47
المجلد 12 العدد 2 يونيو 2011

دراسة استكشافية لملاءمة نماذج نظرية الاستجابة للمفردة في بناء اختبار تحصيلي من إعداد المعلم

د. هند عبد المجيد الحموري
قسم علم النفس التربوي
كلية العلوم التربوية- الجامعة الهاشمية

دراسة استكشافية لملاءمة نماذج نظرية الاستجابة للمفردة في بناء اختبار تحصيلي من إعداد المعلم

د. هند عبد المجيد الحموري
قسم علم النفس التربوي
كلية العلوم التربوية- الجامعة الهاشمية

الملخص

هدفت هذه الدراسة إلى استقصاء إمكان استخدام نماذج نظرية الاستجابة للمفردة في بناء اختبار تحصيلي من إعداد المعلم. تم بناء اختبار تحصيلي مكون من (٢٣) فقرة من نمط الاختيار من أربعة بدائل تقيس العدد نفسه من الأهداف المقصودة. طبق الاختبار على (٢٨٤) طالباً وطالبة. وحللت البيانات الناتجة من الاستجابة للفقرات وفق نظرية الاستجابة للمفردة. وكشفت نتائج الدراسة عدم إمكان استخدام هذه النظرية لبناء الاختبارات ذات العدد القليل من الفقرات.

الكلمات المفتاحية: الاستجابة للمفردة، بناء اختبار، اختبار تحصيلي من إعداد المعلم.

* تاريخ تسلم البحث: ٢٠٠٩/١٠/٢٦م
* تاريخ قبوله للنشر: ٢٠١٠/٣/٢٣م

An Exploratory Study of the Applicability of Item Response Theory Models to Construct Teacher Made Achievement Test

Dr. Hind A. Alhammouri
Faculty of Educational Sciences
Hashemite University

Abstract

This study aimed at investigating the applicability of item response theory (IRT) models to construct a teacher made achievement test. A test consisting of 23-multiple choice items measuring 23 intended goals was constructed and applied to (284) pupils. Item responses were analyzed according to IRT. Results of the study revealed non-applicability of the IRT to construct tests consisting of few numbers of items.

Key words: item response theory, test construction, teacher made achievement test.

دراسة استكشافية لملاءمة نماذج نظرية الاستجابة للمفردة في بناء اختبار تحصيلي من إعداد المعلم

د. هند عبد المجيد الحموري
قسم علم النفس التربوي
كلية العلوم التربوية- الجامعة الهاشمية

المقدمة

تؤدي الاختبارات دوراً مهمًّا في العملية التربوية. ولما كانت الاختبارات التحصيلية أكثر أدوات ووسائل التقويم استخداماً في المدرسة، فقد نالت عملية بنائها وتطويرها اهتماماً كبيراً من التربويين، ذلك لأنه يتم الاعتماد عليها، غالباً، في صنع القرارات التي تتعلق بعمليتي التعلم والتعليم. وتعتمد صحة القرارات المتخذة على نوع ودقة المعلومات والنتائج والتغذية الراجعة التي تزود الاختبارات بها. وعليه، فإن ضبط الاختبارات يؤدي إلى ضبط عملية التعلم. إلا أن العديد من الدراسات وتقارير الندوات والاجتماعات المتعددة للجان وجمعيات التقويم في العالم بعامة، والعالم العربي بخاصة، تدل على وجود ضعف واضح في معرفة المعلمين بمواصفات الاختبار الجيد، ووجود قصور في إعدادها (الصيداوي، ٢٠٠٤). فهي تفتقر إلى توافر الجوانب الفنية عند بنائها وفي أثناء عمليات التطبيق، وما يترتب على ذلك من آثار سلبية على مصداقية النتائج المستمدة من تلك الاختبارات (الثبيتي، ٢٠٠٢).

لقد سيطرت النظرية الكلاسيكية في القياس، ولفترة طويلة من الزمن، على عملية بناء الاختبارات وتحليلها وتفسير نتائجها، وتحديد مقدار السمات الكامنة وراء استجابات وأداء المفحوصين في الاختبارات المختلفة. غير أن هذه النظرية تعاني من بعض العيوب، فقيم برامترات الفقرة تعتمد على عينة المفحوصين؛ كما أن درجة الفرد–التي يفترض فيها أنها تعبر عن قدرته– تعتمد على عينة الفقرات التي يختبر بها (Hambleton & Swaminathan, 1985). هذا بالإضافة إلى أن قياس الأفراد في ذيلي التوزيع (ذوي القدرة العليا والدنيا)، وفق هذه النظرية، قد يكون أقل دقة من قياس ذوي القدرة المتوسطة (Crocker & Algina, 1986)، ومن ثمّ يصعب مقارنة نتائج المفحوصين الذين يتعرضون لاختبارات مختلفة. وفضلاً عن ذلك، فإن علامة المفحوص في الموقف الاختباري تمثل متغيراً عشوائيا؛ بمعنى أن احتمال الحصول على علامة معطاة في الاختبار تحدد بشكل مستقل عن التوزيع المختلف

لكل مفحوص، ومركب الخطأ الكلي يعتمد على أخطاء القياس التي أثرت في علامات المفحوصين بشكل مختلف (Shultz & Whitney, 2005).

وبهدف التغلب على عيوب النظرية الكلاسيكية، تم تقديم «نظرية الاستجابة للمفردة». وهذه النظرية تأخذ باعتبارها الأنماط الملاحظة للاستجابة على كل فقرة (Weiss, 1995). فقد يجيب مفحوص إجابة صحيحة عن فقرات صعبة ويتميز بشكل كبير ويحصل على علامة أعلى على السمة الكامنة من آخر أجاب عن نفس العدد من الفقرات ذات الصعوبة والتمييز المنخفضين. وعليه، فإن علامة المفحوص وفق هذه النظرية تكون حساسة للفروق في أنماط استجابة الأفراد، وتعطي تقديرا أفضل للمستوى الحقيقي على متصل السمة من مجرد جمع العلامات كما في النظرية الكلاسيكية (Santor & Ramsay, 1998). وهذه المعلومة قد تكون مفيدة لمطور الاختبار؛ فهي تزوده بصورة كاملة عن كيفية عمل الفقرة (Crocker & Algina, 1986) الأمر الذي يفيد في تحسين قرارات تطوير الاختبار.

تفترض نظرية الاستجابة للمفردة نماذج رياضية للسمات الكامنة، فهي تفترض أن القيمة الاحتمالية لاستجابة مفحوص لفقرة اختبارية دالة لكل من قدرة المفحوص، التي يفترض أن الاختبار يقيسها، وخصائص الفقرة التي يحاول المفحوص الإجابة عنها. ومن ثمّ، فإن الدالة المميزة لخصائص الفقرة تحدد العلاقة بين مقدار هذا الاحتمال وبين البرامترات المناظرة لقدرة المفحوص وخصائص الفقرة. ووفق نماذج هذه النظرية، فإن تقديرات قدرة المفحوصين متحررة من تقديرات برامترات الفقرات، كما أن الأخيرة متحررة من خصائص المفحوصين الذين طبقت عليهم (Hays, Morales & Reise, 2000)، وهذه الميزة مهمة بشكل خاص عند استخدام الاختبار أكثر من مرة (Embretson & Reise, 2000). ونظرا لأن استجابة المفحوص للفقرة تتضمن أخطاء، فإن هذه النظرية تهدف للتوصل إلى قيم تقديرية لكل من البرامترات، ومن ثم استخدام هذه القيم في تقدير احتمال الاستجابة الصحيحة لكل فقرة من فقرات الاختبار.

تتخذ الدالة المميزة لخصائص الفقرة شكل المنحنى اللوغاريتمي التراكمي logistic ogive curve، الذي يطلق عليه «منحنى خصائص الفقرة» ويمثل هذا المنحنى الانحدار غير الخطي للدرجات التي يحصل عليها المفحوصون في فقرة من فقرات الاختبار على القيم التقديرية للسمة أو القدرة المقيسة. ومن منحنى خصائص الفقرة تتضح كيفية عمل الفقرات، كما تتضح قدرة المفحوص المناظرة لاحتمال معين للأداء على فقرة ما (Shultz & Whitney, 2005; Wiberg, 2004).

تزود نظرية استجابة الفقرة بدالة معلومات لكل من الفقرات وللاختبار كاملا. ودالة المعلومات دليل يشير إلى مدى من مستوى القدرة التي يكون فيها الاختبار أو الفقرة أكثر فائدة للتمييز بين الأفراد (Shultz & Whitney, 2005)؛ وتشير المعلومات الأعلى إلى دقة أكبر. ومجموع دوال معلومات الفقرات عند مستوى معين من القدرة هو دالة معلومات الاختبار. وتؤدي دالة معلومات الاختبار دورا مهما في هذه النظرية، إذ يمكن من خلالها تحديد الخطأ المعياري في التقدير. وهذه الدالة مستقلة عن عينة المفحوصين، الأمر الذي يمكننا من وصف دقة الاختبار كأداة لتقدير قدرة المفحوص عبر مقياس السمة الكامنة، على العكس من النظرية الكلاسيكية التي تركز على الثبات بدلالة العلامة الحقيقية للمفحوص التي تنتج من قياسات متكررة (Doran, 2005). وعليه، فإن بناء اختبار اعتمادا على معايرة الفقرات باستخدام نماذج نظرية الاستجابة للمفردة أكثر فاعلية من مجرد الاعتماد على إحصائيات الفقرة التقليدية (Yen, 1983).

