تدمج بين النموذج الثنائي المعلمة ومعاملات رياضية إضافية، فإذا سلك المفحوص سلوك التخمين خلال إجابته عن فقرة الاختبار يعمل النموذج المختلط، وإذا لم يُخمن يوظف النموذج الثنائي المعلمة، ويحتوي النموذج على معلمة تُحدد لكل فرد العتبة التي يُخمن عندها (Cao and Stokes, 2008).

- **نموذج التخمين القائم على الصعوبة IRT Difficulty-Based Guessing Model:**

هو شكل من أشكال النماذج المختلطة التي تدمج بين النموذج الثنائي المعلمة ومعاملات رياضية إضافية، ويفترض أن المفحوص يجيب عن الفقرات السهلة أولاً، ثم يستخدم التخمين للإجابة عن باقي الفقرات التي مستوى صعوبتها أعلى من مستوى

قدرته (Cao and Stokes, 2008) ..

- النموذج الثنائي المعدل:

هو النموذج الثنائي المعلمة المعدل إلى ثلاثي من خلال إضافة قيمة ثابتة عوضاً عن معلمة التخمين تبلغ 0.20 (Sireci, 1992).

- الكفاءة النسبية Relative Efficiency :

جرى تقدير الكفاءة النسبية باختلاف متغيرات الدراسة من: حجم عينة وعدد فقرات الاختبار والنموذج المستخدم في الدراسة من خلال قسمة تباين معالم الفقرة ومعلمة القدرة التي تم تقديرها من خلال نموذج معين على تباين نظيراتها المقدرة من خلال نموذج آخر فإذا زاد ناتج القسمة عن الواحد الصحيح تكون الكفاءة النسبية لصالح النموذج الأول (علام، 2005).

- الجذر التربيعي لمتوسط مربع الخطأ (RMSE) Root Mean Square Error :

مؤشر لدقة تقدير معالم الفقرات، ويتم حسابه من خلال إيجاد الجذر التربيعي لمتوسط مربعات الفروق بين القيم المولدة (الحقيقية) والقيم المقدرة (Hambleton. and Jones, 1993).

- التحيز (Bias):

تم النظر للتحيز في هذه الدراسة بأنه مؤشر يُعبر عن الدرجة التي تكون فيها القيم المقدرة لمعالم الفقرات وفق النموذج المستخدم في هذه الدراسة قريبة من القيم الحقيقية (Weiss, and Babcock, 2009).

حدود الدراسة ومحدداتها

تتحدد نتائج هذه الدراسة بالمحددات الآتية:

أ- البيانات التي تم استخدامها في إطار هذه الدراسة: بيانات مولدة (محاكاة) لاختبارات موضوعية /اختبار من متعدد بافتراض أربعة بدائل للفقرة، وجرى توليدها باستخدام برنامج (WinGen3) .

ب- النماذج التي وظفت في إطار هذه الدراسة هي:

النموذج الثلاثي المعلمة، والنموذج الثنائي المعدل إلى ثلاثي بإضافة قيمة ثابتة كمعلمة تخمين تساوي (0.20)، ونموذج عتبة التخمين، والنموذج القائم على الصعوبة.

ج- تقدير معالم الفقرات ومعلمة القدرة للمفحوصين باستخدام برنامج: (Winbug)

(Windows Bayesian Using Gibbs Sampling) وباستخدام التقدير البييزي.

منهجية الدراسة

إجراءات توليد البيانات

تم توليد البيانات التي تمثل استجابات المفحوصين باستخدام برنامج ون جن (WinGen) وذلك وفقاً للمتغيرات الخاصة بالدراسة معالم الفقرات (معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين) وطول الاختبار وعدد المفحوصين، وذلك وفقاً للخطوات الآتية:

أ- تحديد عدد الفقرات بحيث يتوافر للباحث اختباران: الأول (أ) بطول (25) فقرة، والاختبار الثاني (ب) بطول (50) فقرة، ويعود سبب اختيار الباحث لهذين الطولين إلى أن الاختبارات النفسية والتربوية عادة ما تستخدم اختبارات بطول (25) فقرة كالاختبارات القصيرة وأخرى بطول (50) فقرة كاختبارات متوسطة (Dawber, Roger and Carbonar, 2009).

ب- توليد استجابات على كل اختبار من الاختبارين السابقين بحيث يكون عدد أفراد العينة الأولى (1500) مفحوص وعدد أفراد العينة الثانية (500) مفحوص ويُعزى اختبار هذين الحجمين المتفاوتين بهدف تحري أثر حجم العينة على معالجة التخمين، وبالتالي تم اختيار عينة متوسطة الحجم، وعينة أخرى كبيرة الحجم. وفي هذا السياق يُشير كل من هامبلتون وكوك (Hambleton and Cook, 1983) إلى أن حجم عينة يبلغ (500) مفحوص يُعتبر ملائماً لإعطاء تقديرات مستقرة لقدرات المفحوصين ومعالم الفقرات في النموذج الثلاثي المعلمة.

كما روعي عند اختيار خصائص معالم الفقرات قبل عملية توليد البيانات أن تكون مطابقة لافتراضات ومتطلبات النموذج الثلاثي المعلمة، بحيث تكون خصائص معالم الفقرات كما يأتي:

- 1- معلمة التمييز: تتوزع توزيعاً منتظماً بقيمة دنيا تساوي (0.40) وقيمة عليا تساوي (1.7) لوجت.
- 2- معلمة الصعوبة: تتوزع توزيعاً طبيعياً بمتوسط حسابي (صفر) لوجت وانحراف معياري (1) لوجت.

3- معلمة التخمين: تتوزع توزيعاً منتظماً بقيمة دنيا تساوي (0.05) وقيمة عليا تساوي (0.40).

تم تخزين كل من معالم الفقرات ومعلمة القدرة للمفحوصين واستجاباتهم على فقرات الاختبارات المولدة في ملفات نصية منفصلة، حيث تم استدعاؤها من خلال الرزمة الإحصائية للعلوم الاجتماعية (SPSS) وتم التعامل مع هذه الملفات على أنها تحوي المعالم الحقيقية للفقرات والمفحوصين.

تقدير معالم الفقرة

تم في هذه المرحلة استدعاء استجابات المفحوصين التي تم توليدها باستخدام برنامج ون جن، وتم استخدام برنامج (Winbug) بهدف تقدير معالم الفقرات وفق كل نموذج من النماذج التي جرى توظيفها في إطار هذه الدراسة وهي: النموذج الثلاثي المعلمة، والنموذج الثنائي المعدل إلى ثلاثي، ونموذج عتبة التخمين والنموذج القائم على الصعوبة، وباختلاف كل من عدد فقرات الاختبار (25 فقرة و 50 فقرة) وحجم العينة (500 مفحوص و 1500 مفحوص).

الإجراءات الإحصائية

استخدمت الإجراءات الإحصائية الآتية للإجابة عن كل سؤال من أسئلة الدراسة :

السؤال الأول: جرى استخدام محكين للدقة: الأول الجذر التربيعي لمتوسط مربعات الخطأ والثاني مؤشر التحيز، وجرى فحص الاختلافات الملحوظة بين متوسطات الجذر التربيعي لمتوسط مربعات الخطأ ومؤشر التحيز باستخدام تحليل التباين الثلاثي للقياسات المتكررة متبعاً باختبار أقل فرق دال للمقارنات البعدية.

للإجابة عن سؤال الدراسة الثاني:

تم حساب التباين للقيم المقدرة وفقاً لكل نموذج من النماذج المستخدمة في هذه الدراسة في المواقع المختلفة للمتغيرات (طول الاختبار وحجم عينة البيانات المولدة) على تباين النماذج الأخرى لكل معلم من معالم الفقرات، فإذا زادت هذه القيمة عن الواحد الصحيح تكون الأفضلية للاختبار أو النموذج (a) وإذا قلّت عن الواحد تكون الأفضلية للاختبار أو النموذج (b).

عرض نتائج الدراسة**النتائج المتعلقة بالإجابة عن سؤال الدراسة الأول:**

هل تختلف دقة تقدير معالم الفقرات (معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين) باختلاف النموذج المستخدم تبعاً لمتغيري طول الاختبار (25 فقرة و 50 فقرة) وعدد المفحوصين (500 مفحوص و 1500 مفحوص).

للإجابة عن هذا السؤال تم إيجاد الجذر التربيعي لمتوسط مربع الخطأ لمعلم الفقرة، حيث تزداد دقة التقديرات كلما اقتربت قيمة هذا المؤشر من الصفر (Swaminathan, Hambleton, Xing, and Rizavi, 2003).

