أثر نموذج نظرية الاستجابة للفقرة ذات الاستجابة المتعددة التدريج على دقة تقدير القدرات للأفراد والخصائص السيكومترية للفقرات والاختبار

د. نضال كمال الشريفين

كلية التربية - جامعة اليرموك

الأردن

الملخص

هدفت الدراسة إلى الكشف عن أثر نموذج نظرية الاستجابة للفقرة ذات الاستجابة المتعددة التدريج المستخدم في التحليل وهي: نموذج الاستجابة الإسمي لبوك، ونموذج التقدير الجزئي العام لموراكي، ونموذج الاستجابة المتدرج لساميجما على دقة تقدير القدرات للأفراد، والخصائص السيكومترية للفقرات، والاختبار. ولتحقيق هدف الدراسة تم استخدام برنامج (WinGen 3) في توليد بيانات خاصة باختبار مكون من ٥٠ فقرة وفق نموذج التقدير الجزئي العام، وكل فقرة مكونة من خمسة مستويات، بأربعة عتبات فارقة/ خطوات. أشارت نتائج الدراسة إلى أن نموذج التقدير الجزئي العام هو أكثر دقة في تقدير القدرات من نموذجي الاستجابة: الإسمي والمتدرج. وأشارت النتائج إلى أن نموذج الاستجابة الإسمي هو نموذج أكثر دقة في تقدير معالم التمييز للفقرات مقارنة بالنموذجين الآخرين، كما بينت النتائج أن نموذج الاستجابة المتدرج هو أكثر دقة في تقدير معالم الصعوبة للعتبات الفارقة من النموذجين الآخرين. وأنه قدم أقصى كمية من المعلومات التي يقدمها الاختبار، وأكثر دقة في تحليل البيانات مقارنة بالنموذجين الآخرين، كما كشفت النتائج عن عدم وجود فروق عند مستوى الدلالة (α=٠,٠٥) بين معاملات ثبات كرونباخ ألفا تبعاً لنموذج الاستجابة المستخدم.

مقدمة

حظيت الفقرات (ذات الاستجابة الثنائية) (Dichotomous) باهتمام كبير في البحوث التربوية، وكان نصيب الفقرات ذات الخطوات المتعددة (متعددة التدريج) (Polytomous) محدوداً، علماً بأن للفقرات المتعددة التدريج أهمية كبيرة في التقويم وبناء الاختبارات الهادفة. فهناك بعض الأهداف يصعب

قياسها باستعمال الفقرات الموضوعية الثنائية التدريج مثل الأهداف التي ترتبط بإنتاج برهان رياضي أو حل مسألة أو التعبير عن الأفكار كتابياً وشفوياً، إذ لا يكون مقبولاً أن يصحح أداء الطالب في موضوع إنشائي أو حل مسألة يتكون حلها من عدة خطوات بطريقة التصحيح الثنائي، لذا يعمل المعلمون ولجان التصحيح في الامتحانات العامة على وضع ما يسمى مفتاح الإجابة الصحيحة للأسئلة متعددة التدريج، حيث يتم تجزئة الاستجابة إلى عدة خطوات، ويمنحون درجات جزئية على هذه الخطوات، وقد يختلف عدد الخطوات من مصحح لآخر. لذا فقد تختلف التقديرات للمفحوص عند التصحيح من مصحح لآخر، مما ينفي ميزة الموضوعية عن هذا النوع من الاختبارات، وانخفاض درجات الثبات والصدق مما يترتب على ذلك قرارات تفتقر إلى المصداقية والموضوعية. لذلك كان لا بد من البحث عن أكثر النماذج النظرية دقة التي تستخدم لتقدير معالمها، وتقديرات القدرة للأفراد بدقة.

وعند الاطلاع على أنواع الفقرات المستخدمة في الاختبارات العامة أو المدرسية يمكن ملاحظة أن معظم الفقرات المستخدمة هي من نوع الفقرات متعددة الخطوات (الأسئلة المقالية)، ولأسئلة المقال فوائد كثيرة منها: سهولة بنائها حيث تحتاج إلى وقت أقل من الوقت الذي يحتاجه بناء أسئلة ذات إجابة منتقاة (الاختيار من متعدد)، وتتطلب من المتعلم أن يعبر عن نفسه بشكل كتابي، وهي تمثل مهارة لفظية لا بد من تنميتها عند المتعلم، وهي داعمة للاختبارات الموضوعية حيث تتطلب من المفحوص أن يربط مكونات الوحدة التعليمية بعضها بعضا، وتعطي المعلم فكرة عن قدرة المتعلم على حل المسألة (problem solving) (عوده، ٢٠١٥)، وهناك دعوات كبيرة لتكوين اختبارات تتضمن استعمال الفقرات متعددة الخطوات بجانب الفقرات الموضوعية، ومع الدعوات المتكررة في بناء بنوك الفقرات فقد أخذت عملية تقدير خصائص الفقرات متعددة التدريج واستخراج معالمها، والإحصاءات الخاصة بها تظهر على السطح وتتحدى المربين في إيجاد النماذج التي تتعامل مع هذا النوع من الفقرات، وترتبط الفقرات متعددة التدريج بنماذج الاستجابة للفقرة التي تتعامل مع الاستجابات المتعددة (Polytomous)، حيث هدفت نماذج نظرية الاستجابة للفقرة إلى تحديد العلاقة بين أداء الفرد على

الفقرة، وبين القدرة أو السمة الكامنة وراء هذا الأداء، ولتحقيق ذلك طورت عبر السنوات الماضية العديد من النماذج سميت أحياناً باسم موجدها مثل: نموذج راش (Rasch)، وأحياناً باسم الدالة الرياضية للنموذج نفسه مثل: نموذج المنحنى الطبيعي، وأحياناً باسم الوظيفة المقترحة للنموذج مثل نموذج التقدير الجزئي (Partial Credit) ونموذج مقياس/ سلم التقدير (Rating Scale Model, RSM)، وأحياناً يضاف تعديل ما على الاسم مثل مقياس التقدير لموراكي (Muraki) (Thissen & Steinberg, 1986).

ونظراً لتزايد عدد النماذج المستخدمة، فقد تم تصنيفها تبعاً لتدريج فقرات المقياس في ثلاث فئات هي: النماذج الخاصة بالاستجابات ثنائية التدريج (Dichotomous) مثل: فقرات الاختيار من متعدد، أو الصواب والخطأ، وتعطى فيها الإجابة الصحيحة الدرجة (١)، والإجابة الخاطئة الدرجة (٠)، والنماذج الخاصة بالاستجابات متعددة التدريج (Polytomous) مثل: تقويم مهارات، حل المسائل، حل المشكلات، وتنفيذ إجراءات، وكتابة مقال، وتجميع جهاز، وتتطلب فقراتها استجابات متعددة التدريج، وكذلك فقرات مقاييس الاتجاهات والشخصية التي تشتمل على ميزان (Scale) متعدد التدريج بدءاً من موافق بشدة وانتهاءً بغير موافق إطلاقاً، وسيتم عرض السؤال التالي الذي يعطى فيه المفحوص تقديرات جزئية على النجاح الجزئي على السؤال. فمثلاً لو كان السؤال: جد المسافة بين النقطتين أ (٢، -٢)، ب (٥، ٢).

الخطـوات	مستويات الحل					
	1	2	3	4	5	6
$\sqrt{(X2-X1)^2+(Y2-Y1)^2}$	الخطوة الأولى					
$\sqrt{(5-2)^2+(2--2)^2}$		الخطوة الثانية				
$\sqrt{(3)^2+(4)^2}$			الخطوة الثالثة			
$\sqrt{(3)^2+(4)^2}$				الخطوة الرابعة		
$\sqrt{25}$					الخطوة الخامسة	
5						

الشكل رقم (١): (العلاقة بين مستويات وخطوات الحل في سؤال يتألف من ست مستويات)

ويلاحظ من السؤال أن الرتب الجزئية هي (١، ٢، ٣، ٤، ٥، ٦)، أي أن هذا السؤال يتألف من ستة مستويات أو ستة مهمات جزئية. يصل المفحوص إلى أي مستوى منها بعد أن يقوم بالخطوة التي تؤدي إلى هذا المستوى، فلا يصل إلى المستوى الثاني إلا بعد أن يقوم بالخطوة الأولى، ولا يصل إلى المستوى الثالث إلا بعد أن يقوم بالخطوة الثانية، وهكذا يصل المفحوص إلى المستوى السادس بعد أن يقوم بالخطوة الخامسة. وهناك أسلوبان للتعامل مع خطوات السؤال، فالأسلوب الأول يعدّ امتداداً لأسلوب الفترات المتتابعة (Successive Intervals) لثيرستون في تقدير معالم الصعوبة للعتبات/ الخطوات (Thresholds) بين مختلف المستويات، حيث تكون درجة صعوبة أي فاصل أعلى من صعوبة الفاصل الذي يسبقه؛ حيث يستعمل الرمز (δij) للدلالة على صعوبة الفاصل بين المستوى (j) والمستوى (j-1) للسؤال (i). والأسلوب الآخر يتعامل على أساس أن كل خطوة من خطوات السؤال عبارة عن سؤال ثنائي التدريج تكون الإجابة عنها إما (١) عندما يصل إليها المفحوص أو (صفراً) عندما لا يصل إليه المفحوص، ويتعامل مع صعوبة كل خطوة من خطوات السؤال الواحد بشكل منفصل مقارنة بالخطوة السابقة لها؛ حيث يستعمل الرمز (δij) للدلالة على صعوبة الخطوة (j) في السؤال (i)، ويتعامل مع هذا الأسلوب كل من نموذج الاستجابة المتدرجة لساميجما (Samejima) (Samejima Graded Response Model, GRM)، ونموذج التقدير الجزئي العام (General Partial Credit Model, GPCM) الأسلوب. وهما من النماذج المستخدمة في هذا الدراسة.

والنماذج الأخرى تتعلق بالاستجابات المتصلة (Continuous)، وهو امتداد للاستجابات المتعددة التدريج، بافتراض أن عدد أقسام الاستجابة لا نهائي، حيث يقوم الفرد بوضع إشارة على نقطة واقعة على متصل التدريج (Sijtsma & Hemker, 2000). وقد وضع ثيزن وستينبيرغ (Thissen & Steinberg) المشار إليهما في لوستينا (Lustina, 2004) ثلاثة تصنيفات خاصة بالنماذج المتعددة التدريج، وذلك بناءً على الافتراضات والقيود على معالمها، وهي، أولاً: نماذج الفرق (Difference Models) وهي تلائم الاستجابات في المستوى الرتبي، ومنها:

نموذج الاستجابة المتدرجة لساميجما (GRM)، ونموذج مقياس التقدير لموراكي (Muraki Rating Scale Model, MRSM)، وثانياً: نماذج القسمة على المجموع (Divide-by-Total Model)، وهي تلائم البيانات الواقعة بمستويي القياس الإسمي، والرتبي، ومنها: نموذج الاستجابة الإسمي لبوك (Nominal Response Model, NRM)، ونموذج التقدير الجزئي العام لموراكي (Generalized Partial Credit Model, GPCM)، ونموذج التقدير الجزئي لماسترز (Partial Credit Model, PCM)، ونموذج الفقرات المتتابعة لروست (Successive Interval Model, SIM)، ونموذج مقياس التقدير لأندريش (Andrich Rating Scale Model, ARSM)، وثالثاً: نماذج القسمة على المجموع مع إضافة الجانب الأيسر (Left-Side Added Divide By Total)، وهي تلائم بيانات اختبارات الاختيار من متعدد مع التخمين، ومنها: نموذج الاختيار من متعدد لساميجما كأحد مكونات النموذج الإسمي لبوك، ونموذج الاختيار من متعدد لثيزن وستينبيرغ، ونموذج سمبسون حيث تم اشتقاقها لتتضمن التخمين في فقرات الاختيار من متعدد.

وبشكل عام، فإن اختيار النموذج المناسب لاستخدامه في موقف تطبيقي معين، ومطابقته لمجموع بيانات واقعية لا يعد أمراً يسيراً، فقد أشار علام (٢٠٠٥) إلى مجموعة من المحكات التي يمكن الاستناد إليها في هذا الشأن، منها؛ أولاً: واقعية افتراضات النموذج. ثانياً: مَنَعة النموذج (Robust) لمخالفة افتراضاته؛ أي أن انتهاك بعض افتراضات النموذج ربما لا تؤثر في دقة التقديرات الناتجة عن استخدام هذا النموذج في مواقف تطبيقية معينة. ثالثاً: أنواع البيانات الاختبارية المراد تحليلها، وعموماً فإن اختيار النموذج المناسب لتحليل البيانات يعتمد في معظم الأحيان على اعتبارات عملية، مثل: خبرة الباحث باستخدام نموذج معين، وتوافر برامج حاسوبية لتنفيذ ما يتطلبه النموذج من عمليات إحصائية وحسابية. وهناك بعض الدراسات الأجنبية التي تناولت موضوع المقارنة بين نماذج نظرية الاستجابة للفقرة متعددة التدريج، مثل دراسة دي أيالا، ودود وكوك (De Ayala, Dodd & Koch, 1992) التي بينت أن نموذج الاستجابة المتدرجة أكثر دقة من نموذج التقدير الجزئي في تقدير قدرة

الأفراد ومعالم الفقرات، بينما أكدت دراسة آدمز (Adams, 1988) أن نموذج التقدير الجزئي أكثر دقة من نموذج الاستجابة المتدرجة في عمليات تقدير القدرة للمفحوصين في ظل متغير عدد فئات الاستجابة، ومتغير مستوى التحيز للمفردات، وكذلك طول الاختبار. أما دراسة سانغ وكانغ (Sung & Kang, 2006) فقد أشارت إلى أن الفرق بين النماذج متعددة التدريج في تقدير معالم القدرة للأفراد والفقرات يعتمد على الأوضاع الخاصة للبيانات التي يتم توليدها، لذا فإن الانتشار الواسع لاستخدام النماذج ذات الاستجابات المتعددة التدريج في البحث ولّد نقاشاً ملحوظاً حول أي النماذج يُعد هو النموذج المناسب في دقة التقديرات للاستجابات متعددة التدريج، لذا جاءت هذه الدراسة لتقدم للباحثين نموذج الاستجابة متعدد التدريج الذي يتصف بدقة تقديراته لمعالم الفقرات، ولتقديرات القدرة للمفحوصين.

دراسات سابقة

أجرى دود (Dodd, 1984) دراسة هدفت إلى تقييم الدقة النسبية للخصائص ذات الصلة بين نموذج الاستجابة المتدرجة (GRM)، ونموذج التقدير الجزئي (PCM)؛ باستخدام فقرات مقياس اتجاهات متعدد التدريج، وتم استخدام مقياس اتجاهات من نوع ليكرت؛ بينت النتائج أن نماذج التقدير الجزئية، والاستجابة المتدرجة أعطت تقديرات متكافئة لمستويات الاتجاهات المقيسة، ومعالم صعوبة الفقرة لكل من مجموعتي البيانات (الحقيقية، والمولدة)، بالرغم من الاختلاف التام بينهم في تعريف معالم صعوبة الفقرة.