اقترح ثايسن وأورلاندو (Thissen & Orlando, 2001) منحيين لبناء النموذج في نظرية الاستجابة للمفردة. الأول، يتضمن الحصول على خصائص محددة للمقياس يعرفها النموذج الذي يجب أن يطابق البيانات. وإذا لم تناسب فقرة أو مفحوصًا خصائص نموذج الاستجابة للمفردة، يتم استبعاد المفحوص أو الفقرة؛ وهذا المنحى يستخدم نماذج راش. والمنحى الآخر، تطوير النموذج الأنسب well-fitting model لبيانات استجابات الفقرات. والهدف من هذا المنحى هو تحليل الفقرات؛ أي تقدير برامتراتها وتقييم جودة مطابقة درجات الفقرات للنموذج المعين، أي تحديد ما إذا كانت القيم المقدرة لبارامترات النموذج تحقق الافتراضات التي يرتكن إليها النموذج؛ فإذا تحققت هذه الافتراضات يمكن القول عندها إن النموذج يفسر أو يصف ما يحدث من تفاعل بين المفحوصين وفقرات الاختبار.

لقد تم إجراء العديد من الدراسات التي استخدمت فيها النماذج المختلفة لنظرية الاستجابة للمفردة وفق منحيي بناء النموذج لأغراض متعددة. منها: استقصاء درجة مطابقة البيانات الناتجة عن اختبارات عالمية أو وطنية لنماذج الاستجابة للمفردة (Kingstone, Leary & Wightman, 1985; Meijer & Sijtsma, 2008)، أو تفسير العلامات الناتجة عن الاختبارات محكية المرجع (Burket, 1984; Stone, Weissman & Lane, 2005)؛ أو دراسة أثر عدم مطابقة البيانات للنموذج في عدم اختلاف برامترات الفقرة أو القدرة (Fan & Ping, 1999)؛ أو تطوير مقاييس الأداء العادي (Probst, 2003; Reid, Kolakowsky-Hayner, Lewis & Armstrong, 2007)؛ أو استقصاء الخصائص السيكومترية لمقاييس

الأداء العادي (Zagorsek, Stough & Jaklic, 2006). إلا أن هناك ندرة في الدراسات العربية المواكبة لتطورات هذه النظرية، وبخاصة في مجال بناء اختبارات التحصيل. فهناك دراسات اعتمدت منحى الحصول على خصائص محددة للمقياس يعرفها النموذج الذي يجب أن يطابق البيانات، واستخدم فيها نموذج راش. ومن بين هذه الدراسات: دراسة كاظم (١٩٨٨) التي هدفت إلى بناء اختبار تحصيلي في علم النفس. ودراسة علام (١٩٩٥) التي هدفت إلى بناء اختبار تشخيصي محكي المرجع للمعارف الأساسية في إعداد خطة البحوث التربوية والنفسية. ودراستا دعنا حيث هدفت الأولى التي أجريت عام (٢٠٠٢) إلى بناء اختبار مفصّل متعدد المستويات للمفاهيم الرياضية الأساسية لطلبة الصفوف الأساسية (الرابع إلى التاسع) في الأردن وفق نموذج راش؛ في حين هدفت الدراسة الثانية التي أجريت عام (٢٠٠٥) إلى بناء اختبار مفصل في الرياضيات للصف الثامن. ودراسة الشريفين (٢٠٠٦) التي هدفت إلى تقدير الخصائص السيكومترية لاختبار محكي المرجع في القياس. ودراسة الفرجات (٢٠٠٤) التي هدفت إلى بناء بنك أسئلة في مبحث الكيمياء للصف الثاني الثانوي العلمي وفق النظرية الكلاسيكية ونموذج راش.

وهناك دراسات أخرى أجريت في الأردن استخدمت المنحى الآخر، ومن بينها دراسة حرز الله (٢٠٠٤) التي هدفت إلى بناء بنك أسئلة في الرياضيات، والتحقق من فاعليته في انتقاء فقرات اختبار محكي المرجع في مستوى امتحان شهادة الدراسة الثانوية العامة في الأردن. ودراسة مهيدات (٢٠٠٥) التي هدفت إلى بناء بنك أسئلة للمهارات الرياضية في نهاية المرحلة الأساسية والتحقق من فاعلية الاختبارات التي يمكن أن تسحب منه. ودراسة عثمان (٢٠٠٦) التي هدفت إلى بناء بنك أسئلة في الرياضيات للصف الثاني الثانوي العلمي باستخدام نظرية الاستجابة للفقرة. ودراسة النجار (٢٠٠٦) التي هدفت إلى بناء بنك أسئلة في مهارات الحاسوب للمرحلة الثانوية في الأردن باستخدام نماذج نظرية استجابة الفقرة بمعلمة ومعلمتين. ودراسة العطيوي (٢٠٠٦) التي هدفت إلى تطوير بنك فقرات في العلوم العامة باستخدام أساليب المعادلة الأفقية المستندة إلى النظرية الحديثة في القياس.

والمتفحص للدراسات الأجنبية والمتعلقة باختبارات التحصيل، يجد أنها إما استخدمت بيانات جاهزة من الاستجابات على اختبارات وطنية أو عالمية معدة سابقاً، أو استخدمت بيانات مولدة أو محاكاة data simulation. في حين يجد أن الدراسات التي عرضت أعلاه وأجريت في الأردن قد استخدمت بيانات حقيقية؛ غير أن معظم دراسات المنحى الثاني في بناء النموذج كانت رسائل ماجستير أو أطروحات دكتوراه، ومعظمها هدف إلى بناء بنوك

اسئلة، على الرغم من أن بنوك الأسئلة ما زالت غير مستخدمة في المدارس الأردنية، فضلاً عن عدم شيوع استخدام نظرية الاستجابة للمفردة، ذلك أنه بالكاد يتم استخدام النظرية الكلاسيكية في بناء الاختبارات المدرسية.

وعلى الرغم من أن المؤسسات المسؤولة عن تطوير الاختبارات لا بد أن تعمل على اختيار وتطوير الفقرات المناسبة لغرض الاختبار الذي يحتوي هذه الفقرات، وذلك من خلال اتباع النظريات الحديثة في القياس (Zenisky, Hambleton & Sireci, 2003)، فإن المتابع لما تحويه المؤسسات التربوية الأردنية، يلاحظ خلوها من اختبارات تحصيلية مبنية لغرض معين وفق نظرية الاستجابة للمفردة، على الرغم من توافر البرمجيات الإحصائية المتعلقة بالقياس وفق هذه النظرية. وفي سعيها لتحسين وتطوير العملية التربوية، أنشأت وزارة التربية والتعليم في الأردن مديرية للامتحانات والاختبارات من أهدافها تطوير اختبارات لضبط جودة التعليم (وزارة التربية والتعليم، ٢٠٠٥). وفي العام الدراسي ٢٠٠٦/٢٠٠٧، أجرت هذه المديرية اختبارا وطنياً لضبط جودة تعليم طلبة الصف العاشر في أربعة مباحث، هي اللغة العربية، واللغة الإنجليزية، والرياضيات والعلوم. وقد طبق هذا الاختبار على حوالي ٤٢٠٠٠ طالب وطالبة من الصف العاشر. غير أن نتيجة تحليل البيانات الناتجة عن الاستجابات لفقرات هذه الاختبارات التي أجراها يعقوب وآخرون عام ٢٠٠٨، تظهر أن هناك خللا في عملية بناء هذه الاختبارات وفي صدقها، الأمر الذي يضع علامة استفهام حول مناسبتها للأغراض التي أعدت من أجلها.