يُبين الجدول (1) المتوسطات الحسابية والانحرافات المعيارية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز باختلاف النموذج المستخدم وحجم عينة المفحوصين (500 مفحوص و 1500 مفحوص) وطول الاختبار (25 فقرة و 50 فقرة).

الجدول (1)

المتوسطات الحسابية والانحرافات المعيارية لمؤشر الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز في حالة النماذج الأربعة المختلفة المستخدمة وتبعاً لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

طول الاختبار				النموذج	حجم العينة
50 فقرة		25 فقرة			
الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي		
0.095	0.060	0.184	0.100	الثلاثي	500 مفحوص
0.254	0.122	0.269	0.126	الثنائي المعدل	
0.308	0.167	0.279	0.173	عتبة التخمين	
0.269	0.165	0.634	0.312	القائم على الصعوبة	
0.124	0.073	0.218	0.124	الثلاثي	1500

حجم العينة	النموذج	طول الاختبار			
		25 فقرة		50 فقرة	
مفحوص	الثنائي المعدل	المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري
		0.123	0.202	0.116	0.172
	عتبة التخمين	0.148	0.228	0.142	0.241
		0.152	0.227	0.130	0.199
القائم على الصعوبة	على				

يُلاحظ من نتائج الجدول (1) أن أقل قيمة للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز قد بلغت (0.060) في حالة النموذج الثلاثي المعلمة وعند حجم عينة (500) مفحوص وطول اختبار (50) فقرة ، يليه أيضا النموذج الثلاثي المعلمة بقيمة بلغت (0.073) عند حجم عينة (1500) مفحوص وطول اختبار (50) فقرة.

وللوقوف على أن التغييرات في قيم متوسطات الجذر التربيعي لمربعات الخطأ في تقدير معلمة التمييز تختلف تبعاً لمتغيرات الدراسة: نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المولدة لهم معلمة التمييز والتفاعلات بينها، جرى استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز، والجدول (2) يُبين ملخص نتائج هذا التحليل.

الجدول (2)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم العينة والتفاعلات بينها على متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز للفقرات.

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجات الحرية	متوسط المربعات	ف	الدلالة الإحصائية
داخل الفقرات	النموذج	0.762	3	0.254	6.645	*0.000
	النموذج* عدد الفقرات	0.141	3	0.047	1.226	0.300
	النموذج* حجم العينة	0.255	3	0.085	2.224	0.085
	النموذج* عدد الفقرات* حجم العينة	0.107	3	0.036	0.930	0.426
	الخطأ	16.748	438	0.038		
بين الفقرات	عدد الفقرات	0.100	1	0.100	0.186	0.667
	حجم العينة	0.575	1	0.575	1.072	0.302
	عدد الفقرات × حجم العينة	0.079	1	0.079	0.148	0.701
	الخطأ	78.335	146	0.537		

يُلاحظ من نتائج الجدول (2) عدم وجود أثرٍ للتفاعلات الثلاثية والثنائية لمتغيرات الدراسة لأثر النموذج المستخدم ، وعدد الفقرات المولد لهم بيانات على تقدير الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز حيث كانت جميع قيم (ف) ليس لها دلالة إحصائية عند مستوى الدلالة ($\alpha=0.05$)، بينما يُلاحظ وجود أثرٍ دالٍ إحصائياً يُعزى للنموذج المستخدم، حيث كانت قيم الدلالة الخاصة بـ(ف) أقل من (0.05)، أما بالنسبة للتباين بين الأفراد، تُشير نتائج الجدول إلى عدم وجود أثرٍ للتفاعل الثنائي ولعدد الفقرات ولعدد المفحوصين المُولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائياً عند مستوى الدلالة ($\alpha=0.05$)، ولمعرفة أي الفروق بين متوسطات قيم الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز تبعاً للنماذج المستخدمة كان له تأثير دالٍ إحصائياً فقد تم استخراج نتائج اختبار أقل فارق دال (LSD) التي تبدو في الجدول (3).

الجدول (3)

نتائج اختبار أقل فرق دال (LSD) للمقارنات الثنائية بين متوسطات الجذر التربيعي لمتوسط مربع الخطأ لتقدير معلمة التمييز باختلاف النموذج المستخدم .

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (2-1)	مستوى الدلالة
الثلاثي المعلمة	الثنائي المعدل	*-0.091	*0.007
	عتبة التخمين	*-0.099	*0.000
	القائم على الصعوبة	-0.222	*0.000
الثنائي المعدل	الثلاثي المعلمة	0.091	*0.007
	عتبة التخمين	-0.008	0.809
	القائم على الصعوبة	-0.131	*0.010
عتبة التخمين	الثلاثي المعلمة	0.099	*0.000
	الثنائي المعدل	0.008	0.809
	القائم على الصعوبة	-0.123	*0.005
القائم على الصعوبة	الثلاثي المعلمة	0.222	*0.000
	الثنائي المعدل	0.131	*0.010
	عتبة التخمين	0.123	*0.005

يُلاحظ من نتائج الجدول (3) أن الفروق بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز، كانت دالة إحصائياً في حالة استخدام النموذج الثلاثي المعلمة مقارنة مع كل من النماذج الأخرى المستخدمة (الثنائي المعدل وعتبة التخمين والقائم على الصعوبة)، حيث كانت قيمة مستوى الدلالة أقل من (0.05) في كل حالة؛ الأمر الذي يُشير إلى أن النموذج الثلاثي المعلمة أدق في تقدير معلمة التمييز من النماذج الأخرى. كما ويُلاحظ أن الفروق بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التمييز كانت دالة إحصائياً عند مستوى الدلالة (0.05) في حالة مقارنة النموذج الثلاثي المعدل مع النموذج القائم على الصعوبة؛ مما يُشير إلى أن النموذج الثلاثي المعدل أدق من هذا النموذج في تقدير معلمة التمييز، وكذلك تُشير النتائج إلى أنه عند مقارنة المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ لنموذج عتبة التخمين مع المتوسطات الحسابية نفسها الخاصة بالنموذج القائم على الصعوبة كانت الفروق دالة إحصائياً، الأمر الذي يُشير إلى أن نموذج عتبة التخمين أدق من النموذج القائم على الصعوبة في تقدير معلمة التمييز. واستكمالاً للإجابة عن السؤال الأول، تم حساب المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التمييز باختلاف كل من حجم العينة والنموذج المستخدم وطول الاختبار والجدول (4) يُبين نتائج هذا المؤشر الإحصائي.

الجدول (4)

المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التمييز والانحرافات المعيارية المناظرة لها في حالة النماذج الأربعة المختلفة المستخدمة وتبعاً لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

حجم العينة	النموذج	طول الاختبار			
		25 فقرة		50 فقرة	
		متوسط التحيز	الانحراف المعياري	متوسط التحيز	الانحراف المعياري
500 مفحوص	الثلاثي	0.034	0.017	0.011	0.012
	الثنائي المعدل	0.049	0.024	0.021	0.017
	عتبة التخمين	0.041	0.031	0.039	0.022

طول الاختبار				النموذج	حجم العينة
50 فقرة		25 فقرة			
الانحراف المعياري	متوسط التحيز	الانحراف المعياري	متوسط التحيز		
0.026	0.049	0.037	0.057	القائم على الصعوبة	1500 مفحوص
0.012	0.004	0.017	0.030	الثلاثي	
0.017	0.011	0.024	0.035	الثنائي المعدل	
0.022	0.034	0.031	0.039	عتبة التخمين	
0.026	0.030	0.037	0.043	القائم على الصعوبة	

يُلاحظ من نتائج الجدول (4) عند الانتقال إلى العينة المكونة من (1500) مفحوص الانخفاض الظاهري لجميع قيم المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التمييز مفحوص وبغض النظر عن طول الاختبار، ففي حالة النموذج الثلاثي المعلمة بلغ المتوسط الحسابي (0.017)، يليه النموذج الثنائي المعدل بقيمة بلغت (0.023)، ثم نموذج عتبة التخمين بقيمة بلغت (0.036)، وأخيرًا النموذج القائم على الصعوبة بقيمة بلغت (0.037).

وللوقوف على أن التغييرات في قيم متوسطات مؤشر التحيز في تقدير معلمة التمييز تختلف تبعًا لمتغيرات الدراسة نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المولدة لهم معلمة التمييز والتفاعلات بينها، تم استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ لتقدير معلمة التمييز، والجدول (5) يبين ملخص نتائج هذا التحليل.