كما أجرى دي أيالا ودود وكوك (De Ayala, Dodd & Koch, 1992) دراسة قارنت بين نموذجي التقدير الجزئي والاستجابة المتدرجة في مجال الاختبارات التكيفية (Computerized Adaptive Testing, CAT)، حيث إن معظم الاختبارات التكيفية تعتمد على نماذج نظرية الاستجابة للفقرة ثنائية التدريج، وبعض الباحثين استخدم نماذج نظرية الاستجابة للفقرة متعددة التدريج؛ مثل نموذج الاستجابة المتدرجة GRM، ونموذج التقدير الجزئي PCM في الاختبارات

التكيفية، أشارت نتائج الدراسة إلى دقة معقولة في تقدير التقدير الجزئي للقدرة حيث بلغت قيمة معامل الارتباط بين تقديرات القدرة في النموذجين أكبر أو تساوي (٠,٩٢١)، على الرغم من أن الاختبارات التكيفية تضمنت ٠,٤٥٪ من الفقرات غير المطابقة، وأن دقة تقديرات القدرة باستخدام النموذجين للاختبارين المتماثلين في الطول أظهرت دقة أكثر لنموذج الاستجابة المتدرجة من نموذج التقدير الجزئي.

وفي دراسة قام بها كوك ودود وفيتسباترك (Cook, Dodd & Fitzpatrick, 1999) هدفت للمقارنة بين ثلاثة نماذج لنظرية الاستجابة للفقرة متعددة التدريج في سياق تصحيح نتائج اختبار قصير Testlet، وأجريت الدراسة على مجموعتين من البيانات الأولى أرشيفية، والأخرى تقليدية (محاكاة)، تبين أن نموذج PCM أفضل من أداء النموذجين الآخرين GRM، PCM لمجموعة بيانات اختبار الاستعداد المدرسي Scholastic Aptitude Test I (SAT I) المطبق خريف عام ١٩٩٤، أما مجموعة بيانات المحاكاة كانت مطابقة النموذجين GPCM، GRM أفضل بقليل من نموذج PCM.

وفي دراسة أجراها أوستني (Ostini, 2001) هدفت إلى تحديد فروق القياس الموضوعية بين نماذج مختلفة من نماذج نظرية الاستجابة للفقرة متعددة التدريج، تم الحصول على نتائج القياس باستخدام ثمانية نماذج استجابة لفقرات متعددة التدريج، تم تقديرها باستخدام سبعة برامج مختلفة، أشارت نتائج الدراسة إلى أن توزيعات السمة الناتجة متشابهة في مجموعتي البيانات بشكل لافت للنظر، وكانت الارتباطات في مجموعتي البيانات صغيرة، وكانت قيم مؤشري الجذر التربيعي لمتوسط مربع الأخطاء (Root Mean Squared Error, RMSD)، ومعدل الفرق المحدد (Average Signed Difference, ASD). لمعالم الفقرات أكبر مقارنة بمعالم الأفراد، والقيم المطلقة لـ RMSD و ASD أكبر لدوال النموذج؛ (في: Walford, Tucker & Viswanathan (ed), 2010).

وأجرت سوزان (Susan, 2001) دراسة لتقصي فاعلية ثلاثة نماذج في

نظرية الاستجابة للفقرة في تقدير قدرات الأفراد ودرجة الصعوبة للفقرات باستخدام نموذج راش ذي المعلمة الواحدة، ونموذج التقدير الجزئي، ونموذج الاستجابة المتدرجة. فقد قامت الباحثة ببناء اختبارات محكية المرجع تتضمن مزيجا من الفقرات من نوع الاختيار من متعدد، وأسئلة متعددة التدريج مكونة من أربع خطوات، وتم تدريج الفقرات باستخدام النماذج اللوجستية: الأحادي المعلمة، والثنائي المعلمة، والثلاثي المعلمة، وكذلك تدريج الفقرات متعددة التدريج باستخدام نموذج التقدير الجزئي. أشارت نتائج الدراسة إلى فاعلية نموذج التقدير الجزئي في معايرة الفقرات من النماذج اللوجستية، وذلك عند مستويات القدرة العالية، في حين أن النماذج اللوجستية كانت أكثر فاعلية في معايرة الفقرات عند مستويات القدرة المتدنية.

وفي دراسة أخرى قام بها وانغ ووانغ (Wang & Wang, 2002) هدفت إلى المقارنة بين طريقة تقدير الأرجحية الموزونة Weighted Likelihood Estimate (WLE)، وطرق تقدير الأرجحية القصوى Maximum Likelihood Estimate (MLE)، التقدير البعدية المتوقعة Expected A Posteriori Estimate (EAP)، التقدير البعدية القصوى Maximum A Posteriori Estimate (MAP) ، وفق نموذج التقدير الجزئي المعمم Generalized Partial Credit Model (GPCM)، ونموذج الاستجابات المتدرجة Graded Response Model (GRM) بينت النتائج أن طريقة WLE لها أقل تحيز وأقل خطأ معياري على تقدير القدرة في النموذجين GPCM، GRM، وأن طريقتي MLE، WLE أظهرتا تحيزاً أقل من طريقتي بييز وللنموذجين. كما أظهرت الطريقتان MLE، WLE تحيزاً أقل من طريقتي بييز، وذلك عند مستويات القدرة المتطرفة، وأن لهما أقل خطأ معياري لمتوسط المربعات RMSE، وأن الخطأ المعياري والتحيز يقل في جميع الطرق المستخدمة بزيادة طول الاختبار وثباته، وبزيادة عدد المتغيرات المستقلة الاختلافات.

وفي دراسة أخرى أجراها الزعبي (٢٠٠٣) حول فاعلية نموذج التقدير

الجزئي في بناء فقرات بنك أسئلة متعددة الخطوات في مادة الكيمياء لطلبة الصف الثاني الثانوي العلمي، تم استقصاء فاعلية نموذج التقدير الجزئي في تقدير قدرات الأفراد، ومعايرة أسئلة متعددة الخطوات، وتوصلت الدراسة إلى أن الخطأ المعياري يزداد بزيادة عدد خطوات الاستجابة، وعدم وجود فروق ذات دلالة إحصائية في متوسطات الفروق المعايرة لتقديرات معالم الصعوبة للأسئلة، ووجود فروق دالة إحصائياً بين تقديرات القدرة المعايرة، وأن الكفاءة النسبية للاختبار بشكل عام تزداد بزيادة عدد خطوات الاستجابة.

وأجرت لوستينا (Lustina, 2004) دراسة هدفت إلى المقارنة بين نموذجين من نماذج نظرية الاستجابة للفقرة متعددة التدريج هما: نموذج أندريش (ARSM) ونموذج الفترات المتتابعة لروست (SIM) في قياس الاتجاهات، حيث تمت المقارنة بين النموذجين باستخدام اختبار مبني وفق طريقة القياس التكيفي المحوسبة، وباستخدام مجموعتين من البيانات إحداهما حقيقية، والثانية كانت مولدة، وتم استخدام معامل ارتباط بيرسون والخطأ المعياري في التقدير لأغراض المقارنة بين النموذجين، أظهرت النتائج أن كلا النموذجين لديه مزايا ضمن ظروف الاختبار التكيفي المعطى عن طريق الحاسوب.

كما قام هاريس ولان وموسينسون (Harris, Laan & Mossenson, 2009) بدراسة هدفت إلى تطبيق أسلوب التقدير الجزئي في بناء اختبارات الكتابة الروائية، حيث تم كتابة مجموعة من المهام الروائية التي تستخدم مع الأطفال من سن الثالثة وحتى العاشرة استخدم نموذج التقدير الجزئي في معايرتها، وتم تقييم القصص الروائية التي قام الطلبة بكتابتها باستخدام ثمانية مقاييس (مثل: الجلسة، استخدام التماسك، بناء القصة)، كما تم التعامل مع هذه المقاييس كثماني فقرات. بينت النتائج دقة التقدير باستخدام نموذج التقدير الجزئي، وأن مقياس الكتابة المعاير يتوجب استخدامه كمقياس عالمي لقياس كفاءة الكتابة الروائية؛ لأن بإمكانها تفسير القدرة على الكتابة الروائية بشكل دقيق وحسب الخصائص الأكثر احتمالا لكتابة كل طالب نظراً لدقة أسلوب التحليل المستخدم.

وفي دراسة قام بها هوجنز والجينا (Huggins & Algina, 2015) بغرض إثبات حجم التشابه بين نموذج التقدير الجزئي PCM، ونموذج التقدير الجزئي العام GPCM، ونموذج الاستجابة الإسمي NRM، من خلال استخدام برمجية Mplus المستخدمة في تحليل بيانات نموذج NRM، من أجل معايرة البيانات باستخدام نموذج التقدير الجزئي، ونموذج التقدير الجزئي العام، حيث إن هناك عدد من الباحثين غير متيقنين فيما إذا كان النموذجان PCM، GPCM يمكن تقدير بياناتهما باستخدام برنامج Mplus المخصصة لتحليل بيانات نموذج الاستجابة NRM، أفضت النتائج إلى نتائج تحليل متشابهة، نظراً لعدم وجود فروق ذات دلالة إحصائية في تقديراتهما، وبينت أن النموذجين PCM، GPCM هما نموذجان متضمنان في نموذج NRM، ويمكن تقدير بياناتهما باستخدام آخر إصدار للبرمجية، وتم عرض مثال مع بيانات لعينة مكونة من ٥٢٢ فرد على مجموعة جزئية من فقرات مقياس الكفاءة الذاتية في الرياضيات Math Self- Efficacy، وتبين بأن استخدام برمجية Mplus هي وسيلة ناجعة في تحليل وتقدير بيانات النموذجين.

وفي دراسة أجراها جيانغ ووانغ وويس (Jiang, Wang & Weiss, 2016) للكشف عن حجم العينة المناسب في تحليل البيانات الخاصة باستخدام نموذج الاستجابة المتدرج المتعدد (Multidimensional Graded Response Model, MGRM)، تم إجراء دراسة محاكاة عن طريق توليد بيانات بأحجام عينات مختلفة، وأطوال اختبار مختلفة، أشارت نتائج الدراسة أن الغالبية العظمى من الدراسات التي كان حجم العينة فيها ٥٠٠ أشارت إلى دقة في تقدير معالم الفقرات باستثناء الاختبارات المكونة من ٢٤٠ فقرة فقد كان حجم العينة ١٠٠٠ للحصول على تقديرات دقيقة لمعالم الفقرات، وأن زيادة حجم العينة إلى أكثر من ١٠٠٠ لا يزيد من دقة تقديرات المعالم الخاصة بالنموذج MGRM.

في ضوء ما تقدم عرضه من الدراسات السابقة يمكن ملاحظة: أن معظم الدراسات الأجنبية قد تمت من خلال بيانات محاكاة باستخدام البرامج الإحصائية، وليست باستخدام بيانات فعلية، كما أن تلك الدراسات لم تحسم

الجدل حول النموذج المتعدد التدريج المفضل في تقدير معالم الفقرات وتقديرات القدرة للأفراد، وطريقة تقدير الثبات، ولا يوجد اتفاق بين تلك الدراسات حول هذه النقاط تحديداً، ولم تتطرق معظمها إلى أثر نموذج الاستجابة متعدد التدريج كمتغير يؤثر في تلك التقديرات، كما أنه - في حدود علم الباحث - لا توجد دراسات عربية حول الموضوع نفسه، لذا ونظراً لقلة الدراسات التي عالجت مثل هذا النوع من المشكلات البحثية، فإنه يؤمل من هذه الدراسة أن تقدم إضافة معرفية جديدة حول النموذج الأكثر دقة في تقدير المعالم المختلفة، والكشف عن أثر نموذج الاستجابة المتعدد التدريج في خصائص الاختبار وخصائص فقراته السيكومترية.

أما عن سبب اختيار الباحث للنماذج الثلاثة لتحقيق غرض هذه الدراسة، يتمثل في أنها تفترض أن معلمة التخمين ثابتة، وهي صفة خاصة بالأفراد ذوي القدرة المتدنية في حين أن معلمتي التمييز والصعوبة هما متغيران في هذه النماذج، وهما المعلمتان اللتان بالإمكان المقارنة بينهما كمعالم مشتركة بين النماذج الثلاثة، في حين أن النماذج الأخرى مختلفة فعلى سبيل المثال يفترض نموذج PCM أن معلمة الصعوبة هي المعلم الوحيد المتغير، ونموذج RSM يفترض أن معلمتي: الصعوبة والتخمين هما المتغيران في حين أن التمييز هي نفسها لكل الفقرات في النموذج، وكذلك توافر برنامج WinGen 3 الخاص بتوليد بيانات بالمواصفات المطلوبة؛ حيث يمكن بواسطته توليد بيانات تتقاطع فيما بينها في النماذج الثلاثة، كما أنها من أكثر النماذج المستخدمة في الدراسات (Dodd, Ayala & Koch, 1995)، وتعد من أكثرها واقعية (Muraki, 1990)، كما أن الدراسات المشار إليها آنفاً لم تحسم الجدل حول الطريقة الأكثر دقة في تقدير القدرة للفرد، وتقدير الخصائص السيكومترية للفقرات.