هدف الدراسة

يتبين مما سبق أن الدراسات السابقة الأجنبية المتعلقة بنظرية الاستجابة للمفردة استخدمت اختبارات عالمية –تتضمن عددا كبيرا من الفقرات–، أو استخدمت بيانات مولدة أو محاكاة. في حين أن الدراسات التي أجريت في الأردن واعتمدت منحى تطوير النموذج الأنسب لبيانات استجابات الفقرة، استخدمت بيانات حقيقية، وهدفت في معظمها إلى بناء بنوك أسئلة؛ وكأن هناك افتراضاً ضمنياً بأن إجراءات بناء وتطوير اختبار تحصيلي وفق نظرية الاستجابة للمفردة واضحة لدى المعلمين وواضعي الاختبارات، على الرغم من أن العديد من واضعي الاختبارات لا يمتلكون مهارات بناء وتطوير اختبار صادق وثابت، ولا يستطيعون استخراج برامترات فقراته وفق النظرية الكلاسيكية في القياس، على الرغم من توافر البرامج الإحصائية. وتزداد المشكلة حدة لدى الطلب إليهم استخدام نظرية الاستجابة

للمفردة في بناء وتطوير الاختبارات المدرسية. وبالإضافة إلى ما سبق، تكاد تخلو المؤسسات التربوية الأردنية من اختبارات تحصيلية مبنية وفق نظرية الاستجابة للمفردة، فما زالت عملية البناء والتطوير وفق هذه النظرية في مراحلها الأولى. ومن المعلوم أن عدد فقرات الاختبارات المدرسية – سواء اليومية أو الشهرية أو النهائية– قليل. فهل بالإمكان استخدام هذه النظرية في بناء اختبار تحصيلي مدرسي من إعداد المعلم؟

وفي ضوء ما سبق، ونظراً لدور الاختبارات التحصيلية المدرسية في القرارات المتعلقة بمسيرة الطالب العلمية ومستقبله، وفي محاولة لوضع النظرية موضع التطبيق، جاءت هذه الدراسة لتستقصي إمكان استخدام نظرية الاستجابة للمفردة في بناء الاختبارات التحصيلية التي يعدها المعلمون–ذات العدد القليل من الفقرات. هذا بالإضافة إلى توضيح طريقة بناء الاختبار وتطويره وفق هذه النظرية.

أسئلة الدراسة

إذا أخذ بعين الاعتبار أنه تم بناء الاختبار وتطبيقه على عينة من الطلبة، فإن تحليل نتائج الاستجابة للفقرات يستهدف الإجابة عن الأسئلة الآتية:

١– ما أنسب نموذج (الأحادي أو الثنائي أو الثلاثي البراميترات) من نماذج استجابة الفقرة ثنائية التصحيح (العلامة على الفقرة إما (٠) أو (١)، ولا يوجد علامة جزئية) لنتائج الاستجابة الفقرات؟

٢– ما درجة تحقيق نتائج استجابة الفقرات لافتراضات هذا النموذج؟

٣– ما درجة تحقق افتراضي عدم اختلاف تقديرات القدرة وبرامترات الفقرة لهذا النموذج؟

٤– ما مدى تحقق معايير الدقة في تقدير قدرة المفحوص ممثلة بالخطأ المعياري في التقدير ودالة المعلومات للفقرات وللاختبار كاملاً؟

منهجية الدراسة وإجراءاتها:

لتحديد محتوى الاختبار الذي سيتم استخدامه وسيلة لاستقصاء إمكان استخدام نماذج الاستجابة للمفردة في بناء اختبار تحصيلي من إعداد المعلم، ولتوضيح كيفية بناء الاختبار وفق هذه النظرية، فقد تم اختيار موضوع العمليات الأربع على الكسور العادية. وقد أظهرت العديد من الدراسات (أبو لبدة، ٢٠٠٣؛ الرفيع وآخرون، ٢٠٠٧) وجود ضعف لدى

الطلبة في هذا الموضوع. ولأن منهاج الصف الخامس يغطي موضوع العمليات الأربع على الكسور العادية، فقد تم التخطيط لتطبيق الاختبار على طلبة الصف الخامس.

عينة الدراسة

تكون أفراد الدراسة من طلبة من الصف الخامس. وتكونت العينة الرئيسة من (٢٨٤) طالباً وطالبة، في ثماني شعب تم اختيارها عشوائيا، من شعب الصف الخامس في مديرية تربية وتعليم الزرقاء الأولى. وهناك أربع عينات تجريبية غير العينة الرئيسة استخدمت لأغراض معينة أثناء بناء الاختبار وتطويره، وردت في الخطوات ذوات الأرقام ٨، ٩، ١٣، ١٤ من بند «أداة الدراسة».

أداة الدراسة

تكونت أداة الدراسة في صورتها النهائية التي تم استخدام نماذج الاستجابة للمفردة عليها من اختبار تحصيلي مكون من ٢٣ فقرة من نمط الاختيار من أربعة بدائل، يقيس العمليات الأربع على الكسور العادية. وحيث إنّ إجراءات بناء الاختبار هي نفسها وفق النظريتين التقليدية والحديثة (Green, Yen & Burket, 1989; Hambleton & Jones, 1993)، فقد تم بناء الاختبار وفق الخطوات الآتية:

١- تكوين فريق عمل مكون من ثلاثة من معلمي الرياضيات للصف الخامس، ومشرفين اثنين لمادة الرياضيات، وعضوي هيئة تدريس أحدهما يحمل الدكتوراه في أساليب تدريس الرياضيات، والآخر في القياس والتقويم. وقد تضمنت مهام هذا الفريق الحكم على:

أ- درجة مطابقة النتاجات التعلمية للمحتوى المستهدف.

ب- درجة تمثيل جدول المواصفات لمجتمع السلوك المستهدف.

ت- مناسبة كل من الفقرات التي عددها ٤٦ للنتاج التعلمي المستهدف.

ث- أنسب فقرة من الفقرتين على كل هدف، في ضوء تمييزها وفاعلية بدائلها.

ج- صدق المحتوى المتعلق بتمثيل الاختبار الذي يحقق افتراضات نظرية الاستجابة للمفردة لمجتمع السلوك المستهدف.

٢- تحديد مجال الاختبار، وقد تطلب ذلك تحليل محتوى موضوع العمليات الأربع على الكسور العادية وتحديد النتاجات التعلمية المتوقع من الطلبة بلوغها.

٣- عرض النتاجات التعلمية على فريق العمل للحكم على درجة مطابقة النتاجات التعلمية

للمحتوى المستهدف ،وقد تم إجراء التعديلات عليها في ضوء التغذية الراجعة الناتجة من فريق العمل.

٤– إعداد جدول المواصفات، وتحديد النتاجات التعلمية التي سيتم قياسها.

٥– عرض الجدول على فريق العمل للحكم على درجة تمثيل الجدول لمجتمع السلوك المستهدف، وقد تم إجراء التعديلات في ضوء التغذية الراجعة.

٦– الخروج بالنتاجات التعلمية التي اتفق المحكمون عليها في الخطوة السابقة، وقد تضمنت ثلاثة وعشرين نتاجا تعلميا.

٧– بناء فقرتين من نمط الاختيار من أربعة بدائل على كل نتاج تعلمي–وبمعنى آخر، بناء (٤٦) فقرة– وعرضها على فريق العمل للحكم على مناسبة كل منها للنتاج التعلمي المستهدف، وقد تم تعديل الفقرات التي تحتاج إلى تعديل وفق التغذية الراجعة من فريق العمل، سواء جذر الفقرة أو بدائلها. كما تم الاتفاق على إعطاء الوزن نفسه للفقرات، ومن ثمّ تحصل الفقرة على العلامة (١) إذا كانت الإجابة عنها صحيحة والعلامة (٠) إذا كانت خطأ.

٨– إعداد اختبار مكون من الفقرات التي أقرها فريق العمل–في الخطوة السابقة رقم ٧– وعددها (٤٦) فقرة، وتطبيقه على عينة تجريبية خارج العينة الرئيسة مكونة من شعبتين تضمان ٤٥ طالباً وطالبة، وذلك لمعرفة درجة وضوح الفقرات ووضوح التعليمات وإجراء تحليل للفقرات وفق النظرية الكلاسيكية. وقد تم تعديل بعض بدائل الفقرات بناءً على التغذية الراجعة من عملية التطبيق هذه.

٩– إعادة تطبيق الاختبار المعدل على عينة تجريبية أخرى خارج العينة الرئيسة مكونة من شعبتين تضمان ٧١ طالباً وطالبة، لإعطاء فكرة عن صعوبة الفقرات وتمييزها والتحقق من فاعلية بدائلها.

١٠– حساب صعوبة الفقرات وتمييزها وفاعلية بدائلها وفق النظرية الكلاسيكية، وقد تراوحت الصعوبة بين ٠,٢٦ و ٠,٩٣؛ في حين تراوح تمييز الفقرات بين ٠,٣٢ و ٠,٨٧؛ كما تبين أن جميع البدائل فاعلة، فقد اجتذبت من الفئة الدنيا أكثر مما اجتذبت من الفئة العليا. وعليه، يمكن اعتبار أن فقرات الاختبار فاعلة.

١١– لأن اختبارات التحصيل المدرسية، في معظمها، معيارية المرجع؛ فإن من الطبيعي أن يتم قياس الهدف المعين بفقرة اختبارية واحدة. وكي يكون الاختبار مماثلا لما يتم في المدرسة، فقد طلب إلى فريق العمل اختيار أنسب فقرة من الفقرتين على كل هدف، في ضوء تمييزها وفاعلية بدائلها، وقد كانت الفروق في التمييز بين كل زوجين من أزواج الفقرات المتناظرة

–اللتين تقيسان الهدف نفسه– لا تزيد عن ٠,٠٤؛ كما كانت الفروق في نسبة اختيار البديل لا تزيد عن ٠,٠٢ لكل بدائل أزواج الفقرات المتناظرة. وقد اختار فريق العمل الفقرات ذات التمييز الأعلى. ومن ثم نوقشت الاختيارات وتم الاتفاق عليها، جميعاً، باستثناء فقرة تقيس مفهوم ضرب الكسور فقد تم اختيار الفقرة المناظرة من زوجي الفقرات الذي يقيس الهدف نفسه التي كان تمييزها يقل عن تمييز الفقرة النظير بمقدار ٠,٠٣ ولكن كانت جميع بدائلها أكثر فاعلية من بدائل الفقرة المناظرة.