الجدول (5)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم العينة والتفاعلات بينها على متوسطات مؤشر التحيز في تقدير معلمة التمييز.

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجات الحرية	متوسط المربعات	ف	الدلالة الإحصائية
داخل الفقرات	النموذج.	0.610	3	0.203	6.011	*0.001
	النموذج* عدد الفقرات.	0.063	3	0.021	0.611	0.601
	النموذج* حجم العينة.	0.033	3	0.051	0.322	0.814
	النموذج* عدد الفقرات* حجم العينة.	0.009	3	0.003	0.088	0.967
	الخطأ	14.814	438	0.034		
بين الفقرات	عدد الفقرات	0.100	1	0.100	0.908	0.342
	حجم العينة	0.029	1	0.029	0.258	0.612
	عدد الفقرات × حجم العينة	0.012	1	0.012	0.112	0.738
	الخطأ	16.156	146	0.111		

يُلاحظ من نتائج الجدول (5) عدم وجود أثرٍ للتفاعلات الثلاثية أو الثنائية لمتغيرات الدراسة النموذج المستخدم وعدد الفقرات وعدد المفحوصين المولد لهم بيانات على متوسطات مؤشر التحيز في تقدير معلمة التمييز حيث كانت جميع قيم (ف) لها ليست دالة إحصائيًا، بينما يُلاحظ أن هناك أثرًا دالاً إحصائيًا يُعزى للنموذج المستخدم على متوسطات مؤشر التحيز في تقدير معلمة الصعوبة فقد كانت قيمة (ف) دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$).

وبالنسبة للتباين بين الأفراد، تُشير نتائج الجدول إلى عدم وجود أثرٍ للتفاعل الثنائي ولعدد الفقرات أو لحجم عينة المفحوصين المولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$)، ولمعرفة أي النماذج المستخدمة كان له تأثير دال إحصائيًا على متوسطات مؤشر التحيز في تقدير معلمة التمييز، تم استخراج نتائج اختبار أقل فرق دال (LSD) التي تبدو في الجدول الآتي:

الجدول (6)

نتائج اختبار أقل فرق دال (LSD) للمقارنات الثنائية بين متوسطات مؤشر التحيز في تقدير معلمة التمييز وباختلاف النموذج المستخدم .

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (1 - 2)	مستوى الدلالة
الثلاثي المعلمة	الثنائي المعدل	-0.039	0.076
	عتبة التخمين	-0.090	*0.000
	القائم على الصعوبة	-0.069	*0.002
الثنائي المعدل	الثلاثي المعلمة	0.039	0.076
	عتبة التخمين	-0.051	*0.020
	القائم على الصعوبة	-0.030	0.223
عتبة التخمين	الثلاثي المعلمة	0.090	*0.000
	الثنائي المعدل	0.051	*0.020
	القائم على الصعوبة	0.021	0.368
القائم على الصعوبة	الثلاثي المعلمة	0.069	*0.002
	الثنائي المعدل	0.030	0.223
	عتبة التخمين	-0.021	0.368

يُلاحظ من نتائج الجدول (6) أن الفروق بين المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التمييز للنماذج المستخدمة كانت دالة إحصائيًا في حالة استخدام النموذج الثلاثي المعلمة مقارنة مع النموذجين (عتبة التخمين والقائم على الصعوبة)؛ الأمر الذي يُشير إلى دقة النموذج الثلاثي المعلمة في تقدير معلمة التمييز مقارنة بهذين النموذجين، وعند مقارنة متوسطات مؤشر التحيز للنموذج الثنائي المعدل مع النماذج الأخرى فإن الفروق دالة إحصائيًا في حالة المقارنة مع نموذج عتبة التخمين فقد كانت مستوى الدلالة أقل من (0.05) الأمر الذي يُشير إلى دقة التقدير للنموذج الثنائي المعلمة المعدل مقارنة مع نموذج عتبة التخمين، وعند مقارنة نموذج عتبة التخمين مع النموذج القائم على الصعوبة فإن حجم متوسطات مؤشر التحيز أعلى في حالة نموذج عتبة التخمين إلا أنها غير دالة إحصائيًا.

الجدول (7)

المتوسطات الحسابية لمؤشر الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة في حالة النماذج الأربعة المختلفة المستخدمة والانحرافات المعيارية المناظرة لها تبعاً لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

حجم العينة	النموذج	طول الاختبار			
		25 فقرة		50 فقرة	
		المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري
500 مفحوص	الثلاثي	0.068	0.013	0.043	0.009
	الثنائي المعدل	0.061	0.012	0.052	0.008
	عتبة التخمين	0.153	0.039	0.148	0.028
	القائم على الصعوبة	0.202	0.044	0.155	0.031
1500 مفحوص	الثلاثي	0.029	0.012	0.016	0.009
	الثنائي المعدل	0.052	0.013	0.046	0.010
	عتبة التخمين	0.146	0.039	0.112	0.028
	القائم على الصعوبة	0.142	0.043	0.132	0.031

يُلاحظ من نتائج الجدول (7) عند الانتقال إلى العينة المكونة من (1500) مفحوص وبغض النظر عن طول الاختبار، يُلاحظ الانخفاض الظاهري لجميع قيم المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة، ففي حالة النموذج الثلاثي المعلمة بلغ المتوسط الحسابي (0.022)، يليه النموذج الثنائي المعدل بقيمة بلغت (0.047)، ثم نموذج عتبة التخمين بقيمة بلغت (0.129)، وأخيراً النموذج القائم على الصعوبة بقيمة بلغت (0.137) وللوقوف على أن التغييرات في قيم متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة تختلف تبعاً لمتغيرات الدراسة: نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المولد لهم معلمة الصعوبة والتفاعلات بينها، جرى استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على المتوسطات الحسابية للجذر التربيعي لمتوسط مربعات الخطأ في تقدير معلمة الصعوبة، والجدول (8) يبين ملخص نتائج هذا التحليل.

جدول (8)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم العينة والتفاعلات بينها على متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة للفقرات.

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	ف	الدلالة الإحصائية
داخل الفقرات	النموذج	26.778	3	8.926	16.317	*0.000
	النموذج* عدد الفقرات	1.395	3	0.465	0.850	0.467
	النموذج* حجم العينة	0.555	3	0.185	0.338	0.798
	النموذج* عدد الفقرات* حجم العينة	0.171	3	0.057	0.104	0.958
	الخطأ	239.596	438	0.547		
بين الفقرات	عدد الفقرات	6.920	1	6.920	2.830	0.095
	حجم العينة	1.387	1	1.387	0.567	0.453
	عدد الفقرات × حجم العينة	0.622	1	0.622	0.254	0.616

		2.445	146	356.980	الخطأ	
--	--	-------	-----	---------	-------	--

يُلاحظ من الجدول (8) عدم وجود أثرٍ للتفاعلات الثلاثية والثنائية لمتغيرات الدراسة النموذج المستخدم وعدد الفقرات وعدد المفحوصين المولد لهم بيانات على متوسطات الجذر التربيعي لمتوسط مربع الخطأ لتقدير معلمة الصعوبة، حيث كانت جميع قيم (ف) لها ليست دالة إحصائية، بينما يُلاحظ أن هناك أثرًا دالاً إحصائياً يُعزى للنموذج المستخدم على المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة فقد كانت قيمة (ف) دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$). وبالنسبة للتباين بين الأفراد تُشير نتائج الجدول إلى عدم وجود أثر للتفاعل الثنائي أو لعدد الفقرات أو لحجم عينة المفحوصين المولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$). ولمعرفة أي النماذج المستخدمة كان له تأثير دالٍ إحصائياً على متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة، تم إجراء اختبار أقل فرق دال (LSD) والذي تبدو نتائجه فيما يأتي:

جدول (9)

نتائج اختبار أقل فرق دال (LSD) بين متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة تبعاً للنموذج المستخدم.