مشكلة الدراسة

هناك نماذج متعددة تستخدم في تقدير معالم الفقرات، حيث تعد عملية اختيار نموذج نظرية الاستجابة للفقرة المتعددة التدريج الملائم في تقدير معالم

الفقرات والأفراد من الخطوات الأساسية والهامة لتطبيق هذه النظرية، إذ تباينت وجهات النظر حول استخدام الطريقة الفضلى ما بين مؤيد ومعارض. فمنهم من يرى أن التقديرات التي يتم الحصول عليها متشابهة باختلاف النموذج المستخدم، ومن هنا تنشأ الحاجة إلى مثل هذه الدراسة التي تهدف إلى المقارنة بين نماذج نظرية الاستجابة المتعددة التدريج على دقة تقدير معالم الفقرة وتقديرات القدرة للأفراد والخصائص السيكومترية للاختبار، وباستخدام نماذج الاستجابة المتعددة التدريج (GRM, GPCM, NRM)، وهي النماذج الأكثر استخداماً في الدراسات الأجنبية (Dodd, Ayala & Koch, 1995)، حيث إن هذه النماذج الثلاثة تتشارك في افتراض معلمة التمييز ومعلمة الصعوبة، وهي تناسب البيانات الواقعة بمستوى القياس الرتبي، إذ أن مصداقية النتائج قد تكون موضع تساؤل عندما لا يتحقق فرض وقوع البيانات على مستوى فئوي، الأمر الذي يبرره استخدام هذه النماذج التي تستطيع تحليل بيانات على مستوى القياس الرتبي، بعكس النماذج التي تشترط وقوع البيانات على مستوى القياس الفئوي على الأقل؛ بسبب خاصية التصنيف الهرمي في مستويات القياس، ولكن قد يكون هناك اختلاف بين النماذج في دقة التقديرات (Sijtsma & Hemker, 2000)، والدراسات المشار إليها آنفاً لم تحسم الجدل حول الطريقة الأكثر دقة في تقدير التقديرات المختلفة للفقرات والاختبار، كما أن هذه النماذج لم تحظ باهتمام كبير في الدراسات العربية في هذا المجال، لذا وفي ضوء ما تقدم ارتأى الباحث استخدام هذه النماذج لتحقيق أهداف الدراسة، وبالتحديد فإن هذه الدراسة سعت للإجابة عن التساؤلات الآتية:

١ - ما أثر نموذج الاستجابة المتعدد التدريج نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM، على دقة تقدير معالم القدرة للأفراد؟

٢ - ما أثر نموذج الاستجابة المتعدد التدريج نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM، على دقة تقدير معالم عتبات الصعوبة للفقرات؟

٣ - ما أثر نموذج الاستجابة المتعدد التدريج نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM، على دقة تقدير معالم التمييز للفقرات؟

٤ - ما أثر نموذج الاستجابة المتعدد التدريج نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM، على ثبات درجات الاختبار؟

٥ - ما أثر نموذج الاستجابة المتعدد التدريج نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM، على دالة المعلومات للاختبار؟

أهمية الدراسة

تتمثل أهمية هذه الدراسة في جانبين: الأول نظري يتوقع من خلال هذه الدراسة أن تكشف عن النموذج المناسب في تقدير معالم الفقرات، وخصائص الاختبار، وقدرات الأفراد من خلال بيانات محاكاة. والجانب الثاني تطبيقي يتوقع أن توفر هذه الدراسة أدلة إمبريقية حول النموذج الأكثر دقة في تقدير معالم الفقرات وخصائص الاختبار، وتقديرات القدرة للأفراد؛ بمعنى محاولة التوصل إلى تبريرات عملية تقدم إلى الباحثين يمكن اعتمادها بناءً على أساس تجريبي، من خلال بيانات محاكاة حول النموذج المناسب الذي يمكن استخدامه من قبل الباحثين في دراسات أخرى لتقدير معالم الفقرات، وقدرات الأفراد عند استخدام نظرية الاستجابة للفقرة للاستجابات متعددة التدريج، كما يمكن أن يستفيد مطورو الاختبارات الذين يلجؤون إلى نظرية الاستجابة للفقرة لحل المشكلات المتعلقة ببناء الاختبارات لاختيار النموذج الأنسب عند تقدير معالم الفقرات، وهي: الصعوبة، والتمييز، وتقديرات القدرة للأفراد، وبالتحديد عندما تكون الأسئلة متعددة التدريج، لا سيما وأنها لم تحظَ بالاهتمام الكافي في الدراسات العربية.

التعريفات الاصطلاحية

نماذج نظرية الاستجابة للفقرة المتعددة التدريج: هي النماذج التي يمكن تطبيقها على فقرات متعددة التدريج ذات شكل استجابة لوجستي محدد، وتفترض أن البيانات فيها واقعة بمستوى القياس الفئوي، ولا يوجد عدد محدد لهذه النماذج التي يمكن أن تتولد في إطار نظرية الاستجابة للفقرة، وفي هذه الدراسة سيتم استخدام ثلاثة نماذج لنظرية الاستجابة للفقرة المتعددة التدريج (GRM, GPCM, NRM) التي تتعامل مع البيانات كما لو كانت بمستوى القياس الإسمي أو الرتبي.

الخصائص السيكومترية للفقرات: تتمثل في تقدير معالم الصعوبة، ومعالم التمييز، التي سيتم تقديرها وفق نماذج نظرية الاستجابة للفقرة متعددة التدريج (GRM, GPCM, NRM) حيث إن هذه النماذج تشترك في تقدير قيم معالم الصعوبة للعتبات الفارقة، والتمييز فقط باستخدام برنامج WinGen 3 المستخدم في توليد البيانات في هذه الدراسة.

الخصائص السيكومترية للاختبار: تتمثل في تقدير كمية المعلومات التي يقدمها الاختبار من خلال دالة المعلومات للاختبار، وتقدير الثبات لدرجات الاختبار، وفي هذه الدراسة تم تقديرها وفق نماذج نظرية الاستجابة للفقرة متعددة التدريج (GRM, GPCM, NRM).

معالم الأفراد: هي معالم قدرات الأفراد، وتمثل مقدار ما يمتلكه الفرد من السمة التي يتم قياسها تبعاً لنماذج نظرية الاستجابة للفقرة المتعددة التدريج المستخدمة في هذه الدراسة (GRM, GPCM, NRM).

دالة المعلومات للفقرة: يُعد من المفاهيم الأساسية في نظرية الاستجابة للفقرة سواء كانت الفقرة ثنائية التدريج، أو تتطلب استجابات متعددة التدريج، وتساعد هذه الدالة في بناء أدوات قياس تجعل دقة القياس أو المعلومات أكبر ما يمكن، وتبين مدى مساهمة الفقرة في دالة المعلومات للاختبار بشكل مستقل

عن الفقرات الأخرى، وفي هذه الدراسة تم تقديرها باستخدام دالة الأرجحية العظمى.

دالة المعلومات للاختبار: وهي دالة رياضية تعبر عن مجموع دوال المعلومات لجميع فقرات الاختبار عند مستوى معين من القدرة، وفي هذه الدراسة كانت عبارة عن مجموع دوال المعلومات على الـ ٥٠ فقرة المولدة باستخدام برنامج WinGen 3..

تقدير المعالم للفقرات والأفراد: هي عملية التعبير الكمي عن المعالم باستخدام برنامج IRTPRO المعد من قبل كاي وثيزن وتويت (Cai, Thissen & Toit, 2011).

دقة التقدير: هو تعبير يشير لجودة التقدير التي يميزها الاحتمالية الكبيرة في أن التقدير قريب من القيمة الحقيقية، وفي هذه الدراسة تم استخدام مؤشر التحيز للمعلمة (Bias)، ومؤشر الجذر التربيعي لمعدل المربعات للخطأ (Root Mean Square Error, RMSE)، وذلك للحكم على دقة التقدير.

مؤشر التحيز Bias للمعلمة: وهو الفرق بين المعلمة المقدرة (estimated)، والمعلمة المولدة (الحقيقية)، والمعالم المولدة في هذه الدراسة هي المعالم الخاصة بالقدرات، والمعالم الخاصة بالفقرات (تمييز الفقرة، وصعوبة العتبة/ الخطوة) باستخدام المعادلة الآتية:

$$BIAS_x = \frac{\sum_{i=1}^{n}(\hat{x}_{estimated} - X_{true})}{n}$$

حيث $\hat{X}_{estimated}$: القيمة المقدرة للمعلمة، X_{true}: القيمة الحقيقية للمعلمة، n: عدد المعالم المقدرة.

الإحصائي (RMSE) : هو الجذر التربيعي لمعدل المربعات للخطأ (Root Mean Square Error, RMSE)، بين القيمة المقدرة والقيمة الحقيقية، وهو مؤشر لدقة تقدير المعالم، حيث يتم إيجاد القيمة المطلقة للفرق بين القيم

المولدة (الحقيقية)، والقيم المقدرة بأي طريقة من طرق التقدير، وفقا للمعادلة الآتية:

$$RMSE_x = \sqrt{\frac{\sum_{i=1}^{n} (\hat{X}_{estmated} - X_{true})^2}{n}}$$

محددات الدراسة

- اقتصرت الدراسة على اعتماد نموذج الاستجابة المتدرجة لساميجما (Graded Response Model, GRM)، ونموذج الاستجابة الإسمي لبوك (Nominal Response Model, NRM)، ونموذج التقدير الجزئي لموراكي (Generalized Partial Credit Model, GPCM)، من نماذج نظرية الاستجابة للفقرة المتعددة التدريج، ولذلك لا يمكن تعميم النتائج على كل النماذج.
- اقتصرت الدراسة على شكل واحد من أشكال فقرات الاختبار، وهي الفقرات متعددة التدريج.
- اقتصرت الدراسة على استخدام بيانات مولدة باستخدام برنامج WinGen 3.
- اقتصرت الدراسة على مؤشرين للحكم على دقة التقدير؛ هما: مؤشر التحيز، والجذر التربيعي لمعدل المربعات للخطأ.
- المصطلحات المستخدمة في هذه الدراسة، محددة في التعريفات الإجرائية، وبالتالي فإن إمكانية تعميم النتائج تتحدد في ضوء هذه التعريفات.

منهج الدراسة

توليد البيانات: تم توليد البيانات باستخدام برنامج (WinGen 3)، وهو من تصميم هان وهامبيلتون (Han & Hambelton, 2007)، ويمكن من خلال هذا البرنامج توليد استجابات لاختبارات أحادية البعد (ثنائية التدريج، متعددة التدريج)، بالإضافة إلى استجابات متعددة الأبعاد. كما أنه باستخدام البرنامج يمكن توليد استجابات لأكثر من مجموعة من المفحوصين، وبتوزيعات مختلفة،

وتستخدم ثلاثة توزيعات لتوليد قدرات المفحوصين وهي: التوزيع الطبيعي (Normal Distribution)، والتوزيع المنتظم (Uniform Distribution)، وأخيراً توزيع بيتا (Beta Distribution)، الذي من خلاله يمكن توليد توزيعات ملتوية سواء كانت ملتوية التواءً موجباً أو التواءً سالباً. وتمت عملية توليد البيانات وفق المراحل التالية:

المرحلة الأولى: تم توليد بيانات خاصة باختبار مكون من ٥٠ سؤالاً وفق نموذج التقدير الجزئي العام ((Generalized Partial Credit Model (GPCM باستخدام برنامج WinGen 3، وتم اختيار هذا الطول للاختبار حتى يكون مناسباً للوصول إلى دقة في التقديرات وفق النموذج (GPCM, GRM, NRM)، وقد تم توليد معالم فقرات الاختبار وفق الشروط التالية: تم توليد بيانات خاصة باختبار في عدد خطواته (خمس خطوات)، مكون من ٥٠ سؤالاً، بحيث كانت المسافة بين صعوبة الخطوة الأولى والخطوة الأخيرة (الخامسة) لها متساوية، حيث تراوحت بين -٣ و+٣ لوجيت، وفق التوزيع الطبيعي ((Normal~(0,1)، ومعلمة التمييز وفق التوزيع المنتظم بقيمة صغرى مقدارها ٠,٤، وقيمة عظمى مقدارها ١,٦.

المرحلة الثانية: تم توليد استجابات ١٥٠٠ مفحوص، وفق التوزيع الطبيعي ((Normal-(0,1)، حيث كان الوسط الحسابي والانحراف المعياري لهذا التوزيع ٠,٠٣٦٩، ٠,٩٩٨٣٢٤ على الترتيب، وكانت درجة الالتواء والتفلطح له ٠,٠٠٥، ٠,١٠٨ على الترتيب، بالاستناد إلى دراسة دراسجو (Drasgow, 1982) للوصول إلى دقة في التقديرات عند استخدام نموذج التقدير الجزئي العام، وكذلك وفق رأي لورد (Lord, 1980) بأن يكون عدد المفحوصين يتجاوز الـ ١٠٠٠ للحصول على أفضل التقديرات.

تحليل البيانات: قبل معالجة البيانات للإجابة عن أسئلة الدراسة تم التحقق من افتراض أحادية البعد للسمة المقيسة، باعتبار أن أحادية البعد أحد الافتراضات الأساسية التي يجب توافرها في البيانات عند تطبيق نموذج

التقدير الجزئي العام لتحقيق الهدف من الدراسة، قام الباحث باختيار استجابات ١٥٠٠ مفحوص على الاختبار رباعي الخطوات، وقد تم الحفاظ على طول الاختبار (٥٠)، وحجم العينة (١٥٠٠) لما لها من أثر في دقة التقديرات، وقد استخدم الباحث برنامج SPSS لتحليل البيانات المولدة، وتمت عملية التحليل من خلال التحقق من افتراضات النموذج؛ حيث تم التحقق من افتراضات نموذج التقدير الجزئي العام وهو أحادية البعد، لما له من أثر في دقة التقديرات، وذلك باستخدام التحليل العاملي للاختبار رباعي الخطوات باستخدام طريقة المكونات الرئيسة. وجرى التدوير باستخدام طريقة التدوير المتعامد (Varimax Rotation) للعوامل التي كانت قيمة الجذر الكامن لها أكبر من واحد. والجدول رقم (١) يبين قيم الجذر الكامن ونسب التباين المفسر للعاملين الأول والثاني وناتج قسمة قيمة الجذر الأول على العامل الثاني في الاختبار رباعي الخطوات.

جدول رقم (١)

قيم الجذر الكامن ونسب التباين المفسر للعامل الأول والثاني وناتج قسمة قيمة الجذر الأول على العامل الثاني في الاختبار رباعي الخطوات

الإحصائي	العامل الأول	العامل الثاني	ناتج القسمة
الجذر الكامن	١٠,٧٩٧	١,١٥٠	٩,٣٨٩
التباين المفسر	٢١,٥٩٣	٢,٣٠٠	

يتضح من الجدول رقم (١) أن قيمة الجذر الكامن للعامل الأول (١٠,٧٩٧)، وهي قيمة مرتفعة إذا ما قورنت بقيم الجذور الكامنة لبقية العوامل، وباعتماد قيمة محك الجذر الكامن كمؤشر على أحادية البعد، فقد ذكر لورد (As cited in Albanese & Forsyth, 1984) أن الفقرات تكون أحادية البعد، إذا كانت نسبة قيمة الجذر الكامن الأول إلى العامل الثاني كبيرة وتزيد على (٢)، وهذا متحقق في هذه الدراسة. والمحك الآخر هو نسبة التباين المفسّر، إذ بلغت نسبة التباين المفسّر للعامل الأول (٢١,٥٩٣٪) من التباين

الكلي، مما يؤكد وجود عامل طاغ في المقياس، وهو ما يتماشى مع ما اقترحه ريكاس (Rechase) عام ١٩٨٥، من أنه: إذا أمكن أن يفسّر العامل الأول (٢٠٪) من التباين المفسّر على الأقل فإن ذلك يعد مؤشراً كافياً لأحادية البعد (Hambelton & Swaminathan, 1985; Hattie, 1985).