١٢– إعداد اختبار مكون من الفقرات التي تم الاتفاق عليها من فريق العمل وعددها ثلاث وعشرون فقرة تقيس الأهداف المخططة.

١٣– اختيار شعبتين من خارج العينة الرئيسة، تضمنتا (٦٩) طالب وطالبة وطبق عليهما الاختبار بهدف تحديد الزمن اللازم، وقد كان متوسط الزمن الذي استغرقه الطلبة في الإجابة عن الاختبار ثلاثين دقيقة، وعليه، فقد اعتمد زمن الاختبار خمس وثلاثون دقيقة (بحيث يتم أخذ زمن توزيع الأوراق وجمعها بعين الاعتبار).

١٤– اختيار ست شعب من خارج العينة الرئيسة، تضمنت (٢٠٠) طالبا وطالبة، بهدف تجريب الاختبار عليهم وتحليل فقراته، وحساب ثباته، ووجد أن الفقرات واضحة وتراوحت صعوبتها بين ٠,٣١ و ٠,٨٨، في حين تراوح تمييز الفقرات بين ٠,٣٨، ٠,٨٤. ويمكن أن يعزى الفرق بين النتيجتين الواردتين في هذه الخطوة مع الخطوة رقم ١٠ بسبب أن العينتين غير عشوائيتين–وحدة الاختيار كانت الشعبة وليس الفرد؛ فضلاً عن أن حجم العينة في الخطوة ١٤ أكبر من حجمها في الخطوة ١٠. وحسب ثبات الاختبار باستخدام معادلة كرونباخ ألفا وكان مساوياً ٠,٨٦ .

١٥– إعادة تطبيق الاختبار على العينة نفسها، الواردة في الخطوة السابقة رقم ١٤، بعد أسبوعين، وحسب ثبات استقرار السمة عبر الزمن، وكانت قيمة معامل ارتباط بيرسون ٠,٩٢، الأمر الذي يشير إلى أن هذه الأداة مناسبة لأغراض هذه الدراسة.

إجراءات التنفيذ

١– طبق الاختبار على العينة الرئيسة، وفق الزمن المحدد، وصححت فقراته تمهيدا لاستخدامها في تطويره وفق نظرية الاستجابة للمفردة.

٢– وبهدف الإجابة عن سؤال الدراسة الأول المتعلق بتحديد أنسب نموذج من نماذج استجابة الفقرة ثنائية التصحيح (الأحادي أو الثنائي أو الثلاثي البراميترات) لنتائج الاستجابة

للفقرات، استخدم اختبار مطابقة النموذج.

٣- وللإجابة عن سؤال الدراسة الثاني المتعلق بدرجة تحقيق نتائج استجابة الفقرات لافتراضات النموذج ،تم استخدام التحليل العاملي بنوعيه الاستكشافي والتوكيدي لاختبار أحادية البعد؛ واختبار Q3 لفحص الاستقلال الموضعي.

٤- وللإجابة عن سؤال الدراسة الثالث الذي يهدف إلى التحقق من افتراضي عدم اختلاف كل من: تقديرات القدرة وبرامترات الفقرة لهذا النموذج، تم حساب الدليل الآتي (the negative of 2 loglikelihood (-2LLH).

٥- وبهدف الإجابة عن السؤال الرابع المتعلق بالتحقق من معايير الدقة في تقدير قدرة المفحوص ممثلة بالخطأ المعياري في التقدير ودالة المعلومات للفقرات وللاختبار كاملاً، تم إيجاد دالة معلومات الاختبار والخطأ المعياري في التقدير.

نتائج الدراسة

هدفت هذه الدراسة إلى استقصاء إمكان استخدام نماذج نظرية الاستجابة للمفردة في تطوير اختبار تحصيلي مدرسي من إعداد المعلم -يتكون من عدد قليل من الفقرات-. وبعد بناء الاختبار وتطبيقه على أفراد الدراسة، تم تحليل استجابات الطلبة على الفقرات الناتجة من التطبيق باستخدام نماذج نظرية الاستجابة للمفردة بهدف الإجابة عن أسئلة الدراسة.

أولاً: نتائج السؤال الأول

نص السؤال الأول على ما يأتي: «ما أنسب نموذج (الأحادي أو الثنائي أو الثلاثي البراميترات) من نماذج استجابة الفقرة ثنائية التصحيح لنتائج الاستجابة للفقرات؟».

وقد تناولت الإجابة عن هذا السؤال ما يأتي:

١- تحديد النموذج الأنسب.

٢- تفحص جودة مطابقة الفقرات للنموذج الأنسب.

وفيما يأتي توضيح لكيفية تناول كل منهما:

١. تحديد النموذج الأنسب

وقد تم تحديد النموذج الأنسب من خلال اختبار مطابقة النموذج model fit. ولأن الاختبار الذي أعد في هذه الدراسة من نمط الاختيار من أربع بدائل، فمن المتوقع أن يتضمن النموذج براميتر التخمين، أي من المتوقع أن يكون النموذج الأنسب هو الثلاثي البراميتر. إلا

أن تضمين النموذج برامترات إضافية سيقلل من دقة التقدير لأنه سيتطلب تقدير برامترات ليست بحاجة إلى التقدير (Jiao & Lau, 2003). ولأن النموذج الأحادي (ويسمى النموذج الصفري null model) مضمن nested في النموذج الثنائي كونه يتضمن برامتراً إضافياً عن النموذج الأحادي (تمييز الفقرة)، والنموذج الثنائي مضمن في النموذج الثلاثي (ويسمى النموذج البديل model alternative) لأنه يتضمن برامتراً إضافيا عن النموذج الثنائي (التخمين) (Zimowski, Muraki, Mislevy & Bock, 2003)، فقد تم اختبار الفرضيتين الصفريتين الآتيتين:

الفرضية الأولى، وتنص على أن تضمين النموذج الأحادي برامترات إضافية لا يؤدي إلى تحسن دال إحصائياً عند مستوى دلالة (α=٠,٠٥) في مطابقة النموذج الأحادي للاستجابات عن الفقرات، بمعنى أن تضمينها في النموذج الثنائي لا يؤدي إلى تحسن دال إحصائياً على مستوى دلالة (α=٠,٠٥) في مطابقة النموذج الأحادي البراميتر للاستجابات عن الفقرات.

والفرضية الثانية، وتنص على أن تضمين النموذج الثنائي برامترات إضافية لا يؤدي إلى تحسن دال إحصائياً على مستوى دلالة (α=٠,٠٥) في مطابقة هذا النموذج للاستجابات عن الفقرات.

ولاختبار كل من الفرضيتين، فقد تم استخدام برمجية BILOG-MG وحساب الإحصائي التالي ((the negative of 2 loglikelihood (-2LLH) لكل من النماذج الثلاثة، ولنفس استجابات الفقرات؛ ومن ثم، وتم إيجاد الفرق بين قيمتي (-2LLH) لكل من أزواج النماذج الثلاثي والثنائي، والثنائي والأحادي. وولأن الفرق الموجب بين هذه القيم له توزيع كاي تربيع $\chi 2$ بدرجات حرية تساوي الفرق بين عدد برامترات النموذجين (Zimowski et al., 2003)، فقد تم حساب القيمة الناتجة من الفرق بين قيمتي (-2LLH) للنموذجين الأحادي والثنائي البراميتر، وقد كانت تساوي ١٠٢,٧٧. كما تم حساب الفرق بين قيمتي (-2LLH) للنموذجين الثنائي والثلاثي البراميتر، وكانت تساوي (١٨,٦٢٨). ومن ثم تم مقارنة كل من القيمتين مع قيمة $\chi 2$ الحرجة بثلاث وعشرين درجة حرية ومستوى دلالة ٠,٠٥ والتي تساوي (٣٥,١٧)، ووفق هذه النتائج فقد تم رفض الفرضية الصفرية الأولى، وقبول الفرضية الصفرية الثانية. أي أن تضمين النموذج الأحادي برامترات إضافية يؤدي إلى تحسن دال إحصائياً (α=٠,٠٥) في مطابقة النموذج للاستجابات عن الفقرات؛ في حين أن تضمين النموذج الثنائي برامترات إضافية لم يؤد إلى تحسن دال في مطابقة النموذج للاستجابات عن الفقرات. الأمر الذي يشير إلى أن النموذج ثنائي البراميتر هو النموذج الأنسب للاستجابات على الفقرات.

٢. تفحص جودة مطابقة الفقرات للنموذج الأنسب.