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (1-2)	الدلالة الإحصائية
الثلاثي المعلمة	الثنائي المعدل	-0.013	*0.049
	عتبة التخمين	-0.101	*0.000
	القائم على الصعوبة	-0.119	*0.000
الثنائي المعدل	الثلاثي المعلمة	0.013	*0.049
	عتبة التخمين	-0.088	*0.000
	القائم على الصعوبة	-0.106	*0.000
عتبة التخمين	الثلاثي المعلمة	0.101	*0.000
	الثنائي المعدل	0.088	*0.000
	القائم على الصعوبة	0.018	0.147
القائم على الصعوبة	على الثلاثي المعلمة	0.119	*0.000
	الثنائي المعدل	0.106	*0.000
	عتبة التخمين	-0.018	0.147

يُلاحظ من نتائج الجدول (9) وعند مقارنة المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة للنموذج الثلاثي المعلمة بالنماذج الأخرى فإن الفروق دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$) عند كل مقارنة، الأمر الذي يُشير إلى أن النموذج الثلاثي المعلمة أدق في تقدير معلمة الصعوبة مقارنة مع تلك النماذج. وعند مقارنة متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة للنموذج الثنائي المعدل مع كل من النماذج الأخرى المستخدمة، كانت الفروق دالة إحصائية، الأمر الذي يُشير إلى دقة النموذج الثنائي المعدل في تقدير معلمة الصعوبة، وعند مقارنة متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة الصعوبة للنموذج القائم على الصعوبة مع نموذج عتبة التخمين لم تكن الفروق دالة إحصائية.

كما جرى حساب المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة الصعوبة باختلاف كل من حجم العينة والنموذج وطول الاختبار، والجدول (10) يبين نتائج هذا المؤشر الإحصائي.

جدول (10)

المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة الصعوبة والانحرافات المعيارية المناظرة لها في حالة النماذج الأربعة المختلفة المستخدمة تبعاً لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

طول الاختبار				النموذج	حجم العينة
50 فقرة		25 فقرة			
الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي		
0.027	0.037	0.038	0.032	الثلاثي	500 مفحوص
0.031	0.049	0.044	0.057	الثنائي المعدل	
0.052	0.067	0.073	0.069	عتبة التخمين	
0.054	0.070	0.076	0.073	القائم على الصعوبة	
0.024	0.003	0.032	0.012	الثلاثي	1500 مفحوص
0.031	0.046	0.041	0.046	الثنائي المعدل	
0.052	0.060	0.061	0.065	عتبة التخمين	
0.054	0.065	0.068	0.068	القائم على الصعوبة	

يُلاحظ من نتائج الجدول (10) أنه في حالة العينة المكونة من (1500) مفحوص وبغض النظر عن طول الاختبار، يُلاحظ الانخفاض الظاهري لجميع قيم المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة الصعوبة، ففي حالة النموذج الثلاثي المعلمة بلغ المتوسط الحسابي (0.008)، يليه النموذج الثنائي المعدل بقيمة بلغت (0.046)، ثم نموذج عتبة التخمين بقيمة بلغت (0.063)، وأخيراً النموذج القائم على الصعوبة بقيمة بلغت (0.066).

وللوقوف على أن التغييرات في قيم متوسطات مؤشر التحيز في تقدير معلمة الصعوبة تختلف تبعاً لمتغيرات الدراسة: نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المولدة لهم معلمة الصعوبة والتفاعلات بينها جرى استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة الصعوبة والجدول (28) يبين ملخص نتائج هذا التحليل.

جدول (11)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم العينة والتفاعلات بينها على متوسطات مؤشر التحيز في تقدير معلمة الصعوبة للفقرات

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	ف	الدلالة الإحصائية
----------------	--------------	----------------	-------------	----------------	---	-------------------

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	ف	الدالة الإحصائية
داخل الفقرات	النموذج	2.567	3	0.856	35.485	0.000*
	النموذج* عدد الفقرات	0.049	3	0.016	0.680	0.563
	النموذج* حجم العينة	0.076	3	0.025	1.051	0.370
	النموذج* عدد الفقرات* حجم العينة	0.08124	3	0.008	0.328	0.805
	الخطأ	10.563	438	0.024		
بين الفقرات	عدد الفقرات	0.096	1	0.096	1.464	0.228
	حجم العينة	0.181	1	0.181	2.761	0.099
	عدد الفقرات × حجم العينة	0.020	1	0.020	0.301	0.584
	الخطأ	9.565	146		0.066	

يُلاحظ من نتائج الجدول (11) عدم وجود أثر ذي دلالة إحصائية للتفاعلات الثلاثية أو الثنائية لمتغيرات الدراسة وهي النموذج المستخدم وعدد الفقرات وعدد المفحوصين المولد لهم بيانات على متوسطات مؤشر التحيز في تقدير معلمة الصعوبة، فقد كانت قيم (ف) لها ليست دالة إحصائية، بينما يُلاحظ أن هناك أثرًا يُعزى للنموذج المستخدم فقد بلغت قيمة ف (35.485)، وهي دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$). وبالنسبة للتباين بين الافراد تُشير نتائج الجدول إلى عدم وجود أثرٍ للتفاعل الثنائي أو لحجم العينة أو لعدد الفقرات أو لعدد المفحوصين المولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$).

ولمعرفة أي النماذج المستخدمة كان له تأثير دال إحصائية على قيم متوسطات مؤشر التحيز في تقدير معلمة الصعوبة، تم استخراج نتائج اختبار أقل فرق دال (LSD) التي تبدو في الجدول (12).

جدول (12)

نتائج اختبار أقل فرق دال (LSD) للمقارنات الثنائية بين متوسطات مؤشر التحيز في تقدير معلمة الصعوبة تبعًا للنموذج المستخدم.

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (1 - 2)	مستوى الدلالة
الثلاثي المعلمة	الثنائي المعدل	-0.043	*0.002
	عتبة التخمين	-0.155	*0.000
	القائم على الصعوبة	-0.158	*0.000
الثنائي المعدل	الثلاثي المعلمة	0.043	*0.002
	عتبة التخمين	-0.113	*0.000
	القائم على الصعوبة	-0.116	*0.000
عتبة التخمين	الثلاثي المعلمة	0.155	*0.000
	الثنائي المعدل	0.113	*0.000
	القائم على الصعوبة	-0.003	0.894
القائم على الصعوبة	الثلاثي المعلمة	0.158	*0.000
	تثبيت معلمة التخمين	0.116	*0.000
	عتبة التخمين	0.003	0.894

يُلاحظ من نتائج الجدول (12) أن الفروق بين المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة الصعوبة لنماذج معالجة التخمين كانت دالة إحصائيًا في حالة مقارنة النموذج الثلاثي المعلمة مع كل من النموذج الثنائي المعدل ونموذج عتبة التخمين والنموذج القائم على الصعوبة حيث كانت قيمة مستوى الدلالة أقل من (0.05) عند كل مقارنة؛ الأمر الذي يُشير إلى دقة التقدير للنموذج الثلاثي المعلمة مقارنة بهذه النماذج المستخدمة ، وعند مقارنة متوسطات مؤشر التحيز للنموذج الثنائي المعدل مع النماذج الأخرى، وُجد أن الفروق دالة إحصائيًا عند المقارنة مع كل من النموذج القائم على الصعوبة ونموذج عتبة التخمين، الأمر الذي يُشير إلى دقة التقديرات للنموذج الثنائي المعدل في تقدير معلمة الصعوبة للمفحوصين مقارنة بهذين النموذجين، وعند مقارنة نموذج عتبة التخمين مع النموذج القائم على الصعوبة فإن حجم المتوسطات الحسابية لمؤشر التحيز أعلى في حالة النموذج القائم على الصعوبة إلا أنها غير دالة إحصائيًا.

الجدول (13)

المتوسطات الحسابية والانحرافات المعيارية المناظرة لها لمؤشر الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين في حالة النماذج الأربعة المختلفة المستخدمة تبعًا لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

حجم العينة	النموذج	طول الاختبار			
		25 فقرة		50 فقرة	
		المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري
500 مفحوص	الثلاثي	0.077	0.010	0.058	0.012
	الثنائي المعدل	*	*	*	*
	عتبة التخمين	0.098	0.011	0.086	0.008
	القائم على الصعوبة	0.124	0.015	0.078	0.012
1500 مفحوص	الثلاثي	0.057	0.007	0.031	0.006
	الثنائي المعدل	*	*	*	*
	عتبة التخمين	0.086	0.008	0.081	0.007
	القائم على الصعوبة	0.077	0.010	0.068	0.011

* تم تثبيت معلمة التخمين عند القيمة 0.20

يُلاحظ من نتائج الجدول (13) أنه في حالة العينة المكونة من (1500) مفحوص وبغض النظر عن طول الاختبار، يُلاحظ الانخفاض الظاهري لجميع قيم الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين، ففي حالة النموذج الثلاثي المعلمة بلغ المتوسط الحسابي (0.044)، يليه النموذج القائم على الصعوبة بقيمة بلغت (0.073)، وأخيرًا نموذج عتبة التخمين بقيمة بلغت (0.084).