أما فيما يتعلق بافتراض الاستقلال الموضعي (Local Independence)، وبهدف التحقق من هذا الافتراض تم استخدام برنامج (Local Dependence Indices for Polytomous Item) (LDIP)، وهو برنامج مكتوب بلغة فورتران ٩٥، يتم من خلاله حساب: Q_3، ZQ_3، G^2. وهي مؤشرات تستخدم للكشف عن الاستقلال الموضعي، حيث اقترحت ين (Yen, 1993) المؤشر Q_3 كمؤشر للكشف عن الارتباط الموضعي بين فقرات الاختبار، وهو معامل الارتباط للبواقي لزوج من الفقرات لجميع المفحوصين بعد ضبط السمة. أما مؤشر G^2 المستخدم من قبل شن وثيزن (Chen & Thissen, 1997) فقد صمم للكشف عن الفروق بين ما هو متوقع وما هو ملاحظ لتكرار الاستجابات لزوج من الفقرات، وهو يتوزع تقريباً حسب توزيع مربع كاي بدرجة حرية واحدة. أما مؤشر ZQ_3 الذي استخدمه شن (Shen, 1997) للكشف عن انتهاك افتراض الاستقلال الموضعي، حيث حملت دراسته عنوان "Quantifying Item Dependency By Fishers Z" وتم اعتماد نموذج راش كأساس في هذه الدراسة، ويركز هذا المعيار على حساب الارتباط الموضعي بين قيم ZQ_3 الفشرية التي تم حسابها للفقرات المستقلة، والمنطق من وراء توزيع قيم ZQ_3 للفقرات المستقلة ينبغي أن يكون توزيعاً طبيعياً بوسط قدره صفر، وبالتالي فإن أي قيمة من قيم ZQ_3 للفقرات إذا كانت تزيد بمقدار انحرافين معياريين لقيم ZQ_3 يمكن اعتبارها فقرات غير مستقلة موضعياً يتم اعتماد المؤشر ZQ_3 كحد فاصل بين أزواج الفقرات التي بينها انتهاك لافتراض الاستقلال الموضعي عن غيرها، فإذا كان المتوسط الحسابي لمؤشرات ZQ_3 الملاحظ الخاص بأزواج الفقرات واقعا بين الحد الأدنى لمؤشر ZQ_3 الخاص بأزواج الفقرات والحد الأعلى لمؤشر ZQ_3 الخاص بأزواج الفقرات، فهذا مؤشر على تحقق افتراض الاستقلال الموضعي (Kim, Cohen & Lin, 2005)، حيث تبين

أن قيمة الوسط الحسابي لمؤشر ZQ_3 الخاصة بأزواج الفقرات التي تقع بين الحد الأدنى وبين الحد الأعلى لمؤشر ZQ_3 الخاصة بأزواج الفقرات البالغ عددها (١٢٢٥) ويبين الجدول رقم (٢) مؤشرات الاستقلال الموضعي تبعاً لنظرية الاستجابة للفقرة.

جدول رقم (٢)

مؤشرات الاستقلال الموضعية تبعاً لنظرية السمات الكامنة

الإحصائي	القيمة
عدد الفقرات	٥٠
عدد أزواج الفقرات	١٢٢٥
القيمة الصغرى الملاحظة لمؤشر ZQ_3	-٠,٤٥٢١٦٣
القيمة العظمى الملاحظة لمؤشر ZQ_3	٠,٨٥٧٢١٠
المتوسط الحسابي لمؤشرات ZQ_3 الملاحظة الخاص بأزواج الفقرات	٠,٠٥٧٨٤٦٣

يلاحظ من الجدول رقم (٢) أن قيمة المتوسط الحسابي لمؤشر ZQ_3 الخاصة بأزواج الفقرات قد بلغ (٠,٠٥٧٨٤٦٣)، وهي واقعة بين الحد الأدنى والحد الأعلى لمؤشر ZQ_3 الخاصة بأزواج الفقرات، وهذا مؤشر على تحقق افتراض الاستقلال الموضعي (Kim, Cohen & Lin, 2005).

كما تم حساب عدد أزواج الفقرات التي وقعت ضمن فترة الثقة المحققة لشروط الاستقلال الموضعي، والجدول رقم (٣) يبين مؤشرات الاستقلال الموضعي تبعاً لنظرية الاستجابة للفقرة.

جدول رقم (٣)

مؤشرات الاستقلال الموضعي تبعا لنظرية الاستجابة للفقرة

حالة الاستقلال الموضعي	عدد الأزواج	النسبة المئوية
التبعية الموضعية	١٣٧	١١,١٨
الاستقلالية الموضعية	١٠٨٨	٨٨,٨٢
الكلي	١٢٢٥	١٠٠

يلاحظ من الجدول رقم (٣) أن عدد أزواج الفقرات التي وقعت ضمن حالة الاستقلالية الموضعية قد بلغت (١٠٨٨) حالة بنسبة مئوية بلغت (٨٨,٨٢٪)، في حين أن عدد الأزواج التي وقعت خارج حالة الاستقلالية الموضعية (التبعية الموضوعية) (١٣٧) وبنسبة مئوية بلغت (١١,١٨٪)، وهذا يبين أن عدد أزواج الفقرات التي حققت الاستقلالية أعلى من ثمانية أمثال عدد أزواج الفقرات التي حققت التبعية الموضعية. وهذا مؤشر على تحقق افتراض الاستقلال الموضعي (Kim, Cohen & Lin, 2005)، بالإضافة لذلك فقد أشار هامبلتون وسوامنيان (Hambelton & Swaminathan, 1985) إلى أن افتراض الاستقلال الموضعي يكافئ افتراض أحادية البعد، وهذا يعني انه إذا تحقق افتراض أحادية البعد في المقياس، فإن المقياس يحقق افتراض الاستقلال الموضعي.

نتائج الدراسة ومناقشتها

للإجابة عن أسئلة الدراسة تم استخدام برنامج (IRTPRO)؛ وذلك للحصول على تقديرات القدرة للأفراد ومعالم الصعوبة للفقرات، ودالة المعلومات للاختبار تبعاً لنموذج الاستجابة متعدد التدريج المستخدم (GPCM, GRM, NRM)، وكانت النتائج على النحو الآتي.

فيما يتعلق بالإجابة عن السؤال الأول: ما أثر نموذج الاستجابة المتعدد التدريج (نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM)، على دقة تقدير معالم القدرة للأفراد؟ فقد تم الحصول على تقدير قدرة واحدة لكل فرد في كل نموذج من النماذج الثلاثة المختلفة (GPCM, GRM, NRM)، وكل منها مكون من ٥٠ سؤالاً، حيث تراوحت بين -٣ و+٣ لوجيت، وتم حساب الأوساط الحسابية والانحرافات المعيارية لتقديرات القدرة للأفراد على المقياس تبعاً لنموذج الاستجابة متعدد التدريج المستخدم (GPCM, GRM, NRM)، وذلك كما في الجدول رقم (٤).

جدول رقم (٤)

الأوساط الحسابية والانحرافات المعيارية لتقديرات القدرة على المقياس تبعاً لنموذج الاستجابة

النموذج	الوسط الحسابي	الانحراف المعياري
GPCM	-٠,٠٠٣١٣٧	٠,٩٣٤٣٩٨
GRM	-٠,٠٠٠٠٠٣	١,٠٠٠٣٣٤
NRM	٠,٠٠٠٠٠١	١,٠٠٠٣٤٤

يلاحظ من الجدول رقم (٤) أن أقرب الأوساط من القيم الحقيقية هو الوسط الحسابي لنموذج الاستجابة الإسمي NRM، حيث كانت القيمة الحقيقية لتقديرات القدرة للمفحوصين، كما تبين من المرحلة الثانية من مراحل توليد البيانات المذكورة آنفاً تساوي (٠,٠٣٦٩)، فربما يكون هو النموذج الأكثر دقة في تقديراتهم بالرغم من عدم وجود فروق بين الأوساط الحسابية لتقديرات القدرة للأفراد على المقياس تبعاً للنموذج المستخدم؛ كما كشفت عنها نتائج تحليل التباين الأحادي للقياسات المتكررة لتقديرات القدرة تبعاً لنموذج الاستجابة متعدد التدريج المستخدم، المبينة في الجدول رقم (٥).

جدول رقم (٥)

نتائج تحليل التباين الأحادي للقياسات المتكررة لتقديرات القدرة على المقياس تبعاً لنموذج الاستجابة

آثار الاختبارات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل النماذج	القدرة	٠,٠١٠	١,٥٦٩	٠,٠٠٦	٠,٧٣٨	٠,٤٤٨
Greenhouse-Geisser	الخطأ (القدرة)	١٩,٩٦٨	٢٣٥١,٧٤٣	٠,٠٠٨٤٩١		
(٠,٧٨٤)	الكلي	١٩,٩٧٧	٢٣٥٣,٣١٢	٠,٠٠٨٤٨٩		
بين الأفراد	الخطأ	٤٢٨٨,٨٤١	١٤٩٩,٠٠٠	٢,٨٦١		
الكلي		٤٣٠٨,٨١٩	٣٨٥٢,٣١٢	١,١١٩		

ولمزيد من الدقة في معرفة النموذج الأكثر دقة في تقدير القدرة، فقد تم حساب الأوساط الحسابية والانحرافات المعيارية الخاصة بمؤشر التحيز BIAS كمؤشر من مؤشرات دقة التقدير لقدرات المفحوصين لنماذج نظرية الاستجابة للفقرة متعددة التدريج، وذلك كما هو مبين في الجدول رقم (٦).

جدول رقم (٦)

قيم مؤشر التحيز لمعلمة القدرة تبعاً لنموذج الاستجابة متعددة التدريج

مؤشر تحيز النموذج:	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
GPCM	٢,٤٦-	٠,٩٤	٠,٠٤٠٠٤-	٠,٢٨٨٣٧
GRM	٠,٩٢-	٠,٩٣	٠,٠٣٦٩١-	٠,٢٦٠٤٣
NRM	٢,٣٧-	١,٠٦	٠,٠٣٦٩٠-	٠,٢٨٥٩٠

يلاحظ من الجدول رقم (٦) أن القيمة المطلقة للوسط الحسابي لمؤشر التحيز كمؤشر دقة لتقديرات معلمة القدرة للمفحوصين قد كانت أقل ما يمكن لنموذج الاستجابة الإسمي NRM، يليه نموذج الاستجابة المتدرجة، وأعلى ما يمكن لنموذج التقدير الجزئي العام؛ مما يعني أن نموذج GPCM أقل دقة في تقدير معلمة القدرة للمفحوصين، وأن نموذج NRM كان هو النموذج الأكثر دقة، وهذه النتيجة تنسجم مع ما أشير إليه حول دقة النموذج في الجدول رقم (٤).

وكذلك فقد تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل المربعات للأخطاء RMSE كمؤشر آخر على دقة تقديرات القدرة للأفراد تبعاً لنموذج الاستجابة المتعدد التدريج (GPCM, GRM, NRM)، والجدول رقم (٧) يبين ذلك.

جدول رقم (٧)

الأوساط الحسابية والانحرافات المعيارية لقيم RSME لتقديرات القدرة للأفراد تبعاً لنموذج الاستجابة

النموذج	الوسط الحسابي	الانحراف المعياري
NRM	٠,٢٥٧١٧٣	٠,٠٣٦٧٥٢
GRM	٠,٢٦٥١٠٧	٠,٠٢٢٢٢٨
GPCM	٠,٢٧٢٠١٥	٠,٠٤٢٥٠١

يلاحظ من الجدول رقم (٧) وجود فروق ظاهرية بين الأوساط الحسابية لقيم RMSE لتقديرات القدرة للأفراد تبعاً لنموذج الاستجابة المستخدم؛ وللكشف عن دلالة الفروق الظاهرية؛ تم إجراء تحليل التباين الأحادي للقياسات المتكررة لقيم RMSE لتقديرات القدرة للأفراد على المقياس تبعاً لنموذج الاستجابة، وذلك كما في الجدول رقم (٨).

جدول رقم (٨)

نتائج تحليل التباين الأحادي للقياسات المتكررة لقيم RMSE لتقديرات القدرة للأفراد تبعاً لنموذج الاستجابة

آثار الاختبارات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل النماذج	الخطأ المعياري للقدرة	٠,١٦٥	١,١١٩	٠,١٤٨	٢٠١,١٥٣	٠,٠٠١
Greenhouse-Geisser	الخطأ (الخطأ المعياري للقدرة)	١,٢٣٣	١٦٧٧,٩٢٠	٠,٠٠٠٧٣٥		
(٠,٥٦٠)	الكلي	١,٣٩٨	١٦٧٩,٠٣٩	٠,٠٠٠٨٣٣		
بين الأفراد	الخطأ	٤,٢٤٠	١٤٩٩,٠٠٠	٠,٠٠٣		
الكلي		٥,٦٣٨	٣١٧٨,٠٣٩	٠,٠٠٢		

يتضح من الجدول رقم (٨) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لقيم RMSE لتقديرات القدرة للأفراد

على المقياس تبعاً للنموذج. وللكشف عن مواقع الفروق تم استخدام اختبار بنفيروني Bonferroni للمقارنات البعدية المتعددة الذي يقلص حجم الخطأ من النوع الأول بين الأوساط الحسابية لقيم RMSE لتقديرات القدرة للأفراد على المقياس، وذلك كما هو مبين في الجدول رقم (٩).

جدول رقم (٩)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لقيم RMSE لتقديرات القدرة تبعاً للنموذج

الخطأ المعياري للقدرة		GPCM	GRM	NRM
Bonferroni	الوسط الحسابي	٠,٢٧٢٠	٠,٢٦٥١	٠,٢٥٧٢
GPCM	٠,٢٧٢٠			
GRM	٠,٢٦٥١	٠,٠٠٧-*		
NRM	٠,٢٥٧٢	٠,٠١٥-*	٠,٠٠٨-*	

يتضح من الجدول رقم (٩) أنَّ هناك فرقاً دالاً إحصائياً بين الوسطين الحسابيين لـ RMSE لتقديرات القدرة للأفراد على المقياس لصالح نموذج الاستجابة الإسمي NRM مقارنة بكلٍّ من النموذجين (GPCM ثم GRM)، وهذا يعني أن نموذج الاستجابة الإسمي (نموذج بوك) NRM هو الأكثر دقة في تقدير القدرة مقارنة مع نموذجي: الاستجابة المتدرج GRM، ونموذج التقدير الجزئي العام GPCM، واختلفت نتائج هذه الدراسة مع دراسة دود (Dodd, 1994)، ودراسة أوستني (Ostini, 2001) اللتين أعطتا نتائج متشابهة لنماذج نظرية الاستجابة للفقرة المستخدمة، وتشابهت جزئياً مع نتائج دراسة كوك ودود وفيتسباترك (CooK, Dodd & Fitzpatrick, 1999)، التي أشارت إلى أنه فيما يخص تقديرات القدرة فإن النموذجين GRM، GPCM هما بنفس الجودة في تقدير معالم القدرات، ويمكن تفسير ذلك بأن نموذج NRM هو من أكثر النماذج عمومية في مجموعة نماذج القسمة على المجموع، وكل النماذج الأخرى هي حالات خاصة (مقيدة بمحددات) من نموذج الاستجابة الإسمي، وهو قابل للتطبيق على

الفقرات التي لها بدائل ولا يمكن ترتيبها؛ لتمثل درجات مختلفة من السمة المقيسة عن طريق الفقرة، ويحاول هذا النموذج زيادة الدقة في تقديرات Θ للأفراد، وخاصة ذوي القدرة المنخفضة باستخدام المعلومات من إجاباتهم غير الصحيحة، كما أن النموذج يتضمن معلم تمييز ومعلم تقاطع يعكس التفاعل بين صعوبة الفئة وتمييز الفئة (Bock, 1972)، وربما يعود ذلك أيضاً إلى أن نموذج NRM مناسب للإجابات الإسمية أو الرتبية، ومن أكثر النماذج عمومية.

فيما يتعلق بالإجابة عن السؤال الثاني: ما أثر نموذج الاستجابة المتعدد التدريج (نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM)، على دقة تقدير معالم التمييز للفقرات؟ تم حساب الأوساط الحسابية والانحرافات المعيارية لتقديرات التمييز لفقرات المقياس تبعاً لنموذج الاستجابة متعدد التدريج المستخدم (GPCM, GRM, NRM)، وذلك كما في الجدول رقم (١٠).