وبعد تحديد أنسب نموذج، فإن هناك حاجة لتفحص جودة مطابقة كل من الفقرات للنموذج ثنائي البراميتر. وقد تم استخدم اختبار احتمالية كاي تربيع likelihood-ratio chi-squared test لاختبار الاختلاف بين الأنماط المتوقعة لاستجابة الفقرة وأنماط الاستجابة الفعلية لها. وتبين أن هنالك أربع فقرات لم تطابق النموذج حيث كانت قيمة كاي تربيع ($\chi 2$) دالة عند مستوى ٠,٠٥، وعليه، فقد تم حذفها، وتبقى تسع عشرة فقرة.

ثانياً: نتائج السؤال الثاني

نص السؤال على: ما درجة تحقيق نتائج استجابة الفقرات لافتراضات هذا النموذج؟

وللإجابة عن السؤال الثاني المتعلق باستقصاء درجة تحقيق نتائج استجابة الفقرات لافتراضات النموذج الأنسب. فقد تم تفحص الافتراضات الرئيسة الآتية (Hambleton & Swaminathan, 1985):

أ) أحادية البعد

ب) الاستقلال الموضعي

ج) الدالة المميزة لكل فقرة والتي تصف العلاقة الوتيرية بين القدرة والأداء على الفقرة، وتسمى "دالة خصائص الفقرة".

د) التوزيع الطبيعي المعياري للقدرة

وفيما يأتي توضيح لكيفية التحقق من هذه الافتراضات:

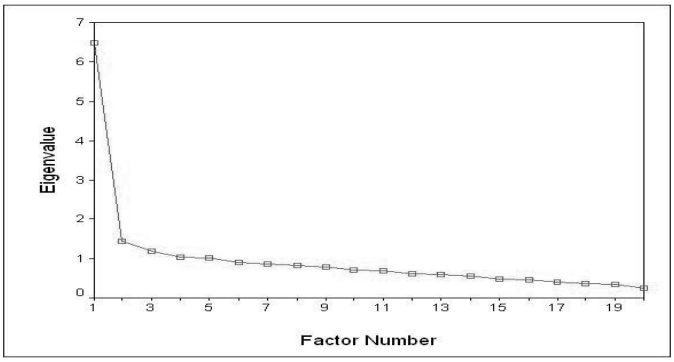
أ) أحادية البعد. فإذا كان الاختبار أحادي البعد، فإن العلامات تتنبأ بالأداء على الفقرات؛ ومن ثمّ يكون من المنطقي تلخيص الأداء على الاختبار بعلامة منفردة (Kane, 2008). أما إذا لو لم يكن الاختبار أحادي البعد، فهناك معلومات إضافية تتعلق بهذا الاختبار الفرعي لم يتم عكسها في العلامة الكلية، ومن ثمّ، فعلامته لا تتنبأ بالأداء على الفقرات أو مجموعات فرعية منها (Schilling, 2007).

وعلى الرغم من أن الأبحاث أشارت إلى أن نماذج نظرية الاستجابة للمفردة متينة نسبياً لانتهاك فرضية أحادية البعد (Lau, 1996; Spray, Abdel-fattah, Huamg & Lau, 1997)، فقد تم اختبار هذا الافتراض بطريقتين: الأولى، باستخدام التحليل العاملي الاستكشافي، ويبين الجدول رقم (١) أن هناك خمسة عوامل الجذر الكامن لكل منها أكبر من ١ ويفسرون معاً ٥٥,٧٥٪ من التباين؛ وأن العامل الأول له جذر كامن قيمته ٦,٤٨ وتفسر ٣٢,٤٠٪ من التباين وهذه القيمة أكبر من ٢٠٪ من ثمّ، يمكن اعتباره عاملاً سائداً

(Hambleton, 2004). وفضلاً عن ذلك، فإن هناك طريقة أخرى لتحديد عدد العوامل من خلال تمثيل منحنى قيم الجذور الكامنة eigenvalues المتتالية بيانياً الذي يظهر في الشكل (١)؛ وباستخدام هذا الشكل فإن القرار حول عدد العوامل وفق لوهلن (Loehlin, 1987) يتم التوصل إليه اعتماداً على النقطة التي يتغير فيها ميل منحنى الجذور الكامنة بسرعة من منحنى يعامد–تقريبا–محور السينات إلى منحنى أفقي–تقريباً–، وهذا يحصل عند العامل الأول. بمعنى أنه يمكن افتراض تحقق افتراض أحادية البعد.

الجدول رقم (١)
الجذور الكامنة لعوامل الاختبار

رقم العامل	الجذر الكامن	نسبة التباين المفسر	نسبة التباين المفسر التجميعي
١	٦,٤٨	٣٢,٤٠	٣٢,٤٠
٢	١,٤٤	٧,١٩	٣٩,٥٩
٣	١,١٨	٥,٩١	٤٥,٥٠
٤	١,٠٤	٥,١٩	٥٠,٦٩
٥	١,٠١	٥,٠٦	٥٥,٧٥
٦	٠,٩٠	٤,٥١	٦٠,٢٦



الشكل رقم (١)
الجذور الكامنة لعوامل الاختبار

وفي محاولة لإعطاء دليل آخر على تحقق افتراض أحادية البعد، فقد تم استخدام التحليل العاملي التوكيدي الذي يختبر الفرضية القائلة بأن النموذج أحادي البعد متسق مع البيانات الناتجة عن الاستجابات لفقرات الاختبار، مقابل الفرضية التي تنص على النموذج ثنائي البعد متسق مع البيانات. وقد استخدمت برمجية ليزرل LISREL، وكانت نتائج الفرضية الأولى كما يأتي: قيمة كاي تربيع ($\chi 2$) تساوي ٦٥,١٦ وهي دالة عند مستوى دلالة ٠,٠٦٢،

بمعنى قبول الفرضية الصفرية الأولى. ودليل جودة المطابقة LISREL goodness-of-fit يساوي ٠,٩٣، ودليل مطابقة المقارنة ((the Comparative Fit Index (CFI يساوي ٠,٩٨٠، وكلاهما أكبر من ٠,٩ (الحد الأدنى المقبول)، كما أن الأخير قريب من الواحد الصحيح، وهذا دليل على مطابقة جيدة (Jöreskog & Sörbom, 1993b)؛ وفضلاً عن ذلك، فقد كانت قيمة الجذر التربيعي لمتوسط مربعات أخطاء التقدير (Root Mean Square Error of Approximation (RMSEA) تساوي ٠,٠٤٤ وهي أقل من ٠,٠٥؛ وهذه النتائج حميعها تشير إلى أن النموذج أحادي البعد متسق مع البيانات الناتجة عن الاستجابات لفقرات الاختبار (Jöreskog & Sörbom, 1993a). في حين كانت نتائج الفرضية الثانية لا تدل على مطابقة جيدة، فقد كانت قيمة كاي تربيع ($\chi 2$) تساوي ٨٥,٢٣ وهي دالة عند مستوى دلالة ٠,٠٠، بمعنى رفض الفرضية الصفرية الثانية التي تنص على النموذج ثنائي البعد متسق مع البيانات. وكان دليل جودة المطابقة LISREL goodness-of-fit يساوي ٠,٨٧، ودليل مطابقة المقارنة ((the Comparative Fit Index (CFI يساوي ٠,٩٠، والأول أقل من ٠,٩؛ وفضلاً عن ذلك، فقد كانت قيمة الجذر التربيعي لمتوسط مربعات أخطاء التقدير ((Root Mean Square Error of Approximation (RMSEA تساوي ٠,٥٢؛ أي لا تعبر عن مطابقة جيدة (Jöreskog & Sörbom, 1993a). وهذه الأدلة، مجتمعة، مؤشر على أن الاختبار يقيس بعداً واحداً (Jöreskog & Sörbom, 1993a; 1993b). وفي ضوء ما سبق، يمكن القول إنّ افتراض أحادية البعد قد تحقق.

ب) الاستقلال الموضعي، ويعني أنه باستثناء القدرة المستهدفة، لا يوجد علاقة بين استجابات فقرات الاختبار غير العلاقة المحددة بالقدرة أو برامترات محددة أخرى للنموذج؛ بمعنى أن الاستجابة على إحدى الفقرات لا تفسر أو تساعد في الإجابة عن أسئلة أخرى. وإذا ما تم انتهاك هذا الافتراض، فسيظهر ما يسمى بالارتباط الموضعي بين الفقرات local item dependence، ووجود مثل هذا الارتباط قد يؤدي إلى تقديرات غير دقيقة لبرامترات الفقرات وإحصائيات الاختبار وقدرة المفحوصين، وفضلاً عن ذلك، فإن هذا الارتباط يضيف بعداً إضافياً (وهو غالباً غير مقصود) للاختبار ويكون على حساب البناء الذي نهتم به (Zenisky, Hambleton & Sireci, 2006).