كما يُلاحظ أنه في حالة الاختبار المكون من (25) فقرة وبغض النظر عن حجم عينة المفحوصين، فقد بلغ المتوسط الحسابي للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين (0.067) في حالة النموذج الثلاثي المعلمة، يليه نموذج عتبة

التخمين بقيمة بلغت (0.104)، واخيرًا النموذج القائم على الصعوبة بقيمة بلغت (0.090). وللوقوف على أن التغيرات في قيم متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين تختلف تبعًا لمتغيرات الدراسة: نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المولد لهم معلمة التخمين والتفاعلات بينها، تم استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين والجدول (14) يبين ملخص نتائج هذا التحليل.

جدول (14)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم العينة والتفاعلات بينها على المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين.

الخطأ	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	ف	الدالة الإحصائية
داخل الفقرات	النموذج	0.014	2	0.007	11.460	*0.000
	النموذج* عدد الفقرات	0.003	2	0.001	2.089	0.126
	النموذج* حجم العينة	0.006	2	0.003	4.427	0.023
	النموذج* عدد الفقرات* حجم العينة	0.003	2	0.002	2.426	0.090
	الخطأ	0.182	292	0.001		
بين الفقرات	عدد الفقرات	0.003	1	0.003	2.529	0.114
	حجم العينة	0.0003	1	0.003	0.028	0.867
	عدد الفقرات × حجم العينة	0.000	1	0.000	0.404	0.526
	الخطأ	0.162	146	0.001		

يُلاحظ من الجدول (14)، عدم وجود أثر للتفاعل الثلاثي بين النموذج المستخدم وعدد الفقرات وحجم العينة أو للتفاعل الثنائي بين النموذج المستخدم وعدد الفقرات على متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين حيث كانت جميع قيم (ف) لها ليست دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$)، بينما يُلاحظ وجود أثر للنموذج المستخدم، فقد كانت قيمة (ف) دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$)، وبالنسبة للتباين بين الأفراد تُشير نتائج الجدول إلى عدم وجود أثر للتفاعل الثنائي أو لعدد الفقرات أو لعدد المفحوصين المولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$)، ولمعرفة أي النماذج المستخدمة كان له تأثير دال إحصائيًا على متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين تم استخراج نتائج اختبار أقل فرق دال (LSD) التي تبدو في الجدول (15).

الجدول (15)

نتائج اختبار أقل فرق دال (LSD) للمقارنات الثنائية بين متوسطات الجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين.

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (1 - 2)	الدالة الإحصائية
الثلاثي المعلمة	عتبة التخمين	-0.032	*0.000
	القائم على الصعوبة	-0.034	*0.000
عتبة التخمين	الثلاثي المعلمة	0.032	*0.000
	القائم على الصعوبة	-0.002	0.797
القائم على الصعوبة	الثلاثي المعلمة	0.034	*0.000

عتبة التخمين	0.002	0.797
--------------	-------	-------

يُلاحظ من الجدول (15) وجود فروق ذات دلالة إحصائية بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ في تقدير معلمة التخمين تبعاً للنموذج المستخدم في حالة النموذج الثلاثي المعلمة عند مقارنته مع كل من نموذج عتبة التخمين والنموذج القائم على الصعوبة، إذ كانت قيمة الدلالة أقل من (0.05) عند كل مقارنة، مما يُشير إلى أن دقة النموذج الثلاثي المعلمة في تقدير معلمة التخمين، بينما لم تكن الفروق دالة إحصائية في حالة مقارنة نموذج عتبة التخمين مع النموذج القائم على الصعوبة، واستكمالاً للإجابة عن السؤال الثاني، تم حساب المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التخمين باختلاف كل من حجم العينة والنموذج المستخدم وطول الاختبار والجدول (16) يُبين نتائج هذا المؤشر الإحصائي.

الجدول (16)

المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التخمين والانحرافات المعيارية المناظرة لها في حالة النماذج الأربعة المختلفة المستخدمة تبعاً لحجم عينة البيانات المولدة وعدد فقرات الاختبار.

حجم العينة	النموذج	طول الاختبار			
		25 فقرة		50 فقرة	
		المتوسط الحسابي	الانحراف المعياري	المتوسط الحسابي	الانحراف المعياري
500 مفحوص	الثلاثي	0.010	0.015	0.019	0.010
	التنائي المعدل	*	*	*	*
	عتبة التخمين	0.052	0.026	0.037	0.014
	القائم على الصعوبة	0.058	0.022	0.011	0.016
1500 مفحوص	الثلاثي	0.001	0.015	0.001	0.015
	التنائي المعدل	*	*	*	*
	عتبة التخمين	0.021	0.020	0.016	0.014
	القائم على الصعوبة	0.015	0.002	0.006	0.016

تم تثبيت معلمة التخمين عند القيمة 0.20

يُلاحظ من نتائج الجدول (16) عند الانتقال إلى العينة المكونة من (1500) مفحوص وبغض النظر عن طول الاختبار، يُلاحظ الانخفاض الظاهري لجميع قيم المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التخمين، ففي حالة النموذج الثلاثي المعلمة ونموذج عتبة التخمين بلغ المتوسط الحسابي (0.010)، ثم النموذج القائم على الصعوبة بقيمة بلغت (0.011). وللوقوف على أن التغيرات في قيم متوسطات مؤشر التحيز في تقدير معلمة التخمين تختلف تبعاً لمتغيرات الدراسة: نموذج معالجة التخمين وطول الاختبار وحجم عينة المفحوصين المُولد لهم معلمة التخمين والتفاعلات بينها، تم استخدام نتائج تحليل التباين الثلاثي للقياسات المتكررة لتأثير هذه المتغيرات والتفاعل بينها على متوسطات مؤشر التحيز، والجدول (17) يُبين ملخص نتائج هذا التحليل.

الجدول (17)

ملخص نتائج تحليل التباين الثلاثي للقياسات المتكررة لأثر متغيرات الدراسة: النموذج المستخدم وطول الاختبار وحجم عينة المفحوصين والتفاعلات بينها على متوسطات مؤشر التحيز في تقدير معلمة التخمين.

اختبار الآثار:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	ف	الدالة الإحصائية
داخل الفقرات	النموذج	0.189	2	0.094	7.351	*0.001
	النموذج* عدد الفقرات	0.011	2	0.005	0.408	0.665
	النموذج* حجم العينة	0.001	2	0.000	0.024	0.976
	النموذج* عدد الفقرات* حجم العينة	0.006	2	0.003	0.230	0.795
	الخطأ	3.753	292	0.013		
بين الفقرات	عدد الفقرات	0.029	1	0.029	1.773	0.185
	حجم العينة	0.031	1	0.031	1.858	0.175
	عدد الفقرات × حجم العينة	0.003	1	0.003	0.161	0.689
	الخطأ	2.404	146	0.016		

يُلاحظ من الجدول (17) عدم وجود أثرٍ للتفاعل الثلاثي بين النموذج المستخدم وعدد الفقرات وحجم العينة أو للتفاعل الثنائي بين النموذج المستخدم وعدد الفقرات على متوسطات مؤشر التحيز في تقدير معلمة التخمين حيث كانت جميع قيم (ف) لها ليست دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$)، بينما يُلاحظ وجود أثرٍ للنموذج المستخدم، فقد كانت قيمة (ف) دالة إحصائية. وبالنسبة للتباين بين الأفراد تُشير نتائج الجدول إلى عدم وجود أثرٍ للتفاعل الثنائي ولعدد الفقرات ولعدد المفحوصين المُولد لهم بيانات؛ فقد كانت جميع قيم (ف) لها ليست دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$)، ولمعرفة أي النماذج المستخدمة كان له تأثير دال إحصائية على متوسطات مؤشر التحيز في تقدير معلمة التخمين، تم استخراج نتائج اختبار أقل فرق دال (LSD) التي تبدو في الجدول (18).

الجدول (18)

نتائج اختبار أقل فرق دال (LSD) للمقارنات الثنائية بين متوسطات مؤشر التحيز في تقدير معلمة التخمين، تبعاً للنموذج المستخدم.