جدول رقم (١٠)

الأوساط الحسابية والانحرافات المعيارية لمعالم التمييز لفقرات المقياس تبعاً لنموذج الاستجابة

النموذج	الوسط الحسابي	الانحراف المعياري
GPCM	٠,٣٨٦٢٠٠	٠,٢٠٤٧٥٩
GRM	٠,٩٨١٠٠٠	٠,٣٥٨٦٩٠
NRM	٠,٣٩٨٢٠٠	٠,٢٠٧٣٥٢

يلاحظ من الجدول رقم (١٠) أن قيم متوسط قيم معالم التمييز باستخدام نموذج الاستجابة المتدرجة البالغ (٠,٩٨١٠٠٠) هو الأقرب من القيمة الحقيقية لمتوسط قيم معالم التمييز والبالغة (٠,٩٨٤)، وربما يعد ذلك مؤشراً على أن نموذج الاستجابة المتدرجة هو الأكثر دقة من النموذجين الآخرين في تقدير معالم التمييز للفقرات، حيث كشفت نتائج تحليل التباين الأحادي للقياسات المتكررة لتقديرات معالم التمييز عن وجود فروق ذات دلالة

إحصائية عند مستوى الدلالة α=٠,٠١ بين الأوساط الحسابية لمعالم التمييز لفقرات المقياس تبعاً لنموذج الاستجابة والجدول رقم (١١) يبين نتائج التحليل.

جدول رقم (١١)

نتائج تحليل التباين الأحادي للقياسات المتكررة لتقديرات معالم التمييز لفقرات المقياس تبعاً لنموذج الاستجابة

آثار المقياس خماسي التدريجات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل النماذج	التمييز	١١,٥٦٠	١,٠٣١	١١,٢٠٧	٤٩٢,٧٩٦	٠,٠٠١
Greenhouse-Geisser	الخطأ (التمييز)	١,١٤٩	٥٠,٥٤١	٠,٠٢٢٧٤٢		
(٠,٥١٦)	الكلي	١٢,٧٠٩	٥١,٥٧٢	٠,٢٤٦٤٣٥		
بين الفقرات	الخطأ	٩,٣١٦	٤٩,٠٠٠	٠,١٩٠		
الكلي		٢٢,٠٢٥	١٠٠,٥٧٢	٠,٢١٩		

وللكشف عن مواقع الفروق، تم استخدام اختبار Bonferroni للمقارنات البعدية المتعددة بين الأوساط الحسابية لمعالم التمييز لفقرات المقياس تبعاً للنموذج المستخدم، وذلك كما هو في الجدول رقم (١٢).

جدول رقم (١٢)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لمعالم التمييز لفقرات المقياس تبعاً للنموذج

التمييز		GPCM	NRM	GRM
Bonferroni	الوسط الحسابي	٠,٣٨٦٢	٠,٣٩٨٢	٠,٩٨١٠
GPCM	٠,٣٨٦٢			
NRM	٠,٣٩٨٢	❖٠,٠١٢		
GRM	٠,٩٨١٠	❖٠,٥٩٥	❖٠,٥٨٣	

يتضح من الجدول رقم (١٢) أنَّ هناك فروقاً دالة إحصائياً بين الأوساط

الحسابية لمعالم التمييز وكان أعلاها لنموذج تقدير الاستجابة المتدرج GRM مقارنة بكلٍّ من النموذجين (GPCM ثم NRM)، وكذلك يتضح من الجدول رقم (١١) أنَّ هناك فرقاً دالاً إحصائياً بين الوسطين الحسابيين لمعالم التمييز لصالح نموذج الاستجابة الإسمي GRM مقارنة بنموذج الاستجابة المتدرج NRM. ولمزيد من الدقة في معرفة النموذج الأكثر دقة في تقدير معالم التمييز فقد تم حساب الأوساط الحسابية، والانحرافات المعيارية الخاصة بمؤشر التحيز BIAS كمؤشر من مؤشرات دقة التقدير لمعالم التمييز للفقرات تبعاً لنماذج نظرية الاستجابة للفقرة متعددة التدريج، وذلك كما هو مبين في الجدول رقم (١٣).

جدول رقم (١٣)

قيم مؤشر التحيز لمعلمة التمييز تبعاً لنموذج الاستجابة متعددة التدريج

مؤشر تحيز التمييز وفق نموذج:	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
GPCM	١,٤٤٨٠-	٠,٦٧٠٠	٠,٥٩٧٠-	٠,٤٦٧٦
NRM	١,٤٤٨٠-	٠,٥٥٠٠	٠,٥٨٥٠-	٠,٤٦٨٢
GRM	١,١٧٨٠-	١,٢٣٠٠	٠,٠٠٢٢-	٠,٥٦٤٠

يلاحظ من الجدول رقم (١٣) أن القيمة المطلقة للوسط الحسابي لمؤشر التحيز كمؤشر دقة لتقديرات معلمة التمييز للفقرات قد كانت أقل ما يمكن لنموذج الاستجابة الإسمي GRM، يليه نموذج الاستجابة الإسمي، وأعلى ما يمكن لنموذج التقدير الجزئي العام؛ مما يعني أن نموذج GPCM أقل دقة في تقدير معلمة التمييز للفقرات، وأن نموذج GRM كان هو النموذج الأكثر دقة، وهذه النتيجة تنسجم مع ما أشير إليه حول دقة النموذج في الجدول رقم (١٠).

وكذلك فقد تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل المربعات للأخطاء RMSE كمؤشر آخر على دقة تقديرات لمعالم التمييز للفقرات تبعاً لنموذج الاستجابة المتعدد التدريج (GPCM، GRM، NRM)، والجدول رقم (١٤) يبين ذلك.

جدول رقم (١٤)

الأوساط الحسابية والانحرافات المعيارية للأخطاء المعيارية لتقديرات التمييز تبعاً لنموذج الاستجابة

النموذج	الوسط الحسابي	الانحراف المعياري
GPCM	٠,٠٤٢٤	٠,٠٢١٥
NRM	٠,٠٦٥٠	٠,٠٠٩٧
GRM	٠,٠٣٠٢	٠,٠١٢٤

يلاحظ من الجدول رقم (١٤) وجود فروق ظاهرية بين الأوساط الحسابية لقيم RMSE لتقديرات التمييز للفقرات تبعاً لنموذج الاستجابة المستخدم؛ وللكشف عن دلالة الفروق الظاهرية؛ تم إجراء تحليل التباين الأحادي للقياسات المتكررة لقيم RMSE لتقديرات التمييز لفقرات المقياس تبعاً لنموذج الاستجابة، وذلك كما في الجدول رقم (١٥).

جدول رقم (١٥)

نتائج تحليل التباين الأحادي للقياسات المتكررة لقيم RMSE لتقديرات التمييز للفقرات تبعاً لنموذج الاستجابة

آثار المقياس خماسي التدريجات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل النماذج	الخطأ المعياري للتمييز	٠,٠٣١	١,٢٧٧	٠,٠٢٤	١٦٥,٦٤٥	٠,٠٠١
Greenhouse-Geisser	الخطأ (RMSE للتمييز)	٠,٠٠٩	٦٢,٥٦٩	٠,٠٠٠١٤٧		
(٠,٦٣٨)	الكلي	٠,٠٤٠	٦٣,٨٤٦	٠,٠٠٠٦٣٣		
بين الفقرات	الخطأ	٠,٠٢٦	٤٩,٠٠٠	٠,٠٠١		
الكلي		٠,٠٦٦	١١٢,٨٤٦	٠,٠٠١		

يتضح من الجدول رقم (١٥) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠١ بين الأوساط الحسابية لقيم RMSE لتقديرات التمييز

لفقرات المقياس تبعاً للنموذج؛ وللكشف عن مواقع الفروق فقد تم استخدام اختبار بنفيروني Bonferroni للمقارنات البعدية المتعددة الذي يقلص حجم الخطأ من النوع الأول بين الأوساط الحسابية للأخطاء المعيارية لتقديرات التمييز لفقرات المقياس، وذلك كما هو مبين في الجدول رقم (١٦).

جدول رقم (١٦)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية للأخطاء المعيارية لتقديرات التمييز للفقرات تبعاً للنموذج

GRM	GPCM	NRM	الخطأ المعياري للتمييز	
٠,٠٣٠٢	٠,٠٤٢٤	٠,٠٦٥٠	الوسط الحسابي	Bonferroni
			٠,٠٦٥٠	NRM
		٠,٠٢٣-*	٠,٠٤٢٤	GPCM
	٠,٠١٢-*	٠,٠٣٥-*	٠,٠٣٠٢	GRM

يتضح من الجدول رقم (١٦) أنَّ هناك فرقا دالاً إحصائياً بين الوسطين الحسابيين للأخطاء المعيارية لتقديرات التمييز لفقرات المقياس لصالح نموذج الاستجابة المتدرجة GRM، مقارنة بكلٍّ من النموذجين (NRM و GPCM)؛ وهذا يعني أن GRM هو النموذج الأكثر دقة في تقدير معالم التمييز للفقرات مقارنة بالنموذجين GPCM و NRM، واختلفت مع نتائج أوستني (Ostini, 2001) التي أعطت نتائج متشابهة في التقدير، ودراسة كوك ودود وفيتسباترك (CooK, Dodd & Fitzpatrick, 1999) التي أشارت إلى أنه فيما يخص تقديرات معالم التمييز فإن النموذجين GRM، GPCM هما بنفس الجودة في تقدير معالم التمييز، ويمكن تفسير ذلك بأن هذا النموذج يعد امتداداً لطريقة ثيرستون (Turnstone's Method) في تحليل الفقرات ذات الاستجابات المتدرجة في الاختبارات التربوية، كما يعد تطبيقاً لنموذج راش وتعميماً لنموذج بيرنبوم الثنائي المعلمة.. وقد اقترحت ساميجما هذا النموذج، وقدمت من خلاله تطبيقات مهمة للقياس مثل: قياس الاتجاه والشخصية، والنواحي المعرفية، حيث يتم منح درجة جزئية على الحلول الصحيحة للمشكلات المقدمة، وهذا

النموذج يمثل العلاقة المنحنية بين مستوى قدرات الأفراد، واحتمال استجابتهم في كل خطوة من خطوات الإجابة، ولا يشترط في هذا النموذج أن تكون جميع فقراته تشتمل على نفس العدد من الخطوات، والتي يرمز لها بالرمز، (B1, B2,....., Bn)، ومجموعة من بارامترات العتبات الفارقة (Thresholds) (a) بارامتر التمييز، وعدد هذه البارامترات أقل من عدد خطوات الاستجابة بواحد صحيح، وهو عدد المنحنيات المميزة.

فيما يتعلق بالإجابة عن السؤال الثالث: ما أثر نموذج الاستجابة المتعدد التدريج (نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM)، على دقة تقدير معالم الصعوبة للعتبات الفارقة (Thresholds) للفقرات؟ تم حساب الأوساط الحسابية والانحرافات المعيارية لتقديرات معالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً لنموذج الاستجابة متعدد التدريج المستخدم (GPCM, GRM, NRM)، وذلك كما في الجدول رقم (١٧).

جدول رقم (١٧)

الأوساط الحسابية والانحرافات المعيارية لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً لنموذج الاستجابة متعدد التدريج

عتبات الصعوبة	العدد	النموذج						الكلي	
		GPCM		GRM		NRM			
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
الأولى	٥٠	-٥,٠١	٧,٦١	-١,٠٩	٠,٨٥	٠,٠٠	٠,٢٣	-٢,٠٤	٢,٩٠
الثانية	٥٠	-١,٥٢	٧,١٤	-٠,٢١	٠,٦٥	-١,٢٩	١,٤٨	-١,٠١	٣,٠٩
الثالثة	٥٠	١,٧٠	٦,٥٠	٠,٣٨	٠,٥٥	٠,٠٣	٠,٩٥	٠,٧٠	٢,٦٦
الرابعة	٥٠	٤,٨٣	٧,٢٦	١,١٣	٠,٦٣	-٠,٠٣	٠,٨٦	١,٩٨	٢,٩٢
الكلي		٠,٠٠	٧,١٣	٠,٠٥	٠,٦٧	-٠,٣٢	٠,٨٨		

يتبين من الجدول رقم (١٧) وجود فروق ظاهرية بين الأوساط الحسابية لمعالم عتبات الصعوبة لفقرات مقياس خماسي التدريج (الأولى، الثانية، الثالثة،

الرابعة)، ناتجة عن اختلاف النموذج المستخدم، كما يلاحظ بأن الأوساط الحسابية لعتبات الصعوبة للفقرات لنموذج الاستجابة المتدرجة GRM كانت هي الأقرب للتقديرات الحقيقية لمتوسطات عتبات الصعوبة للفقرات التي تم توليدها (-٠,٩٢، -٠,٢٩، ٠,٣٣، ١,١٤) على الترتيب، ؛ وكشفت نتائج تحليل التباين الثنائي للقياسات المتكررة لمعالم الصعوبة للعتبات الفارقة للفقرات وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لمعالم الصعوبة للعتبات الفارقة للفقرات تبعاً للنموذج المستخدم، والجدول رقم (١٨) يبين نتائج التحليل.

جدول رقم (١٨)

نتائج تحليل التباين الثنائي للقياسات المتكررة لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً لنموذج الاستجابة

آثار الاختبارات لـ:	آثار الاختبارات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل:	النماذج	النموذج	١٦,٤٩٨	١,٤٢٥	١١,٥٧٧	١٠,١٩٣	٠,٠٠١
	Greenhouse-Geisser (0.713)	الخطأ (النموذج)	٧٩,٣١١	٦٩,٨٢٥	١,١٣٦		
	صعوبة العتبات	صعوبة العتبة	١٤٢٧,٨٥١	٢,٤٨٨	٥٧٣,٨٧٤	٢٢,٥٨٨	٠,٠٠١
	Greenhouse-Geisser (0.829)	الخطأ (صعوبة العتبة)	٣٠٩٧,٣٩٥	١٢١,٩١٦	٢٥,٤٠٦		
	النماذج × الصعوبة	النموذج × صعوبة العتبة	١٤٤٩,٥١٩	٢,٨٢٥	٥١٣,٠٨٥	١٠,٠٥٨	٠,٠٠١
	Greenhouse-Geisser (0.471)	الخطأ (النموذج × الصعوبة)	٧٠٦١,٤٤٣	١٣٨,٤٣٠	٥١,٠١١		
	الكلي لمصدر التباين		٢٨٩٣,٨٦٨	٦,٧٣٨	٤٢٩,٤٧٢		
	الكلي (الخطأ)		١٠٢٣٨,١٤٩	٣٣٠,١٧٢	٣١,٠٠٩		
	الكلي		١٣١٣٢,٠١٧	٣٣٦,٩١٠	٣٨,٩٧٨		
بين الفقرات	الخطأ		٢٩,٨٢٨	٤٩	٠,٦٠٩		
الكلي			٢٦٢٩٣,٨٦٢	٧٢٢,٨٢٠	٣٦,٣٧٧		

يتضح من الجدول رقم (١٨)؛ ولتحديد مصادر الفروق فقد تم استخدام اختبار Bonferroni للمقارنات البعدية المتعددة بين الأوساط الحسابية لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً للنموذج المستخدم، والجدول رقم (١٩) يبين نتائج التحليل.