وقد تم التحقق من هذا الافتراض، من خلال إجراءات بناء الاختبار، وذلك من خلال العمل على عدم وجود فقرة تعطي ملمحاً عن إجابة أية فقرة أخرى. وفضلاً عن ذلك، ولكون النموذج الأنسب يتمتع بأحادية البعد، فقد استخدم الإحصائي Q3 الذي اقترحته ين (Yen, 1984) كمؤشر للكشف عن الاستقلال الموضعي للفقرات. ويعبر عنه بالعلاقة بين

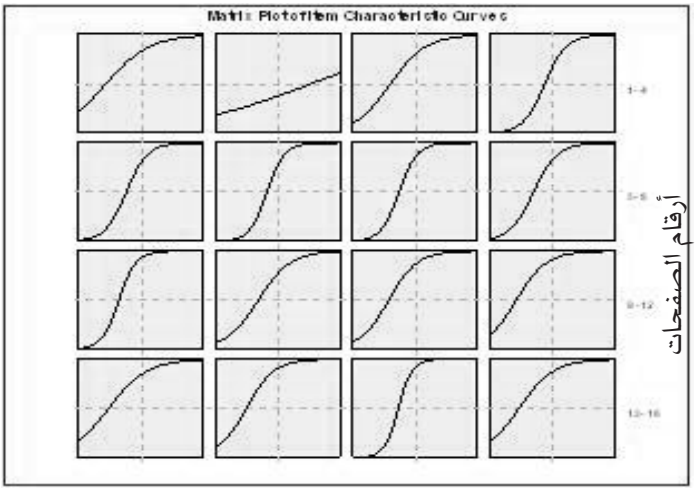
البواقي لزوج من الفقرات بعد ضبط السمة المقدرة. وقد تم حساب هذا الإحصائي، وكانت قيمته أقل من الصفر لجميع أزواج الفقرات باستثناء ثلاث فقرات كانت قيمة هذا الإحصائي لأي زوجين منها أكبر من الصفر. بمعنى أن هذه الفقرات غير مستقلة. وتم حذفها، وبذلك بقيت ست عشرة فقرة.

ج) الدالة المميزة لكل فقرة التي تصف العلاقة الوتيرية بين القدرة والأداء على الفقرة وتسمى «دالة خصائص الفقرة». ولاختبار هذا الافتراض، فلا بد من إيجاد برامترات الفقرات التي تحدد منحنى خصائص كل منها وفق النموذج الثنائي البراميتر الذي تبين أنه الأنسب. وعليه، فقد تم استخدام برمجية BILOG-MG لتقدير برامترات الفقرات الستة عشر (صعوبة الفقرات وتمييزها) وتقدير قدرة الأفراد. ويبين الجدول رقم (٢) معاملات صعوبة وتمييز هذه الفقرات وفق هذا النموذج، ومستوى دلالة قيم كاي تربيع للمطابقة. وبتأمل الجدول رقم (٢)، يتبين أن قيم معاملات صعوبة الفقرات تراوحت بين –١,٧٧ و ١,٠٣، بمتوسط يساوي –٠,٩٥ وانحراف معياري يساوي ٠,٦٨. وبالنسبة لمعاملات التمييز، فقد كانت جميعها موجبة، وتراوحت قيمها بين ٠,٤٨، ٣,٢٧ بمتوسط يساوي ١,٧١ وانحراف معياري يساوي ٠,٦٩. كما يتبين أيضاً، أن جميع الفقرات مطابقة للنموذج.

الجدول رقم (٢)
معاملات صعوبة وتمييز فقرات الاختبار ومستوى دلالة جودة المطابقة

رقم الفقرة	معامل الصعوبة (b_i)	معامل التمييز (a_i)	مستوى دلالة $\chi 2$	رقم الفقرة	معامل الصعوبة (b_i)	معامل التمييز (a_i)	مستوى دلالة $\chi 2$
١	–١,٦٥	١,٠٠	٠,٩٥	٩	–١,٠٠	٢,٢٢	٠,٩٤
٢	١,٠٣	٠,٤٨	٠,٨٤	١٠	–٠,٩٢	١,٤٠	٠,٥٥
٣	–١,١٤	١,٢٩	٠,٤٨	١١	–١,٢٨	١,٤٥	٠,٣٢
٤	–٠,٣٩	٢,١٤	٠,٠٥٢	١٢	–١,٧٧	١,٣٥	٠,٩٩
٥	–٠,٦٦	٢,٠٩	٠,٠٥٩	١٣	–١,٥١	١,١٢	٠,٩٦
٦	–٠,٥٠	٢,٥٦	٠,٤٧	١٤	–١,٤٨	١,٥٦	٠,٥٤
٧	–٠,٧٣	٢,٣١	٠,٧٥	١٥	–٠,٧٠	٣,٢٧	٠,٧٢
٨	–٠,٨٧	١,٨٦	٠,٢٦	١٦	–١,٥٥	١,٣٠	٠,٨٢

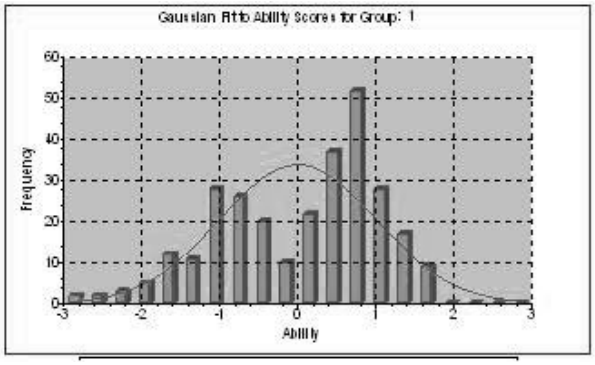
وبعد تحديد برامترات الفقرات التي طابقت النموذج، أمكن رسم منحنى خصائص الفقرة، كما يتبين في الشكل رقم (٢).



الشكل رقم (٢)
مصفوفة منحنيات خصائص فقرات الاختبار

وعليه، فقد تحقق الافتراض الثالث، بمعنى أن هناك دالة مميزة لكل فقرة تصف العلاقة الوتيرية بين القدرة والأداء على الفقرة لكل من الفقرات الستة عشر.

د) التوزيع الطبيعي المعياري للقدرة؛ ولاختبار هذا الافتراض تم استخدام اختبار كولموغروف–سميرنوف Kolmogorov-Smirnov test لاختبار الفرضية الصفرية التي تنص على أنه لا يوجد فرق دال إحصائياً (α=٠,٠٥) بين توزيع القدرة الناتج من الفقرات الستة عشر والتوزيع الطبيعي المعياري؛ حيث كانت قيمة (ز) المحسوبة (Kolmogorov-Smirnov Z) دالة عند مستوى دلالة ٠,٠٣٢، أي أننا نرفض الفرضية الصفرية؛ بمعنى أن توزيع القدرة ليس طبيعيا. ويبين الشكل رقم (٣) الآتي ذلك.



الشكل رقم (٣)
توزيع قدرة المفحوصين نسبة إلى التوزيع الطبيعي المعياري

وعليه، يمكن القول إن القدرة θ غير موزعة توزيعاً طبيعيا.

ويتبين مما سبق أن ثلاثة من الافتراضات الأربعة السابقة لم تتحقق للاختبار المكون من تسع عشرة فقرة، ففي حين نجد أن افتراض أحادية البعد قد تحقق، نجد أن الاستقلال الموضعي تحقق لست عشرة فقرة، فقط؛ في حين لم يتحقق التوزيع الطبيعي للقدرة المقدرة من الست عشرة فقرة. بمعنى أن نتائج استجابة الفقرات لم تحقق افتراضات النموذج ثنائي البراميتر الذي بيّن اختبار (LLH 2-) أنه الأنسب لهذه البيانات.

ثالثاً: نتائج السؤال الثالث

نص السؤال الثالث على ما يأتي: ما درجة تحقق افتراضي عدم اختلاف تقديرات القدرة و برامترات الفقرة لهذا النموذج؟

ويتعلق هذا السؤال بموضوعية المقارنة بين قدرتي مفحوصين استجابا لبند معين، أو المقارنة بين برامترات فقرتين استجاب لهما مفحوص معين (Bond & Fox, 2007). وقد تم التحقق من هذين الافتراضين كما يأتي:

أ) عدم اختلاف تقديرات القدرة؛ ولاختبار هذا الافتراض، استخدم أسلوب كستر وآخرين (Custer, et al., 2008). فقد تم استخدام تصميم المجموعة المشتركة common person design عبر مجموعتين من الفقرات: إحداهما تتكون من الفقرات السهلة والأخرى من الفقرات الصعبة. وقد تطلب هذا الأمر تصنيف الفقرات وفق صعوبتها وفق النموذج الثنائي البراميتر، إذ رتبت من الأصعب إلى الأسهل، ثم قسمت إلى جزأين؛ وقد شكلت الفقرات الثماني الأولى الاختبار الصعب، والفقرات الثماني الأخيرة الاختبار السهل. وعليه، تضمن كل اختبار ثماني فقرات. ثم حللت البيانات الناتجة عن استجابة أفراد الدراسة لكل من الاختبارين (الصعب والسهل)، وفق نماذج استجابة الفقرة الثلاثة (الأحادي والثنائي والثلاثي البراميتر)، ومن ثم تم استخدام اختبار ولككسون للمجموعات المترابطة (Wilcoxon's matched-pairs signed-ranks test) لاختبار أن توزيعي القدرة المقدرة من كل من النماذج الثلاثة عبر الاختبارين السهل والصعب مسحوبان من مجتمعين متطابقين. وكانت النتيجة كما في الجدول رقم (٣) التي تبين أن النموذج الأنسب الذي يحقق محك عدم اختلاف برامترات القدرة هو النموذج الأحادي. مما يشير إلى أن تقديرات القدرات لا تختلف باختلاف هل الفقرات الصعبة أو السهلة هي التي تم استخدامها إذا كان النموذج الأنسب هو الأحادي البراميتر.