النموذج (1)	النموذج (2)	متوسط الفروق بين النموذجين (2 - 1)	مستوى الدلالة
الثلاثي المعلمة	عتبة التخمين	-0.034	*0.000
	القائم على الصعوبة	-0.018	0.057
عتبة التخمين	الثلاثي المعلمة	0.034	*0.000
	القائم على الصعوبة	0.017	0.053
القائم على الصعوبة	الثلاثي المعلمة	0.018	0.057
	عتبة التخمين	-0.017	0.053

يُلاحظ من نتائج الجدول (18) أن الفروق بين المتوسطات الحسابية لمؤشر التحيز في تقدير معلمة التخمين كانت دالة إحصائية في حالة استخدام النموذج الثلاثي المعلمة مع نموذج عتبة التخمين حيث كانت قيمة (ف) دالة إحصائية عند مستوى الدلالة ($\alpha=0.05$)، وعند المقارنة مع النموذج القائم على الصعوبة كانت قيم متوسطات التحيز في تقدير معلمة التخمين أعلى منها في حالة النموذج القائم على الصعوبة إلا أنها لم تكن دالة إحصائية، الأمر الذي يُشير إلى دقة النموذج الثلاثي المعلمة مقارنة بهذه النماذج المستخدمة، وعند مقارنة متوسطات مؤشر التحيز في تقدير معلمة التخمين لنموذج عتبة التخمين مع النموذج القائم

على الصعوبة فإن حجم متوسطات مؤشر التحيز أعلى منها في حالة النموذج القائم على الصعوبة إلا أنها غير دالة إحصائيًا عند مستوى الدلالة ($\alpha=0.05$).

النتائج المتعلقة بالإجابة عن سؤال الدراسة الثاني:

ما أثر النموذج المستخدم في الكفاءة النسبية لتقدير معالم الفقرة (الصعوبة والتمييز والتخمين) في ضوء اختلاف عدد الفقرات، حجم العينة؟ للإجابة عن هذا السؤال تم قسمة تباين القيم المقدرة لمعالم الفقرات (معلمة الصعوبة ومعلمة التمييز ومعلمة التخمين) لكل نموذج على تباين القيم المقدرة لنفس المعالم باستخدام النماذج الأخرى، والجدول (19) يبين ذلك.

الجدول (19)

قيم الكفاءة النسبية لمعالم الفقرة (التمييز، والصعوبة، والتخمين) وفقاً للنموذج المستخدم وباختلاف طول الاختبار وحجم العينة

طول الاختبار	النموذج	الخطأ المعياري للتقدير لمعلمة التمييز		الخطأ المعياري للتقدير لمعلمة الصعوبة		الخطأ المعياري للتقدير لمعلمة التخمين	
		500 مفحوص	500 مفحوص	500 مفحوص	500 مفحوص	1500 مفحوص	1500 مفحوص
25 فقرة	الثلاثي	0.095	0.120	0.060	0.026	0.076	0.058
	الثنائي المعدل	0.116	0.118	0.022	0.024	-	-
	عتبة التخمين	0.168	0.143	0.137	0.130	0.083	0.083
	القائم على الصعوبة	0.307	0.146	0.188	0.125	0.109	0.075
	الكفاءة النسبية لـ الثلاثي على الثنائي المعدل	1.49	0.97	0.13	0.85	-	-
	الكفاءة النسبية لـ الثلاثي على عينة التخمين	3.13	1.42	5.21	25	1.19	2.05
	الكفاءة النسبية لـ الثلاثي على القائم على الصعوبة	10.44	1.48	9.82	23.11	2.06	1.67
	الكفاءة النسبية لـ الثنائي المعدل على عتبة التخمين	2.10	1.47	38.78	29.34	-	-
	الكفاءة النسبية لـ الثنائي المعدل على القائم على الصعوبة	7.00	1.53	73.02	27.13	-	-
	الكفاءة النسبية لـ عتبة التخمين على القائم على الصعوبة	3.34	1.04	1.88	0.92	1.72	0.82
50 فقرة	الثلاثي	0.059	0.073	0.022	0.016	0.54	0.030
	الثنائي المعدل	0.120	0.115	0.020	0.020	-	-
	عتبة التخمين	0.162	0.137	0.132	0.094	0.078	0.079
	القائم على الصعوبة	0.157	0.133	0.138	0.115	0.077	0.068
	الكفاءة النسبية لـ الثلاثي على الثنائي المعدل	4.14	2.48	0.83	1.56	*	*
	الكفاءة النسبية لـ الثلاثي على عتبة التخمين	7.54	3.52	36.00	34.52	2.09	6.93

5.14	2.03	51.66	39.00	3.32	7.08	الكفاءة النسبية لـ الثلاثي على القائم على الصعوبة
-	-	22.09	43.56	1.42	1.82	الكفاءة النسبية لـ الثاني المعدل على عتبة التخمين
-	-	33.06	47.61	1.34	1.71	الكفاءة النسبية لـ الثاني المعدل على القائم على الصعوبة
0.74	0.97	1.50	1.09	0.94	0.94	الكفاءة النسبية لـ عتبة التخمين على القائم على الصعوبة

يُلاحظ من نتائج الجدول (19) أن قيم الكفاءة النسبية للنموذج الثلاثي المعلمة لجميع معالم الفقرات أفضل منها عند المقارنة مع النماذج الأخرى، يليه النموذج الثاني المعدل، ثم النموذج القائم على الصعوبة.

مناقشة النتائج والتوصيات

السؤال الأول:

هل يختلف تقدير معالم الفقرات باختلاف النموذج المستخدم في إطار النموذج الثلاثي المعلمة تبعاً لمتغيري طول الاختبار (25 فقرة و 50 فقرة) وعدد المفحوصين (500 مفحوص و 1500 مفحوص).

أولاً معلمة التمييز:

أظهرت النتائج أن متوسطات الجذر التربيعي لمتوسط مربع الخطأ تقل بزيادة طول الاختبار من (25 فقرة إلى 50 فقرة)، واتفقت هذه النتائج مع نتائج دراسة فو (Fu, 2010) التي أشارت إلى أن النموذج الثلاثي المعلمة ونموذج تثبيت معلمة التخمين، يُعطيان تقديرات دقيقة لمعالم الفقرة حينما يكون هناك مواعمة جيدة بين كل من معلمة الصعوبة للفقرة ومعلمة القدرة للمفحوص، كما اتفقت نتائج الدراسة الحالية مع دراسة عابنه (2004) التي أشارت إلى أن دقة التقديرات لمعالم الفقرات ومعلمة القدرة تزداد بزيادة حجم عينة المفحوصين وبزيادة طول الاختبار، كما اتفقت نتائج هذه الدراسة نسبياً مع دراسة الحاج (2011) التي أشارت إلى أن دقة التقديرات لكل من معالم الفقرة ومعلمة القدرة للمفحوصين أفضل في حالة النموذج الثلاثي المعلمة منها في حالة النموذج الثاني المعدل، كما اتفقت نتائج الإجابة عن هذا السؤال مع دراسة ستيدي (Setiadi, 1997) التي أشارت إلى أن دقة التقدير لمعالم الفقرات كانت أفضل في حالة استخدام النموذج الثلاثي المعلمة مقارنة مع النماذج الأخرى التي استخدمها الباحث وهي النموذج الأحادي المعلمة والنموذج الأحادي المعلمة المعدل، كما اتفقت النتائج التي توصلت لها هذه الدراسة مع ما أشار إليه هامبلتون وكوك (Hambleton and Cook, 1983) من أن حجم عينة المفحوصين المستخدمة في تقدير معالم الفقرة ومعلمة القدرة يؤثر على استقرار ودقة معالم الفقرات ومعلمة القدرة للمفحوصين.

كما اتفقت نتائج هذه الدراسة نسبياً مع دراسة جاو (Gao, 2011) التي أشارت إلى أن دقة تقديرات معلمة الصعوبة في حالة النموذج الثلاثي المعلمة أفضل منها في حالة المقارنة مع النماذج الأخرى المشمولة في الدراسة، كما اتفقت نتائج هذه الدراسة نسبياً مع دراسة ستيزر (Stetzer, 2007) التي أظهرت أن دقة التقدير لمعالم الفقرات في حالة استخدام النموذج الثلاثي المعلمة أفضل منها عند استخدام نموذج توسط الجهد الذي استخدمته تلك الدراسة والذي تمت الإشارة إليه خلال استعراض الأدب النظري في الفصل الثاني من هذه الدراسة.

كما اتفقت نتائج الدراسة الحالية مع نتائج دراسة والر (Waller, 1989) التي أشارت إلى أن دقة التقديرات لمعالم الفقرة للنموذج الثلاثي المعلمة أفضل منها في حالة استخدام نموذج (إزالة التخمين العشوائي) والنموذج الثلاثي المعلمة ذي معلمة التخمين الثابتة.