جدول رقم (١٩)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً للنموذج

النموذج		NRM	GPCM	GRM
Bonferroni	الوسط الحسابي	٠,٣٢٤٤٠-	٠,٠٠٠٣٥-	٠,٠٤٩٧٠
NRM	٠,٣٢٤٤٠-			
GPCM	٠,٠٠٠٣٥-	٠,٣٢*		
GRM	٠,٠٤٩٧٠	٠,٣٧*	٠,٠٥	

يتضح من الجدول رقم (١٩) أنَّ الفرق بين الوسطين الحسابيين لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس قد كان أكبر بفارق جوهري وفقاً للنموذج GRM مقارنة بالنموذج NRM، وكذلك أنَّ الفرق بين الوسطين الحسابيين لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس قد كان أكبر بفارق جوهري وفقاً للنموذج GPCM مقارنة بالنموذج NRM. كما يتضح من الجدول رقم (١٩) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لمعالم الصعوبة تبعاً لمستوى العتبة (الأولى، الثانية، الثالثة، الرابعة)؛ وللكشف عن مواقع الفروق فقد تم استخدام اختبار Bonferroni للمقارنات البعدية بين الأوساط الحسابية لمعالم الصعوبة لفقرات تبعاً لمستوى العتبة الفارقة، والجدول رقم (٢٠) يبين ذلك.

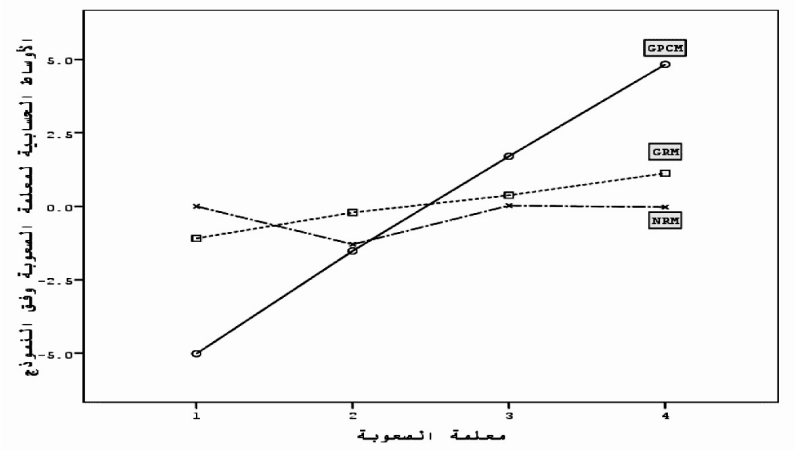
جدول رقم (٢٠)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لمعالم الصعوبة لفقرات المقياس تبعاً لمستوى لعتبة الفارقة للصعوبة

صعوبة العتبة		الأولى	الثانية	الثالثة	الرابعة
Bonferroni	الوسط الحسابي	-٢,٠٣٥	-١,٠٠٨	٠,٧٠١	١,٩٧٦
الأولى	-٢,٠٣٥				
الثانية	-١,٠٠٨	١,٠٣			
الثالثة	٠,٧٠١	٢,٧٤*	١,٧١*		
الرابعة	١,٩٧٦	٤,٠١*	٢,٩٨*	١,٢٧*	

يتضح من الجدول رقم (٢٠) أنَّ الفرق بين الوسطين الحسابيين لمعالم الصعوبة للعتبات الفارقة قد كان أكبر بفارق جوهري تبعاً لمستوى الصعوبة للعتبة الفارقة الرابعة مقارنة بكلٍّ من مستويات الصعوبات للعتبات (الأولى، ثم الثانية، ثم الثالثة)، وكذلك يتضح من الجدول رقم (٢٠) أنَّ الفرق بين الوسطين الحسابيين لمعالم الصعوبة للفقرات قد كان أكبر بفارق جوهري تبعاً لصعوبة العتبة الفارقة الثالثة مقارنة بكلٍّ من مستويات الصعوبات للعتبات (الأولى، ثم الثانية).

وأخيراً؛ يتضح من الجدول رقم (١٨) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية تعزى للتفاعل بين لمعالم الصعوبة للعتبات الفارقة والنموذج الاستجابة المستخدم؛ وللكشف عن مواقع الفروق فقد تم التمثيل بياني الذي يوضح التفاعل بين النموذج وبين مستويات الصعوبة للعتبات الفارقة الأربع، وذلك كما هو مبين في الشكل رقم (٢).



الشكل رقم (٢): تمثيل بياني يوضح التفاعل بين نموذج الاستجابة وبين مستويات العتبات الفارقة للصعوبة الأربع

يلاحظ من الشكل رقم (٢) أن الأوساط الحسابية لمعلمتي الصعوبة لمستويات العتبات الفارقة (الثالثة والرابعة) وفقاً للنموذج NRM هي أقل منها وفقاً للنموذج GRM وهي بدورها أقل منها للنموذج GPCM. كما يلاحظ من الشكل رقم (٢) أن الأوساط الحسابية لمعالم الصعوبة للعتبة الفارقة الثانية تبعاً للنموذج GPCM هي أقل منها للنموذج NRM وهي بدورها أقل منها للنموذج GRM. وأخيراً؛ يلاحظ من الشكل رقم (٢) أن الأوساط الحسابية لمعلمة الصعوبة للعتبة الفارقة الأولى وفقاً للنموذج GPCM هي أقل منها للنموذج GRM وهي بدورها أقل منها للنموذج NRM. ولمزيد من الدقة في معرفة النموذج الأكثر دقة في تقدير معالم الصعوبة للعتبات تبعاً لنموذج الاستجابة متعددة التدريج، فقد تم حساب الأوساط الحسابية والانحرافات المعيارية الخاصة بمؤشر التحيز BIAS كمؤشر من مؤشرات دقة التقدير لمعالم الصعوبة للعتبات تبعاً لنموذج الاستجابة متعددة التدريج، وذلك كما هو مبين في الجدول رقم (٢١).

جدول رقم (٢١)

الأوساط الحسابية لقيم مؤشرات التحيز لمعالم الصعوبة للعتبات تبعاً لنموذج الاستجابة متعددة التدريج

العتبة	النموذج	القيمة الصغرى	القيمة العظمى	الوسط الحسابي	الانحراف المعياري
الأولى	GPCM	-٣٠,٠١٢٠	٣,٩٢٠٠	-٤,٠٩٩٢	٧,٥١٧٠
	GRM	-٣,٠٨٠٠	٢,٤٣٢٠	-٠,١٧٤٤	١,٢٤٤٣
	NRM	-٠,٤٣٨٠	٢,٥٩٣٠	٠,٩١٣٢	٠,٦٥٤٠
الثانية	GPCM	-٢٤,٤٦٦٠	١٨,١٦٢٠	-١,٢٣٣٣	٧,١٢٩٧
	NRM	-٤,٠٤٨٠	٢,٣٩٣٠	-١,٠٠٨٥	١,٣٩٣٩
	GRM	-٢,٠٥١٠	١,٥٩٤٠	٠,٠٧٣٧	٠,٨٥٨٨
الثالثة	NRM	-٢,٨٧٤٠	٢,٨٥١٠	-٠,٣٠١٨	١,١٢٨٠
	GRM	-١,٥٥٣٠	١,٥٤١٠	٠,٠٤٨٢	٠,٧٦٠٨
	GPCM	-١٦,٤٣٩٠	١٨,٣٦٨٠	١,٣٧٦٢	٦,٥٣٩٩
الرابعة	NRM	-٣,٣٩٠٠	١,٦٠٠٠	-١,١٧٠٠	١,٠٧٣٠
	GRM	-٢,٣٩٠٠	١,٢٨٠٠	-٠,٠١٨٢	٠,٩١٠٣
	GPCM	-٣,٩٦٠٠	٢٦,٣٥٠٠	٣,٦٨٥٤	٧,٢٨٩١

يلاحظ من الجدول رقم (٢١) أن القيمة المطلقة للوسط الحسابي لمؤشر التحيز كمؤشر دقة لتقديرات معلمة الصعوبة للعتبات الفارقة قد كانت أقل ما يمكن لنموذج الاستجابة المتدرجة GRM، يليه نموذج الاستجابة الإسمي، وأعلى ما يمكن لنموذج التقدير الجزئي العام؛ مما يعني أن نموذج GPCM أقل دقة في تقدير معلمة الصعوبة للعتبات الفارقة، وأن نموذج GRM كان هو النموذج الأكثر دقة، وهذه النتيجة تنسجم مع ما أشير إليه حول دقة النموذج في الجدول رقم (١٧).

وكذلك تم حساب الأوساط الحسابية والانحرافات المعيارية للجذر التربيعي لمعدل المربعات للأخطاء RMSE كمؤشر آخر على دقة تقديرات

الصعوبة للعتبات تبعاً لنموذج الاستجابة المتعدد التدريج (GPCM، GRM، NRM)، والجدول رقم (٢٢) يبين ذلك.

جدول رقم (٢٢)

الأوساط الحسابية والانحرافات المعيارية للأخطاء المعيارية لمعالم الصعوبة للعتبات الفارقة لفقرات تبعاً لنموذج الاستجابة

الخطأ المعياري لعتبات الصعوبة	العدد	النموذج						الكلي	
		GPCM		GRM		NRM			
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
الأولى	٥٠	٠,٩٣	١,١٣	٠,١١	٠,٠٦	٠,٠٣	٠,٠١	٠,٣٥	٠,٤٠
الثانية	٥٠	١,٠٢	١,١٤	٠,٠٨	٠,٠٤	٠,١٣	٠,٠٧	٠,٤١	٠,٤٢
الثالثة	٥٠	١,١٣	١,٥٠	٠,٠٧	٠,٠٣	٠,١٠	٠,١٠	٠,٤٣	٠,٥٤
الرابعة	٥٠	١,٠٤	١,٦٠	٠,١١	٠,٠٦	٠,١١	٠,٠٨	٠,٤٢	٠,٥٨
الكلي		**١,٠٣**	**١,٣٥**	**٠,٠٩**	**٠,٠٥**	**٠,٠٩**	**٠,٠٦**		

يلاحظ من الجدول رقم (٢٢) وجود فروق ظاهرية بين الأوساط الحسابية لقيم RMSE لمعالم الصعوبة للعتبات الفارقة ناتجة عن اختلاف النموذج المستخدم، ومستويات الصعوبات للعتبات الفارقة (الأولى، الثانية، الثالثة، الرابعة)؛ وللكشف عن جوهرية الفروق الظاهرية؛ تم إجراء تحليل التباين الثنائي للقياسات المتكررة لقيم RMSE لمعالم الصعوبة للعتبات الفارقة لفقرات المقياس تبعاً للنموذج المستخدم، والجدول رقم (٢٣) يبين ذلك.

جدول رقم (٢٣)

نتائج تحليل التباين الثنائي للقياسات المتكررة لقيم RMSE لمعالم الصعوبة للعتبات الفارقة تبعاً للنموذج ومستوى العتبة

آثار الاختبارات لـ:	آثار الاختبارات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل:	النماذج	النموذج	١١٦,٦٠٧	١,٠٠٦	١١٥,٩٥٢	٣٧,٤٤٥	٠,٠٠١
	Greenhouse-Geisser (0.503)	الخطأ (النموذج)	١٥٢,٥٩٢	٤٩,٢٧٧	٣,٠٩٧		
	RMSE لصعوبة العتبات	الخطأ المعياري للصعوبة	٠,٥٥٢	١,٤٧٥	٠,٣٧٤	٠,٦٣٥	٠,٤٨٧
	Greenhouse-Geisser (0.492)	الخطأ (RMSE لصعوبة العتبات)	٤٢,٦٢٤	٧٢,٢٥٢	٠,٥٩٠		
	النماذج × RMSE لصعوبة العتبات	النموذج × RMSE لصعوبة العتبات	٠,٧٩٦	١,٥٢٨	٠,٥٢١	٠,٤٨٤	٠,٥٦٨
	Greenhouse-Geisser (0.525)	الخطأ (النموذج × RMSE لصعوبة العتبات)	٨٠,٦١٥	٧٤,٨٦٨	١,٠٧٧		
	الكلي لمصدر التباين		١١٧,٩٥٥	٤,٠٠٨	٢٩,٤٢٩		
	الكلي (الخطأ)		٢٧٥,٨٣١	١٩٦,٣٩٧	١,٤٠٤		
	الكلي		٣٩٣,٧٨٦	٢٠٠,٤٠٥	١,٩٦٥		
بين الفقرات	الخطأ		٨٩,٠١٧	٤٩	١,٨١٧		
الكلي			٨٧٦,٥٨٩	٤٤٩,٨١١	١,٩٤٩		

يتضح من الجدول رقم (٢٣) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لقيم RMSE لصعوبة العتبات الفارقة لفقرات المقياس تبعاً للنموذج؛ وللكشف عن مواقع الفروق فقد تم استخدام اختبار Bonferroni للمقارنات البعدية المتعددة بين الأوساط الحسابية لـ RMSE لصعوبة العتبات لمعالم الصعوبة للعتبات الفارقة للفقرات تبعاً للنموذج، وذلك كما هو مبين في الجدول رقم (٢٤).

جدول رقم (٢٤)

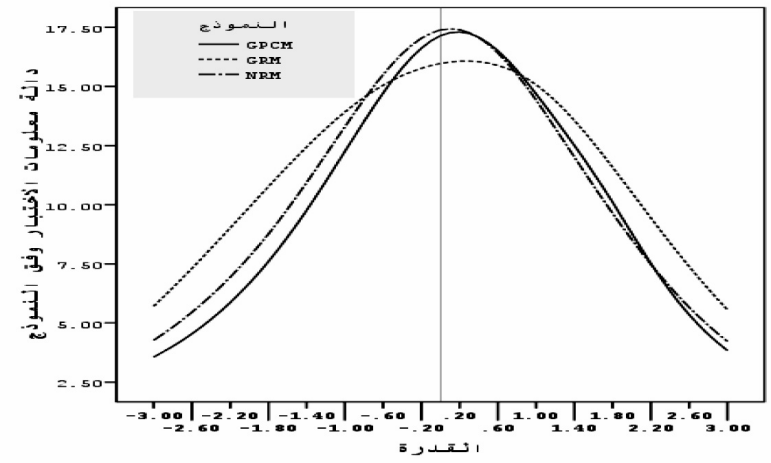
نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لقيم RMSE لمعالم الصعوبة للعتبات الفارقة تبعاً للنموذج

النموذج		GPCM	NRM	GRM
Bonferroni	الوسط الحسابي	١,٠٢٦٢٥	٠,٠٩١٢٥	٠,٠٩٠٩٠
GPCM	١,٠٢٦٢٥			
NRM	٠,٠٩١٢٥	٠,٩٣٥٠-*		
GRM	٠,٠٩٠٩٠	٠,٩٣٥٤-*	٠,٠٠٠٣-	

يتضح من الجدول رقم (٢٤) أنَّ الفرق بين الوسطين الحسابيين لـ RMSE لمعالم الصعوبة للعتبات الفارقة قد كان جوهرياً لصالح نموذج الاستجابة المتدرجة GRM مقارنة بالنموذج GPCM؛ وهذا يعني أن هذا النموذج GRM هو الأكثر دقة في تقدير الصعوبة للعتبات الفارقة من نموذج GPCM، واختلفت نتائج هذه الدراسة مع دراسة دود (Dodd, 1994)، ودراسة أوستني (Ostini, 2001) اللتين أشارتا إلى نتائج متشابهة في تقدير معالم الصعوبة للعتبات، ويمكن تفسير ذلك بأن هذا النموذج يعد من نماذج الفرق (Difference Models)، ويستخدم في هذا النموذج الطرح للحصول على احتمالية الإجابة لفئة معينة، وهي مناسبة للإجابات الرتبية، واستخدمت نماذج المنحنى الطبيعي واللوجستي كوحدات (Lustina, 2004)، ويتصف هذا النموذج بإمكانية الفصل الجبري بين قدرات الأفراد، ومعالم الفواصل بين مختلف مستويات السؤال، وكذلك إمكانية تقدير معالم القدرات للأفراد، ومعالم الفواصل للأسئلة بغياب أي منهما، وتحت شرط تحقق الشرط الكافي لاستخراج المعلم الآخر (Masters, 1988)، وكذلك يتضح من الجدول رقم (٢٥) أنَّ الفرق بين الوسطين الحسابيين لـ RMSE لصعوبة العتبات للعتبات الفارقة كان أكبر بفارق جوهري لصالح نموذج الاستجابة الإسمي NRM مقارنة بالنموذج GPCM؛ أي أن نموذج NRM هو أكثر دقة في تقدير معالم الصعوبة للعتبات الفارقة من نموذج GPCM.