الجدول رقم (٣)
نتائج اختبار ولككسون

النموذج	قيمة Z	مستوى الدلالة
النموذج الأحادي	-٧,٥٥	٠,٤٥
النموذج الثنائي	-١٨,٨٦٩	٠,٠٠٠
النموذج الثلاثي	-١٩,٠٢٩	٠,٠٠٠

وإذا أخذنا بالاعتبار أن النموذج الأنسب للبيانات الناتجة عن الاستجابات للفقرات كان النموذج ثنائي البراميتر، فإن هذا يعني أن الافتراض المتعلق بعدم اختلاف تقديرات القدرة لم يتحقق.

ب) عدم اختلاف تقديرات برامترات الفقرات، ويعني أن الفرق بين تقدير برامترات فقرتين يبقى ثابتاً مهما كان مستوى قدرة الفرد التي يتم من خلالها إيجاد هذا الفرق (التقي، ٢٠٠٩). وإحدى طرق تقويم عدم اختلاف برامترات الفقرة لمجموعة من الفقرات أحادية البعد هي مقارنة ترتيب رتب تقديرات صعوبة الفقرة عبر مجموعات مختلفة من المفحوصين الذين يسحبون من نفس المجتمع. وعندما يكون ترتيب تقديرات صعوبة الفقرات متشابهًا بشكل كبير عبر المجموعات يمكن القول عندها بتحقق افتراض عدم اختلاف تقديرات برامترات الفقرة (Custer et al., 2008; Adedoyin, Nenty & Chilisa, 2008).

وقد تم اتباع أسلوب سيرسي (Sireci, 1991) لاختبار هذا الافتراض وفق الخطوتين الآتيتين. الأولى، وقد تطلبت تقسيم الأفراد إلى فئتين وفق وسيط تقدير قدرتهم (ذات القدرة المرتفعة والمنخفضة). ومن ثم حساب الدليل 2LLH- لمطابقة استجابات أفراد كل من المجموعتين (ذات القدرة المرتفعة والمنخفضة) لكل من النماذج الثلاثة: الأحادي والثنائي والثلاثي البراميتر؛ مع الأخذ بعين الاعتبار أن تقديرات برامترات الفقرة التي تم الحصول عليها من المجموعة كاملة اعتبرت على أنها برامترات مرجعية referent parameters يتم وفقها مقارنة البرامترات الناتجة من تحليل استجابات المجموعتين (ذات القدرة المرتفعة والمنخفضة) على الفقرات.

وتطلبت الخطوة الثانية من التحليل تضمين عضوية المجموعة في نموذج استجابة الفقرة. وعليه، فقد تم إعادة بناء البيانات الناتجة عن الاستجابات للفقرات، بحيث إنّ كل مجموعة من المجموعتين أجابت عن ١٦ فقرة من الاختبار المكون من ٣٢ فقرة (٢×١٦=٣٢). ومن ثمّ، فقد تضمن النموذج المقيد تقييد برامترات الفقرة (الصعوبة، والتمييز) للفقرات من ١ إلى ١٦ لكي تكون مساوية لبرامترات الفقرات من ١٧ إلى ٣٢. ويتطلب النموذج غير المقيد

حساب برامترات الفقرة من خلال التعامل مع البيانات على أنها مكونة من اختبار منفرد، يتكون من ٣٢ فقرة (وكل مجموعة لديها قيم مفقودة على ١٦ فقرة). ويسمح هذا النموذج غيرالمقيد بعمل معايرة آنية منفصلة لبرامترات الفقرات لكل من المجموعتين، في حين أن النموذج المقيد يقيد البرامترات بأن تكون متساوية في كل مجموعة.

وبسبب أن النموذج الثنائي البراميتر كان الأنسب للبيانات الناتجة عن الاستجابات للفقرات، فقد تم حساب الفرق بين دليلي 2LLH- للنموذج ثنائي البراميتر لكل من النموذج المقيد (الاستجابة عن ٣٢ فقرة) والنموذج غير المقيد (الاستجابة عن ١٦ فقرة) ومن ثم مقارنة النتيجة مع الإحصائي كاي تربيع مع ١٦ درجة حرية؛ وكانت القيمة تساوي ٩٠,٧، وهي دالة عند مستوى دلالة أقل من ٠,٠٥ (قيمة كاي تربيع الحرجة= ٢٦,٣٠). وهذا دليل على اختلاف برامترات الفقرة (Sireci, 1991). وتم إعادة الاختبار باستخدام كل من النموذجين الأحادي والثلاثي البراميتر، وكانت النتائج اختلاف برامترات الفقرة.

ولاستقصاء سبب هذا الاختلاف، فقد استخدمت إجراءات رب وزامبو (Rupp & Zumbo, 2004)، حيث تم ترتيب الفقرات وفق صعوبتها الناتجة عن التطبيق على المجموعتين: ذات القدرة المرتفعة والمنخفضة، ويبين الجدول رقم (٤) هذا الترتيب.

الجدول رقم (٤)
ترتيب الفقرات وفق الصعوبة الناتجة عن التطبيق على المجموعتين ذات القدرة المرتفعة والمنخفضة

	المجموعة ذات القدرة المرتفعة		المجموعة ذات القدرة المنخفضة
رتبة الفقرة	رقم الفقرة		رقم الفقرة
١	١٢		١٣
٢	١		١٢
٣	١٦		١٦
٤	١٤		١١
٥	١١		٣
٦	١٣		١٤
٧	٣		٩
٨	٩		٨
٩	١٠		١٠
١٠	٨		١٥
١١	١٥		٧
١٢	٧		٦
١٣	٥		١
١٤	٦		٥
١٥	٤		٢
١٦	٢		٤

ويتضح من الجدول السابق أن هناك فقرتين (رقم ١، ١٣) اختلف ترتيب صعوبتيهما بين المجموعتين اختلافا كبيرا (أكثر من رتبتين). وعليه، فقد تم استبعادهما من كلا النموذجين المقيد وغير المقيد، وتبقى ١٤ فقرة؛ وأعيد حساب الفرق بين دليلي 2LLH- للنموذجين (المقيد وغير المقيد) وفق كل من النماذج الأحادي والثنائي والثلاثي البراميتر، ومن ثم مقارنة النتيجة مع الإحصائي كاي تربيع مع درجات حرية ١٤؛ وكانت القيمة ١٧,٢ للنموذج الثنائي البراميتر، وهي دالة عند مستوى دلالة أكبر من ٠,٠٥ (قيمة كاي تربيع الحرجة=٢٣,٦٩). وهذا دليل على عدم اختلاف برامترات الفقرة بعد حذف هاتين الفقرتين (Sireci, 1991).

وللتحقق من أثر حذف هاتين الفقرتين في عدم اختلاف تقديرات القدرة، تم إعادة الإجراءات الواردة في البند أعلاه بعد حذفهما، وتم التوصل إلى نفس النتيجة السابقة إذ إنّ تقدير برامترات القدرة لا تختلف إذا كان النموذج أحادي البراميتر، وليس الثنائي البراميتر، هو المطابق للبيانات.

وفي ضوء ما سبق، يمكن القول إنّ النموذج الثنائي البراميتر الأنسب للبيانات الناتجة من الاستجابات للفقرات لا يحقق افتراض عدم اختلاف تقديرات القدرة، ولكنه حقق افتراض عدم اختلاف تقديرات برامترات الفقرات للاختبار المكون من أربع عشرة فقرة (بعد حذف الفقرتين).

رابعاً: نتائج السؤال الرابع

نص السؤال الرابع على ما يأتي: ما مدى تحقق معايير الدقة في تقدير قدرة المفحوص ممثلة بالخطأ المعياري في التقدير ودالة المعلومات للفقرات وللاختبار كاملاً؟ وللإجابة عن هذا السؤال، فقد تم إيجاد أكبر قيمة معلومات لكل من الفقرات وللاختبار كاملا، وتم رسم دالة معلومات الاختبار ودالة الخطأ المعياري في التقدير.