واختلفت نتائج هذه الدراسة مع نتائج دراسة سوبادي (Subedi, 2009) التي أظهرت أن دقة التقديرات لكل من معالم الفقرة ومعلمة القدرة كانت أفضل عند استخدام نموذجين مختلفين مقارنة بكل من النموذج الثلاثي المعلمة والنموذج الثاني المعدل، كما اختلفت نتائج هذه الدراسة مع نتائج دراسة (Pelton, 2002) التي توصلت إلى أن النموذج الثاني المعدل يُقدم تقديرات لمعلمة الصعوبة للفقرة أكثر دقة من تلك التي يُقدمها نموذج راش والنموذج الثلاثي المعلمة، ويمكن أن يُعزى تفوق النموذج الثلاثي

المعلمة على النموذج الثنائي المعدل بسبب إضافة ثابت التخمين (0.20) للنموذج الثنائي المعدل؛ الأمر الذي يُعد كأنه خطأ ثابت تمت إضافته للنموذج وهذا يقود إلى أن تكون دقة التقديرات في النموذج الثنائي المعدل أقل من دقة التقديرات في النموذج الثلاثي المعلمة؛ وهذا بالتالي يُفسر أن أخطاء التقدير لمعلمة الفقرة أو للمفحوصين أعلى في حالة النموذج الثنائي المعدل منها في حالة النموذج الثلاثي المعلمة، مما يُفسر أيضاً أن تكون الكفاءة النسبية لمعالم الفقرة أو لمعلمة القدرة أيضاً لصالح النموذج الثلاثي المعلمة مقارنة مع النماذج الأخرى المستخدمة في إطار هذه الدراسة .

كما يمكن تفسير دقة التقديرات لمعالم الفقرات في حالة النموذج الثلاثي المعلمة مقارنة بنموذج عتبة التخمين والنموذج القائم على الصعوبة؛ بسبب وجود معالم إضافية في هذه النماذج وبالتالي حاجتها إلى زمن أطول لإجراء عملية التقدير مقارنة مع النماذج الأخرى، كما أشار إلى ذلك كل من فرايك وستروبل وليش وزيليس (Frick, Strobl, Leish, and Zeileis, 2012)؛ كما أن هذه النماذج تتأثر بشكل كبير بالمعلومات القبلية (Priors Information) التي يتم تزويد البرنامج الحاسوبي المستخدم بها لإتمام عملية التقدير التي تؤثر بشكل كبير على دقة التقديرات لمعالم الفقرة .

أظهرت النتائج أفضلية للنموذج الثلاثي المعلمة عند المقارنة مع النماذج الأخرى المستخدمة ، وأظهر النموذج الثنائي المعدل أفضلية على النموذج القائم على الصعوبة ولم تكن الفروق دالة مع نموذج عتبة التخمين، وأظهر نموذج عتبة التخمين أفضلية على النموذج القائم على الصعوبة، كما أظهرت النتائج انخفاضاً في المتوسطات الحسابية لقيم الجذر التربيعي لمتوسط مربع الخطأ ظاهرياً بزيادة عدد فقرات الاختبار من (25) فقرة إلى (50) وبزيادة حجم عينة المفحوصين من (500) مفحوص إلى (1500) مفحوص على التوالي، إلا أن هذه الفروق لم تكن ذات دلالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$).

ثانياً: معلمة الصعوبة:

أظهرت نتائج الدراسة وجود فروق دالة إحصائية عند مستوى الدلالة ($\alpha = 0.05$) بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ تُعزى للنموذج المستخدم وذلك لصالح النموذج الثلاثي المعلمة، ثم للنموذج الثنائي المعدل على كل من نموذج عتبة التخمين والنموذج القائم على الصعوبة، ولم تكن الفروق ذات دلالة عند مقارنة نموذج عتبة التخمين مع النموذج القائم على الصعوبة، كما أظهرت المتوسطات الحسابية لقيم الجذر التربيعي لمتوسط مربع الخطأ انخفاضاً ظاهرياً في قيمها لجميع النماذج المستخدمة بزيادة حجم عينة المفحوصين من (500) مفحوص إلى (1500) مفحوص وبزيادة عدد فقرات الاختبار من (25) فقرة إلى (50) فقرة.

ثالثاً: معلمة التخمين:

أظهرت النتائج التي توصلت لها هذه الدراسة فيما يتعلق بدقة تقدير معلمة التخمين وجود فروق دالة إحصائية بين المتوسطات الحسابية للجذر التربيعي لمتوسط مربع الخطأ تُعزى للنموذج المستخدم وذلك لصالح النموذج الثلاثي المعلمة، ثم للنموذج القائم على الصعوبة ثم لنموذج عتبة التخمين كما أظهرت المتوسطات الحسابية لقيم الجذر التربيعي لمتوسط مربع الخطأ انخفاضاً ظاهرياً في قيمها بزيادة حجم عينة المفحوصين من (500) مفحوص إلى (1500) مفحوص وبزيادة عدد فقرات الاختبار من (25) فقرة إلى (50) فقرة. اتفقت النتائج التي توصلت لها الدراسة في الإجابة عن هذا السؤال مع نتائج دراسة العباينة (2004) ومع نتائج دراسة الحاج (2011) التي أشارت إلى أن دقة التقدير لمعالم الفقرة كانت أفضل في حالة النموذج الثلاثي المعلمة مقارنة مع النموذج الثنائي المعدل، كما اتفقت نتائج مع ما أشار إليه كل من سوامنثان جيفورد (Swaminathan and Gifford, 1983) و هامبلتون وكوك (Hambleton and, Cook, 1983) و هامبلتون وجونز (Hambelton and Jones, 1994) التي أشارت جميعها إلى زيادة الدقة في تقدير معالم الفقرات بزيادة حجم عينة المفحوصين وزيادة عدد فقرات الاختبار.

السؤال الثاني:

هل تختلف قيم الكفاءة النسبية لمعالم الفقرات ومعلمة القدرة باختلاف متغيرات الدراسة (النموذج المستخدم وعدد الفقرات وحجم العينة)؟

أولاً: معلمة التمييز:

أظهرت نتائج مؤشر الكفاءة النسبية في تقدير معلمة التمييز أفضلية للنموذج الثلاثي المعلمة عند المقارنة مع النماذج الأخرى المستخدمة يليه النموذج الثنائي المعدل، ثم نموذج عتبة التخمين، وأخيراً النموذج القائم على الصعوبة. كما أظهرت النتائج أن قيم الكفاءة النسبية في تقدير معلمة الصعوبة هي لصالح النموذج الثلاثي المعلمة يليه النموذج الثنائي المعدل ثم النموذج القائم على الصعوبة، وأخيراً نموذج عتبة التخمين.

أما بالنسبة لمعلمة التخمين فقد أظهرت النتائج المتعلقة بالإجابة عن هذا السؤال أن قيم الكفاءة النسبية قد كانت لصالح النموذج الثلاثي المعلمة عند مقارنته مع كل من نموذج عتبة التخمين والنموذج القائم على الصعوبة، وجاءت قيم الكفاءة النسبية لصالح نموذج عتبة التخمين في حالة الاختبار المكون من (25) فقرة والعينة المكونة من (500) مفحوص، وفي حالة الاختبار المكون من (50) فقرة كانت قيم الكفاءة النسبية لصالح النموذج القائم على الصعوبة.

ويمكن تفسير نتائج الإجابة عن هذا السؤال من خلال العلاقة التي تربط بين المؤشرات: الجذر التربيعي لمتوسط مربع الخطأ والتحيز والخطأ المعياري للتقدير $RMSE^2 = Bias^2 + SEE^2$ وبالتالي فإن القيمة الكبيرة للجذر التربيعي لمتوسط مربع الخطأ تتوافق مع قيمة كبيرة للخطأ المعياري للتقدير والعكس صحيح وبالتالي فإن النماذج التي كانت متوسطات الجذر التربيعي لمتوسط مربع الخطأ لها كبيرة فإن الخطأ المعياري للتقدير لها سيكون مرتفعاً أيضاً (Wang and Wang, 2002; Weiss and Babcock, 2009).