كما يتضح من الجدول رقم (٢٤) عدم وجود فروق دالة إحصائياً (α=٠,٠٥) بين الأوساط الحسابية لقيم RMSE لمعالم الصعوبة للعتبات الفارقة تبعاً لمستوى العتبة (الأولى، الثانية، الثالثة، الرابعة)؛ وهذا يعني عدم وجود فروق جوهرية في الصعوبة للعتبات الفارقة تعزى للنموذج المستخدم في التحليل. وأخيراً؛ يتضح من الجدول رقم (٢٥) عدم وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية للأخطاء المعيارية لمعالم الصعوبة للعتبات تعزى للتفاعل بين النموذج وصعوبة العتبة.

وفيما يتعلق بالإجابة عن السؤال الرابع حول أثر نموذج الاستجابة المتعدد التدريج (نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM)، على دالة المعلومات للاختبار؟، فقد تم إيجاد منحنيات دالة المعلومات للمقياس تبعاً لنموذج الاستجابة متعدد التدريج (GPCM، GRM، NRM)، ويبين الشكل رقم (٣) هذه المنحنيات.



الشكل رقم (٣): التمثيل البياني لدالة المعلومات لمقياس خماسي التدريج تبعاً لنموذج الاستجابة

يلاحظ من الشكل رقم (٣) أن نموذج الاستجابة NRM قدم أقصى كمية من المعلومات عند مستوى القدرة المتوسطة يليه نموذج الاستجابة GPCM وهو

بدوره قدم أقصى كمية من المعلومات من نموذج الاستجابة GRM. كما يلاحظ من الشكل رقم (٣) أن المقياس حسب نموذج الاستجابة المتدرجة GRM قدم معلومات أكبر لدى الأفراد ذوي القدرة المتدنية القدرة مما هي عليه وفقاً لنموذج الاستجابة NRM وهو بدوره قدم معلومات أكبر مما هي عليه وفقاً لنموذج التقدير الجزئي العام GPCM، وأخيراً؛ يلاحظ من الشكل رقم (٣) أن المقياس لدى الأفراد ذوي القدرة العالية قدم معلومات وفقاً لنموذج الاستجابة NRM أكبر مما هي عليه وفقاً لنموذج الاستجابة GPCM وهو بدوره قدم معلومات أكبر مما هي عليه وفقاً لنموذج الاستجابة NRM، ومن حيث القيم الرقمية فإن الجدول رقم (٢٥) يبين متوسطات كميات المعلومات التي يقدمها المقياس تبعاً لنموذج الاستجابة ومستوى القدرة للمستجيب.

الجدول (٢٥)

الأوساط الحسابية والانحرافات المعيارية لكميات المعلومات تبعاً لنموذج الاستجابة ومستوى القدرة

مستوى القدرة	العدد	GPCM		GRM		NRM		الكلي	
		الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري	الوسط الحسابي	الانحراف المعياري
منخفض	٨٠٠	٠,١١٩	٠,٠٨٠	٠,١٧٧	٠,١٣١	٠,١٣٩	٠,٠٩٩	٠,١٤٥	٠,١٠٣
متوسط	١٤٥٠	٠,٢٩٣	٠,٢١٣	٠,٢٩٩	٠,١٩٣	٠,٣٠٠	٠,٢٠٠	٠,٢٩٧	٠,٢٠٢
مرتفع	٨٠٠	٠,١٥٠	٠,١٩٠	٠,١٨٥	٠,١٤٣	٠,١٤٩	٠,١٣٠	٠,١٦١	٠,١٥٤
الكلي	٣٠٥٠	٠,٢١٠	٠,١٩٨	٠,٢٣٧	٠,١٧٦	٠,٢١٨	٠,١٧٩		

يلاحظ من الجدول رقم (٢٥) أن هناك فروقاً ظاهرية بين متوسطات كمية المعلومات التي يقدمها المقياس تبعاً لنموذج الاستجابة المستخدم، فمثلاً قدم نموذج الاستجابة GRM أكبر كمية من المعلومات عند ذوي القدرة المنخفضة والعالية، في حين أن نموذج الاستجابة NRM قدم أكبر كمية من المعلومات عند ذوي القدرة المتوسطة، وللكشف عن الدلالة الإحصائية للفروق بين كميات المعلومات التي يقدمها المقياس تبعاً لنموذج الاستجابة ومستوى

القدرة للمستجيب، فقد تم استخدام تحليل التباين الثنائي للقياسات المتكررة، والجدول رقم (٢٦) يبين نتائج التحليل.

الجدول رقم (٢٦)

نتائج تحليل التباين الثنائي للقياسات المتكررة تبعاً لنموذج الاستجابة ومستوى القدرة للمستجيب

آثار الاختبارات لـ:	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط مجموع المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
داخل	النماذج	١,٦١	١,٢٥	١,٢٩	٣٧٣,٧٠	٠,٠٠١
Greenhouse-Geisser (0.627)	النماذج× مستويات القدرة	٠,٩٢	٢,٥١	٠,٣٧	١٠٦,٨٦	٠,٠٠١
	الخطأ (النماذج)	١٣,١٥	٣٨٢١,١١	٠,٠٠		
	الكلي	١٥,٦٩	٣٨٢٤,٨٨	٠,٠٠		
بين	مستويات القدرة	٤٧,٨٨	٢	٢٣,٩٤	٢٩٢,١٣	٠,٠٠١
	الخطأ	٢٤٩,٧١	٣٠٤٧	٠,٠٨		
	الكلي	٢٩٧,٥٩	٣٠٤٩	٠,١٠		
الكلي		٣١٣,٢٨	٦٨٧٣,٨٨	٠,٠٥		

يتضح من الجدول رقم (٢٦) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لكميات المعلومات التي يقدمها المقياس تبعاً للنموذج المستخدم؛ وللكشف عن مواقع الفروق فقد تم استخدام اختبار Bonferroni للمقارنات البعدية المتعددة بين الأوساط الحسابية لكميات المعلومات التي يقدمها المقياس تبعاً للنموذج، وذلك كما هو مبين في الجدول رقم (٢٧).

جدول رقم (٢٧)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لكميات المعلومات تبعاً للنموذج

النموذج		GPCM	NRM	GRM
Bonferroni	الوسط الحسابي	٠,٢١٠	٠,٢١٨	٠,٢٣٧
GPCM	٠,٢١٠			
NRM	٠,٢١٨	٠,٠٠٩❖		
GRM	٠,٢٣٧	٠,٠٣٣❖	٠,٠٢٤❖	

يتضح من الجدول رقم (٢٧) أنَّ الفرق بين الوسطين الحسابيين لكميات المعلومات التي يقدمها المقياس كان أكبر بفارق جوهري تبعاً للنموذج لصالح نموذج GRM مقارنة بكل من النموذجين GPCM، NRM، كما يتضح أنَّ الفرق بين الوسطين الحسابيين لكميات المعلومات التي يقدمها المقياس كان أكبر بفارق جوهري تبعاً للنموذج NRM مقارنة بالنموذج GPCM.

كما يتضح من الجدول رقم (٢٧) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لكميات المعلومات التي يقدمها المقياس تبعاً لمستوى القدرة للمستجيب، وللكشف عن مواقع الفروق فقد تم استخدام اختبار Bonferroni للمقارنات البعدية المتعددة بين الأوساط الحسابية لكميات المعلومات التي يقدمها المقياس تبعاً لمستوى القدرة، والجدول رقم (٢٨) يبين ذلك.

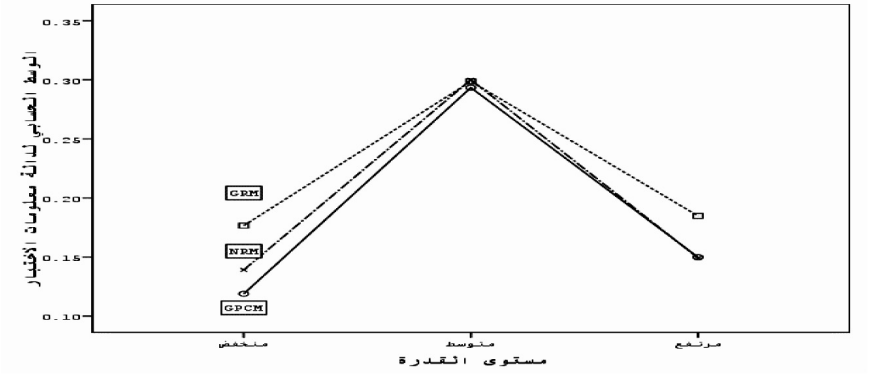
جدول رقم (٢٨)

نتائج اختبار Bonferroni للمقارنات البعدية المتعددة للأوساط الحسابية لكميات المعلومات تبعاً لمستوى القدرة للمستجيب

مستوى القدرة		منخفض	مرتفع	متوسط
Bonferroni	الوسط الحسابي	٠,١٤٥	٠,١٦١	٠,٢٩٧
منخفض	٠,١٤٥			
مرتفع	٠,١٦١	٠,٠١٦		
متوسط	٠,٢٩٧	٠,١٥٣❖	٠,١٣٦❖	

يتضح من الجدول رقم (٢٨) أنَّ الفرق بين الوسطين الحسابيين لكميات المعلومات التي يقدمها المقياس قد كان أكبر بفارق جوهري لصالح الأفراد ذوي مستوى القدرة المتوسطة مقارنة بكل من مستويي القدرة المنخفض والمرتفع، كما يتضح أنَّ الفرق بين الوسطين الحسابيين لكميات المعلومات التي يقدمها المقياس لم يكن جوهرياً تبعاً لمستويي القدرة المنخفض والمرتفع.

وأخيراً؛ يتضح من الجدول رقم (٢٨) وجود فروق دالة إحصائياً عند مستوى الدلالة α=٠,٠٥ بين الأوساط الحسابية لكميات المعلومات التي يقدمها المقياس تعزى للتفاعل بين نموذج الاستجابة المستخدم ومستوى القدرة للمستجيب؛ وللكشف عن مواقع الفروق فقد تم تمثيل ذلك بيانياً الذي يوضح التفاعل بين النموذج وبين مستويات القدرة للمستجيبين، وذلك كما هو مبين في الشكل رقم (٤).



الشكل رقم (٤): التمثيل البياني للتفاعل بين نموذج الاستجابة وبين مستويات القدرة للمستجيبين

يلحظ من التمثيل البياني في شكل رقم (٤) الذي يمثل تفاعلاً رتبياً أن كمية المعلومات التي يقدمها المقياس باستخدام نماذج الاستجابة الثلاثة عند الأفراد ذوي القدرة المتوسطة تفوق كمية المعلومات التي يقدمها المقياس عند الأفراد ذوي القدرة المرتفعة وذوي القدرة المتدنية، وأن نموذج الاستجابة المتدرج

يقدم معلومات أكبر من نموذج الاستجابة الإسمي ونموذج التقدير الجزئي العام عند الأفراد ذوي القدرتين المتدنية والعالية، واختلفت نتائج هذه الدراسة مع دراسة دود (Dodd, 1994) التي أشارت إلى تشابه النماذج المستخدمة في تقديم المعلومات، وتشابهت جزئياً مع نتائج دراسة كوك ودود وفيتسباترك (Cook, Dodd & Fitzpatrick, 1999) التي أشارت إلى أنه فيما يخص كمية المعلومات فإن النموذجين GRM، GPCM هما بنفس الجودة في تقديم المعلومات عند جميع مستويات القدرة، ويمكن تفسير ذلك بما أشار إليه أمبرتسون وريز (Embertson & Reise, 2000) بأن هذا النموذج يمكن بواسطته استخلاص أكبر قدر من المعلومات فيما يتعلق بمستويات القدرة أو السمة المراد قياسها، فهذا النموذج يمثل العلاقة غير الخطية بين مستوى قدرة الفرد، واحتمال استجابته في كل عتبة فارقة، ويتم معالجة السؤال في هذا النموذج كما لو كان سلسة من العتبات/ الخطوات الثنائية، وبالتالي فإن كل خطوة تساهم في زيادة كمية المعلومات التي يقدمها المقياس.

وللإجابة عن سؤال الدراسة حول أثر نموذج الاستجابة المستخدم على دقة التقديرات لكمية المعلومات التي يقدمها المقياس المستخدم، فقد تم حساب محكي: أكيكي للمعلومات Akaike Information Criterion (AIC)، ومحك بيز للمعلومات Bayesian Information Criterion (BIC)، تبعاً لنموذج الاستجابة المتعدد التدريج (GPCM، GRM، NRM)، حيث يكون نموذج الاستجابة المتعدد التدريج ذو القيمة الأقل هو النموذج المناسب لتحليل البيانات (Since both the AIC and BIC are the smallest for a specific model, we conclude that this model provides the better fit of the data. (IRTPRO Guide version: 3.1.21505.4001 may 2015) والجدول رقم (٢٨) يبين ذلك.

جدول رقم (٢٨)

الإحصائيات الخاصة بمحكي أكيكي وبيز للمعلومات تبعاً لنموذج الاستجابة المتعدد التدريج

الموذج	الإحصائيات القائمة على log likelihood	القيمة
GPCM	-2 loglikelihood:	185913.13
	Akaike Information Criterion (AIC)	186413.13
	Bayesian Information Criterion (BIC)	187741.43
GRM	-2 loglikelihood:	185418.65
	Akaike Information Criterion (AIC)	185918.65
	Bayesian Information Criterion (BIC)	187246.95
NRM	-2 loglikelihood:	185440.18
	Akaike Information Criterion (AIC)	186240.18
	Bayesian Information Criterion (BIC)	188365.46

يلحظ من الجدول رقم (٢٨) أن نموذج الاستجابة GRM كان هو الأكثر دقة ومناسبة في تحليل البيانات لأن قيمة محك أكيكي كانت هي الأقل، وكذلك قيمة محك بيز هي الأقل.