المراجع

- الحاج، أ. (2011). أثر التحكم بمعلمة التخمين في نماذج استجابة الفقرة على تقدير معالم الفقرات والأفراد. أطروحة دكتوراه غير منشورة، جامعة اليرموك، اردب، الأردن.
- الشريفين، ن. (2012). أثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة والخصائص السيكمترية للاختبار، في ضوء تغير حجم العينة. المجلة التربوية، 12 (104)، 22-67.
- الشواورة، ش. (2013). دقة تقدير معالم الفقرات بطريقتي الأرجحية العظمى الهامشية وبييز في ظروف مختلفة في عدد الفقرات وحجم العينة والنموذج اللوغاريتمي المستخدم. أطروحة دكتوراه غير منشورة. جامعة اليرموك، اردب. الاردن.
- العابنة، ع. (2004). أثر حجم العينة وطريقة إنتقائها على دقة تقدير معالم الفقرة والقدرة لاختبار قدرة عقلية باستخدام نظرية الاستجابة للفقرة. أطروحة دكتوراه غير منشورة، جامعة عمان للدراسات العليا، عمان، الاردن.
- علام، ص. (2005). نماذج الاستجابة للمفردة الاختبار-أحادية البعد ومتعددة الأبعاد- وتطبيقاتها في القياس النفسي والتربوي. القاهرة: دار الفكر العربي.
- Baldwin, P. (2006). A modified IRT Model Intended to Improve Parameter Estimates under Small Sample Conditions. Paper Presented at the annual meeting of the National Council on Measurement in education, San Francisco, CA: April 8-10.
- Barnes, L. and Wise, S. (1991). The Utility of a Modified One-Parameter IRT Model With Small Samples. Applied Measurement In Education, 4(2), 143-157.
- Cao, J. and Stokes, S. (2008). Bayesian IRT Guessing Models for Partial Guessing Behavior. Psychometrika, 73, 209-230.
- Dawber, T., Rogers, W., and Carbonaro, M. (2009). Robustness of Lord's Formulas for Item Difficulty and Discrimination Conversions Between Classical and Item Theory Models. The Alberta Journal of Educational research, 4(55), 512-533.
- De Ayala, R. (2009). The Theory And Practice Of Item Response Theory. Guilford Press. New York, USA.
- Embretson, S., and Reise, S. (2000). Item Response Theory for Psychologists. New Jersey: Lawrence, Erlbaum Associates, Publishers.
- FU, Q. (2010). Comparing Accuracy of parameter Estimation Using IRT Models in The Presence of Guessing. Unpublished Doctoral Dissertation University of Illinois at Chicago, USA.
- Gao, S. 2011. The Exploration of the Relationship Between Guessing and Latent Ability in IRT Models. Unpublished Doctoral Dissertation University of southern Illinois, USA.
- Gruijter, D., and Kamp L. (2002). Statistical Test Theory For Education and Psychology. Retrieved May 10 2014 from: <http://irt.com.ne.kr>.
- Hambelton, R. (1993). Item Response Theory: A Broad Psychometric Framework For Measurement Advances. Psychometrika, Vol 6, n 3, pp(353-556).
- Hambleton, R., and Jones, R. (1994). Comparison of Classical Test Theory and Item Response Theory and their Applications to test Development. Educational Measurement: Issue and Practice.

- Hambleton, Ro., and Jones, R.(1993). Item Paramter estimation errors and their influence on test information function. *Applied Measurment in Education*.7(3), 171-186.
- Hambleton, R.(1989). Principles and selected Applications of Item Response Theory. In R. L. Linn(Ed), *Educational Measurement 3th*, 147-200. Macmillan New York, USA.
- Hambleton, R.,(1989). Principles and selected Applications of Item Response Theory. In R. L. Linn(Ed), *Educational Measurement 3th*, 147-200. Macmillan New York, USA.
- Hambleton R., and Cook. L. (1983). Robutness of item response models and effects of test length and sample size in the precision of ability estimate. NewYork. In D.J. Weiss(Ed),*New Horizons in Testing*, 31-49.
- Hambleton R., Swaminthan. H, and Rogers. H.(1991). *Fundamentals of item response Theory*. International Educational and professional. Sage Publication, Inc Newbury Park, USA.
- Linden, W., and Hambleton, R.(1997). *Handbook of Modern Item Response Theory*. Springer-Verlage New York Inc, USA.
- Reckase, M.(2009).*Multidimensional Item response Theory*. Springer Dordrecht Heidelberg London.
- Setiadi, H.(1997).*Small Sample IRT Parameter Estimates*. Unpublished Doctoral Dissertation, University of Massachusetts, USA.
- Setzer, J.(2007).*Studying the effect of Rapid-Guessing on low-Stakes test: An Application of the Effort-Moderated IRT Model*. Unpublished Doctoral Dissertation, James Madison University, USA.
- Sireci, S.(1992). The Utility of IRT in Small-Sample Testing Applications. Reterived April. 12.2013 from <http://eric.ed.gov>.
- Stocking, M.(1990). Specifying Optimum Examinees for Item Parameter Estimation in Item Response Theory. *Psychometrika*, 55(3), 461-475.
- Subedi, D.(2009). *Investigating Unobserved Heterogeneity Using Item Response Theory Mixture Models*. Unpublished Doctoral Dissertation, Michigan University, USA.
- Swaminathan, H., Hambleton, R., Sireci, S., Xing, D. and Rizavi, S. (2003) Small Sample Estimation in Dichotomous Item Response Models: Effect of priors Based on Judgemental Information on the Accuracy of Item Parameter Estimates. *Applied Psychological Measurement*, 27, 27-51.
- Waller, M.(1989).*Modeling Guessing Behavior: A comparison of two IRT Models*. *Applied Psychological Measurement*, 13,233-243.
- Wang, S., And Wang, T.(2002). Relative Precision of ability Estimation in Polytomous CAT: A Comparison under the Generalized Partial Credit Model and Graded Response Model. Paper Presented at the annual meeting of the American Educational Research Association, New Orleans.
- Weiss, David J. and Babcock, B.(2009). Termination Critreria in Comuterized Adaptive Tests: Variable-Length CATs Are Not Biased. Presneted at the Realities of CAT Paper Aession, Retrieved April 8.2015 from [http:// www.psych.umn.edu/psylabs/CATCentral](http://www.psych.umn.edu/psylabs/CATCentral).
- Weiss, D.,. and Babcock, B.(2009). Termination Critreria in Comuterized Adaptive Tests: Variable-Length CATs Are Not Biased. Presneted at the Realities of CAT Paper Aession, Retrieved April 8.2015 from [http:// www.psych.umn.edu/psylabs/CATCentral](http://www.psych.umn.edu/psylabs/CATCentral).
- Wise, S., and DeMars,C.(2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, 10,1-17.
- Wise, S. and DeMars, C.(2006).An Application of Item Response Time: The Effort Moderated IRT model. *Journal of Educational Measurement*, 43,19-38.
- Xie, Y.(2002).*An Application of a Special Two-Class Item Response Model Using Markov chain Monte Carlo Method*. Unpublished Doctoral Dissertation, University of California at Berkely, USA.
- Yang, X.(2007).*Methods of Identifying Individual Guessers from Item Response Data*. *Educational and Psychological Measurement*. 67(5), 20-34.

Yen, W., and Fitzpatrick, A.(2006).Education Measurement .Inc. Brennan, Robert L, 4th Educational Measurement. .Praeger Publishers. USA.

The Accuracy of Item Parameters Estimation When Using Four Models within Item Response Theory

*Mueen Slman Al Nsraween **

ABSTRACT

This study aimed at investigating the accuracy of item parameters when using four models within item response theory, as test length vary (25 item and 50 item), and examinee sample size vary (500 and 1500 examinees). The four models were: the three parameter model, the two parameter modified model with a rate of guessing fixed at (0.20) for all individuals, the threshold guessing model, and the difficulty-based guessing model. To achieve the purposes of the study a simulated examinee-response dataset was generated by fitting a three parameter logistic model using (Wingen3) program, according to the following specifics: The difficulty parameters distributed normally with mean of (0) logits and standard deviation of (1) logits. The discrimination parameters were distributed uniformly with their initial value set equal to (0.40) and a maximum-possible value of (1.7) The guessing parameters were generated uniformly with initial value of (0) and a maximum value of (0.45). The results demonstrated statistically-significant differences at ($\alpha=0.05$) between means for (RMSE), and (Bias), when estimating item parameters (discrimination, difficulty, and guessing) when using the three logistic model, followed by the two parameter modified model, The relative-efficiency indicator also demonstrated the same superiority for the three logistic parameter model and the two-parameter modified model, with the results showing that the, RMSE, and Bias, decreased as the sample size increased from (500) to (1500) examinees and as test length increased from (25) to (50) items.

Keywords: Accurate Estimation, Modified Models, Mixture Models.

.

* Ministry of Education. Received on 30/6/2016 and Accepted for Publication on 15/8/2016.