فيما يتعلق بالإجابة عن السؤال الخامس: ما أثر نموذج الاستجابة المتعدد التدريج (نموذج الاستجابة المتدرجة GRM، ونموذج الاستجابة الإسمي NRM، ونموذج التقدير الجزئي العام GPCM)، على ثبات الاختبار؟ تم تقدير الثبات باستخدام طريقة الاتساق الداخلي باستخدام معادلة كرونباخ α الذي يعتمد على متوسطات الأخطاء المعيارية لتقديرات القدرة للمفحوصين، ويظهر ذلك من خلال التحليل في برنامج (IRTPRO)، ويبين الجدول رقم (٢٩) قيم معاملات الثبات المقدرة تبعاً لنموذج الاستجابة المتعددة التدريج.

جدول رقم (٢٩)

معاملات ثبات كرونباخ α تبعاً لنموذج الاستجابة المتعدد التدريج

النموذج	معامل كرونباخ α
GPCM	٠,٩٣٢٠
GRM	٠,٩٣٥٣
NRM	٠,٩٣٦٤

يلاحظ من الجدول رقم (٢٩) أن قيم معاملات الثبات باستخدام معادلة كرونباخ α كانت قيمها متساوية إلى حدٍ ما من خلال نماذج الاستجابة المتعددة التدريج. ولمعرفة دلالة الفروق بين قيم كرونباخ ألفا الثلاثة تبعاً لنموذج الاستجابة، فقد استخدم الباحث الإحصائي (M) المقترح من هاكستين وولن (Hakstin & Whalen, 1976) الذي يتبع توزيع X^2 بدرجات حرية تساوي (عدد المعاملات -١) وتحسب قيمة M من الصيغة الرياضية الآتية:

$$M = \frac{(J-1)(9n-11)^2}{18J(n-1)}\left[k - \frac{\left[\sum_{k=1}^{3}(1-r_k)^{-\frac{1}{3}}\right]^2}{\sum_{k=1}^{3}(1-r_k)^{-\frac{2}{3}}}\right]$$

حيث تمثل الرموز: K عدد نماذج الاستجابة متعددة التدريج، n: عدد أفراد المجموعة في النموذج المستخدم r_k: معامل ثبات النموذج j: عدد فقرات الاختبار في النموذج المستخدم.

وقد كشفت نتائج التقديرات الخاصة بقيم كرونباخ ألفا تبعاً لنموذج الاستجابة المتعدد التدريج عن عدم وجود فروق بين معاملات ثبات كرونباخ ألفا، حيث كانت قيمة الإحصائي المحسوبة (٢,٠٩) وهي أقل من القيمة الحرجة لكاي تربيع بدرجات حرية (٢) عند مستوى الدلالة الإحصائية (α=٠,٠٥).

خلاصة وتوصيات

كشفت نتائج التحليل حول أثر نموذج نظرية الاستجابة للفقرات ذات الاستجابات المتعددة على دقة تقدير معالم الفقرة وتقديرات القدرة للأفراد والخصائص السيكومترية للاختبار عن الآتي:

- أن نموذج الاستجابة الإسمي هو النموذج الأكثر دقة في تقدير القدرة من نموذج الاستجابة المتدرجة، ونموذج التقدير الجزئي العام (نموذج موراكي)، وأن نموذج الاستجابة المتدرجة هو أكثر دقة في تقدير القدرة من نموذج التقدير الجزئي العام.

- أن نموذج الاستجابة المتدرجة GRM هو النموذج الأكثر دقة في تقدير معالم التمييز للفقرات مقارنة بالنموذجين GPCM وNRM، وأن نموذج الاستجابة الإسمي NRM هو الأكثر دقة في تقدير التمييز من نموذج GPCM.

- أن نموذج GRM هو الأكثر دقة في تقدير معالم الصعوبة للعتبات الفارقة من نموذج التقدير الجزئي العام GPCM، وأن نموذج الاستجابة الإسمي NRM هو أكثر دقة في تقدير معالم الصعوبة للعتبات الفارقة من نموذج GPCM.

- أن نموذج GRM قدم أقصى كمية من المعلومات التي يقدمها المقياس مقارنة بكل من النموذجين GPCM، NRM، وأن النموذج NRM قدم أقصى كمية من المعلومات مقارنة بالنموذج GPCM، وأن نموذج الاستجابة المتعدد التدريج هو النموذج الأكثر مناسبة ودقة في تحليل البيانات.

- كشفت نتائج التقديرات الخاصة بقيم كرونباخ ألفا عن عدم وجود فروق بين معاملات ثبات كرونباخ ألفا تبعاً لنموذج الاستجابة المستخدم، ويمكن تلخيص ترتيب النماذج من حيث دقة التقدير كما في الجدول رقم (٣٠).

جدول رقم (٣٠)

ترتيب نماذج الاستجابة المتعددة التدريج من حيث دقة التقدير

الترتيب	النموذج من حيث دقة تقدير				
	القدرة	التمييز	الصعوبة	كمية المعلومات	الثبات
الأول	NRM	GRM	GRM	GRM	لا توجد فروق
الثاني	GRM	NRM	NRM	GPCM	
الثالث	GPCM	GPCM	GPCM	NRM	

واعتماداً على ذلك فإن الباحث يوصي مطوري الاختبارات بإمكانية استخدام نموذج الاستجابة الإسمي (نموذج بوك) في تقدير معالم القدرة للأفراد، وتحديداً عندما تكون عينة المفحوصين كبيرة، حيث تكون دقة التقدير لمعالم القدرة هي الفضلى مقارنة مع نموذجي الاستجابة للفقرة المتعددة (GRM، GPCM)، واستخدام نموذج الاستجابة المتدرجة GRM لتقدير معالم التمييز للفقرات، ومعالم الصعوبة للعتبات الفارقة (Thresholds)، وكمية المعلومات التي يقدمها الاختبار/ المقياس وخاصة للأفراد ذوي القدرات المتدنية والعالية مقارنة مع نموذجي الاستجابة للفقرة المتعددة التدريج (GPCM, NRM)، وربما تساهم هذه التوصية في تمكين مطوري بنوك الأسئلة والاختبارات التكيفية (المفصلة) (Adaptive Testing) عند تفعيل استخدام الاختبارات المتعددة التدريج على نطاق أوسع، كما يمكن أن تقدم هذه الدراسة دليلاً امبريقياً بأن تقدير الثبات باستخدام نماذج الاستجابة للفقرة الثلاثة تبقى قيمته إلى حد بعيد مستقرة، كما يوصي الباحث بإجراء دراسات أخرى مماثلة لكن باستخدام بيانات محاكاة مولدة باستخدام البرامج الإحصائية، ودراسة المقارنة بين الطرق الثلاث في دقة التقدير لمعالم الفقرات والأفراد تحت ظروف خاصة من توزيعات مستويات القدرة (طبيعي، منتظم، ملتو التواءً سالباً، ملتو التواءً موجباً). كذلك أخذ حجم العينة وعدد الفقرات كمتغيرين مستقلين، كما يمكن إعادة الدراسة نفسها باستخدام نماذج أخرى من نماذج نظرية

الاستجابة للفقرة كنماذج: الفترات المتتابعة لروست (Successive Interval Model, SIM)، ونموذج سلم التقدير لأندريش (Andrich Rating Scale Model, ARSM)، ونموذج سلم التقدير لموراكي (Rating Scale Model, MRSM)، وغيرها، وإعادة إجراء الدراسة باستخدام بيانات واقعية من خلال تطبيق مقاييس التقدير والاختبارات المقالية بأشكالها المختلفة للوصول إلى نتائج أكثر قابلية للتعميم.

Impact of Polytomous IRT Model on Accuracy of Estimation of Individuals' Abilities and the Psychometric Properties of Items and Test.

Dr. Nedal K. Al-Shraifin
College of Education - Yarmouk University
K.H.J

Abstract

The study aims at exploring the impact of polytomous IRT model used in analysis (NRM, GPCM, GRM)* on the accuracy of estimation of individuals' abilities and the psychometric properties of items and test. To achieve the purpose of the study, WinGen 3 program was used to simulate data for a test which consists of 50 items according to GPCM model. Each item consists of five levels, with four thresholds. Results indicated that the GPCM model is more accurate in estimating abilities than both GRM and NRM models. Additionally, it was found that the NRM model is more accurate in estimating discrimination parameters of items, compared with the other two models; GRM model was more accurate in estimating the difficulty parameters for thresholds than the other two models; GRM model presents maximum amount of information of the test, compared with the other two models, and it was more relevant and accurate in analyzing data. Results also indicated that there were no significant differences in coefficient values due to Polytomous IRT model.

المراجع

١ - الزعبي، محمد (٢٠٠٣). فاعلية نموذج التقدير الجزئي في بناء فقرات بنك أسئلة متعددة الخطوات في مادة الكيمياء للصف الثاني الثانوي العلمي. رسالة دكتوراه غير منشورة، عمان: جامعة عمان العربية للدراسات العليا.

٢ - علام، صلاح الدين (٢٠٠٥). **نماذج الاستجابة للمفردات الاختيارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي** (ط١). القاهرة: دار الفكر العربي.

٣ - عوده، أحمد (٢٠١٥). **القياس والتقويم في العملية التدريسية** (ط٥)، إربد: دار الأمل للنشر والتوزيع.

4 - Adams, R. J. (1988). Applying the partial credit model to educational diagnosis. **Applied Measurement in Education**, 1(4), 347-378.

5 - Albanese, M. & Forsyth, R. (1984). The one - two and modified two-Parameter Latent trait models: An empirical study of relative fit**. Education and Psychological Measurement**, 44,229- 246

6 - Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more nominal categories**. Psychometrika**, 37, 29-51.

7 - Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). **IRTPRO: User guide**. Lincolnwood, IL: Scientifice Software International.

8 - Chen, W. H. & Thissen, D. (1997). Local independence indexes for item pairs using item response theory. **Journal of Educational and Behavioral Statistics**, 22(3), 265-289.

9 - Cook, K.F., Dodd, B.G., & Fitzpatrick, S.J. (1999). A comparison of three polytomous item response theory models in the context of testtet scoring. **Journal of Outcome Measurement**, 3(1), 1-20.

10 - De Ayala, R. J., Dodd, B. G. & Koch, W. R. (1992). A comparison of partial credit and graded response models in computerized and adaptive testing. **Applied Measurement Education**, 5, 17-34.

11 - Dodd, B.G. (1985). Attitude scaling: A Comparison of the Graded and Partial Credit Latent Models. (Doctoral Dissertation, Unitversity of Texas, 1984), **Dissertation Abstracts International**, 45, 2074A.

12 - Dodd, B., Ayala, R. & Koch, W. (1995). Computerized adaptive testing with Polytomous items. **Applied Psychological Measurement**, 19, 5-22.

13 - Drasgow, H. (1982). What constitutes an adequate sample size for calibrating the graded response model (GRM) and the three - parameter logistic (3PL) model, Retrieved January 15, 2012, from, <http:// www.measuredprogress.org/documents/.../Adequate> Sample Size. pdf.

14 - Embretson, S. & Reise, S. (2000). **Item response theory for psychologists. Mahawah**, NJ: Erlbaum.

15 - Hakstin, A. R. & Whalen, T. E. (1976). A k- sample significance test for independent alpha coefficients. **Psychometrika**, 41(2), 219- 231.

16 - Hambleton, R. K. & Swaminthan, H. (1985). **Item response theory principle and application**. Boston: Kluwer: Nijhoff publishing.

17 - Han, K.T. & Hambleton, R.K (2007). User's Manual for WinGen: Windows Software that Generated IRT Model Parameter and Item Response. Center for Educational Assessment Research Report No 642, Amherst, MA: University of Massachusetts Center for Educational Assessment.

18 - Harris, J., Laan, S. & Mossenson, I. (2009). Applying partial credit analysis to the construction of narrative writing. Retrieved January 15, 2012 from <http://www.informaworld.com/smpp/title~content=t775653631>, Online publication date: 14 December 2009.

19 - Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. **Applied Psychological Measurement**, 9 (2), 139 - 164.

20 - Huggins, A. & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in Mplus. **Structural Equation Modeling: A Multidisciplinary Journal**, University of Florida, online 1-11-2015.

21 - Jiang, Sh., Wang, Ch. & Weiss, D. (2016). Sample size requirements for estimation of item parameters in multidimensional graded response

model. **Frontiers in Psychology**, 7(109), www.forntiersin.org <http:// www.forntiersin.org>, February 2016.

22 - Kim, S. (2001). An evaluation of a Markov Chain Monte Carlo Method for the Rasch model. **Applied Psychological Measurement**, 25(2), 163-176.

23 - Lord, F. M. (1980). **Application of item response theory to practical testing problems**. Hillsadle, New Jersey: Lawrence Erlbaum Associates.

24 - Lustina, M. (2004). **A comparison of Andrich rating scale model and Rosts successive intervals model**. [on- Line]: http:www.lip.utexas.edu/etd/d/2004/Lustinam52242/Lustinam52242.pdf.

25 - Masters, G. N. (1988). The analysis of partial credit scoring. **Applied Measurement in Education**, 1(4), 149-174

26 - Muraki, E. (1990). Fitting a Polytomous item response model to Liket-Type scale data. **Applied Psychological Measurement**, 14, 59 - 71.

27 - Muraki, E. (1992). A Generalized partial credit model: application of an EM algorithm. **Applied Psychological Measurement**, 16, 159- 176.

28 - Samejima, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation**. Psychometrica**, 59 (3), 307-329.

29 - Shen, L. (1997). Quantifying item dependency by fishers Z. Paper presented at the annual meeting of the American educational research associate-Ion, Chicago, IL, March 24-28, 1997 (ERIC Document Reproduction Service No. ED 410241).

30 - Sijtsma, K. & Hemker, B. (2000). Taxonomy of IRT Models for Ordering of Persons and Items Using Simple Sum Scores**. Journal of Educational and Behavioral Statistics**, 25, 391-415.

31 - Sung, H. J., & Kang, T. (2006). Choosing a polytomous IRT model using Bayesian model selection methods. Paper Presented at the National Council on Measurement in Education annual meeting in San Francisco, April, 2006.

32 - Susan, N. B. (2001). To meet or not to meet standards: Proficiency estimation using different Polytomous IRT models. University of Washington. Degree: PHD.

33 - Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. **Psychometrika**, 51, 567-577.

34 - Walford, G.; Tucher, E.; & Viswanathan, M.C. **The SAGE Handbook of Measurement**. Los Angles, SAGE.

35 - Wang, S. & Wang, T. (2001). Precision of RE warms weighted likelihood estimation for Polytomous model in computerized adaptive testing**. Applied Psychological Measurement**, 25, (4), 317-331.

36 - Wang, S. & Wang, T. (2002). **Relative precision of ability estimation in Polytomous CAT: A Comparison under the generalized partial credit model and graded response model**. ED 477926, shudonwang, 19500 Bulverederoad, SanAntonion, TX 28259-3701.

37 - Yen, W. (1993). Scaling performance assessment: strategies for managing local item dependence. **Journal of Educational Measurements**, 30(3), 187-213.