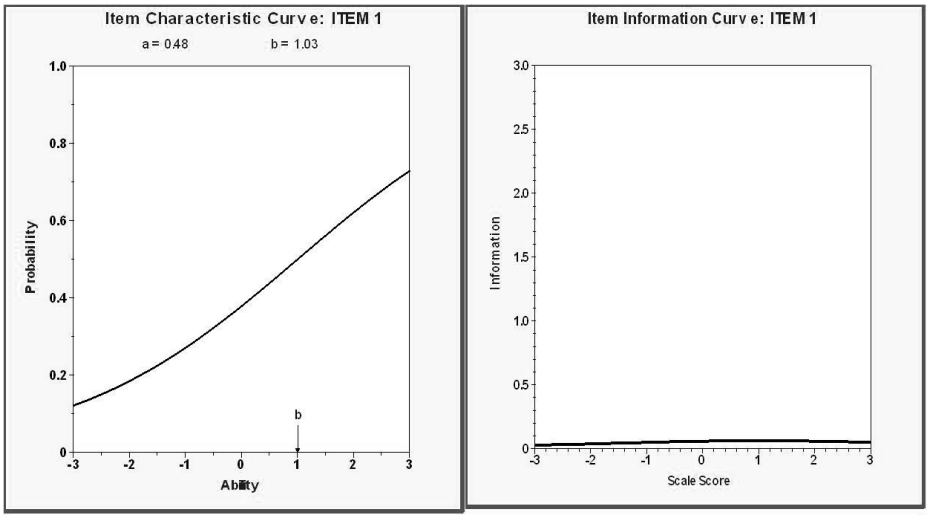
أ) دالة معلومات الفقرة

تزود كل فقرة في الاختبار ببعض المعلومات المتعلقة بقدرة المفحوص؛ وتعتمد كمية هذه المعلومات على مدى مطابقة صعوبة الفقرة مع قدرة المفحوص (Partchev, 2004). وقد تم رسم دوال معلومات كل الفقرات، وحساب أكبر قيمة لدالة معلومات كل منها، كما يظهر في الجدول رقم (٥) الذي يشير إلى أن أكبر قيمة لدوال المعلومات تراوحت بين ٢,٧٢ من الفقرة رقم (١٣) و ٠,٠٨ من الفقرة رقم (١).

الجدول رقم (٥)
أكبر قيمة لدوال معلومات كل من الفقرات

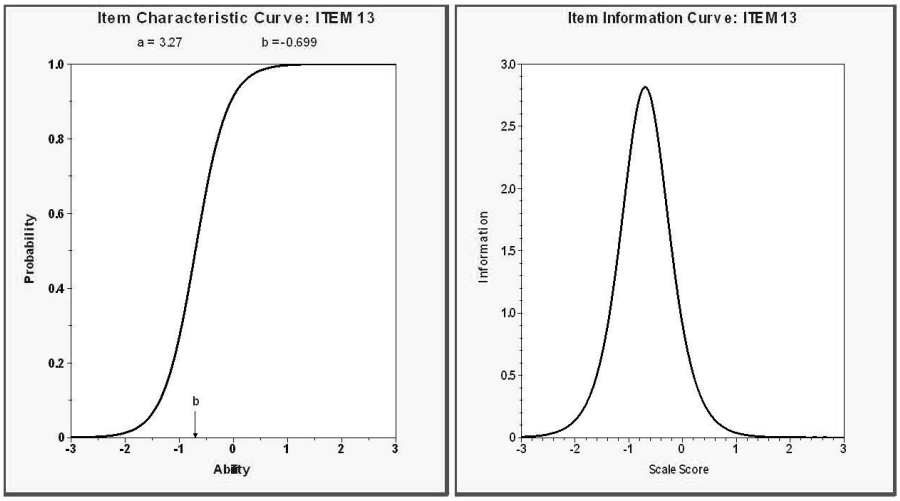
رقم الفقرة	أكبر قيمة	رقم الفقرة	أكبر قيمة	رقم الفقرة	أكبر قيمة	رقم الفقرة	أكبر قيمة
	٠,٠٨		١,٤٥		٠,٥١		٢,٧٢
	٠,٤٢		١,٢٢		٠,٥٢		٠,٤٢
	١,٠٦		٠,٨٠		٠,٤٣		
	١,٠٢		١,١٦		٠,٦٠		

ويبين الشكل رقم (٤) منحنى خصائص الفقرة رقم (١) ودالة معلوماتها. والفقرة ذات صعوبة مرتفعة وتمييزها قليل والمعلومات التي تزود بها قليلة.



الشكل رقم (٤)
منحنى خصائص ومنحنى معلومات الفقرة رقم (١)

ويبين الشكل رقم (٥) منحنى خصائص الفقرة رقم (١٣) ودالة معلوماتها. والفقرة ذات صعوبة متوسطة وتمييزها مرتفع والمعلومات التي تزود بها كبيرة.

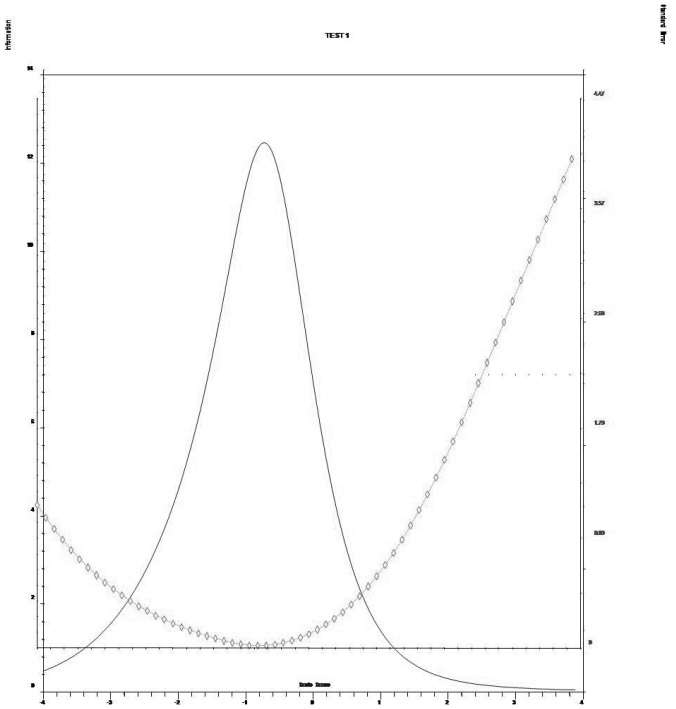


الشكل رقم (٥)
منحنى خصائص ومنحنى معلومات الفقرة رقم (١٣)

ولأن الفقرة رقم (١) تزود بمعلومات قليلة جداً، فقد تم حذفها (Zimowski, et al., 2003)، وعليه، فقد أصبح الاختبار مكوناً من ثلاث عشرة فقرة، فقط.

ب) دالة معلومات الاختبار والخطأ المعياري

ترتبط دالة معلومات الاختبار بالدقة التي نستطيع بها تقدير القدرة. وترتبط المعلومات التي يزود بها الاختبار عند مستوى قدرة معين ارتباطا عكسيا مع خطأ تقدير القدرة في مستوى القدرة المعين. ويبين الشكل رقم (٦) دالة معلومات الاختبار والخطأ المعياري في التقدير.



الشكل رقم (٦)
دالة المعلومات والخطأ المعياري للاختبار
***يشير الخط المتصل إلى دالة معلومات الاختبار، والخط المتقطع إلى الخطأ المعياري للاختبار**

يتبين من الشكل (٦) أن أكبر قيمة لدالة معلومات الاختبار تساوي ١٢,٣٣، وهي قيمة مناسبة لأنها أكبر من ١٠ (Hambleton, 2004). وهذه القيمة مناظرة لأدنى قيمة للخطأ المعياري. فكلما قل الخطأ المعياري، زادت المعلومات المتعلقة بالقدرة التي يزود بها الاختبار. ويظهر أيضاً، أن معظم المعلومات (الدقة في القياس والخطأ القليل) متضمنة في الفترة (−١,٥ إلى −٠,٥)، بمعنى أن هذا المقياس دقيق في الفترة السابقة، في حين أنه يفتقر إلى الدقة عند القدرة (θ) أقل من −١,٥، وأكبر من −٠,٥.

بتأمل النتائج التي تم التوصل إليها في هذه الدراسة، يمكن القول بأننا فشلنا في تطوير اختبار تحصيلي –عدد فقراته قليل– وفق نظرية الاستجابة للمفردة، بحيث يحقق افتراضاتها. وفضلاً عن ذلك، فالاختبار الناتج الذي يحقق معظم الافتراضات المتعلقة بالنظرية يتكون، فقط، من ثلاث عشرة فقرة (من أصل ٢٣ فقرة)، بمعنى أن الاختبار يقيس ١٣ هدفاً من الأهداف المخططة، أي تقريباً يقيس نصف الأهداف فقط. مما وضع تمثيل الاختبار للمحتوى

المستهدف موضع التساؤل. وقد أدت هذه النتيجة إلى العمل على تفحص صدق المحتوى؛ وعليه، فقد طلب من فريق العمل الحكم على درجة كفاية الأهداف التي يقيسها هذا الاختبار وهل يمكن اعتبارها ممثلة لمجتمع الأهداف المقصودة. وكانت النتيجة الإجماع على عدم تمثيلها للمجال المستهدف، فجميع الأهداف التي تقيس عملية القسمة لم تغطها فقرات الاختبار، إضافة إلى أربعة أهداف تقيس العمليات الثلاث الأخرى. بمعنى أن الاختبار الناتج يفتقر إلى صدق المحتوى بدلالة تمثيله للأهداف.

ومن المعلوم أن من استخدامات الاختبارات المدرسية ترتيب الطلبة وفق قدرتهم –ويعبر عنها بالعلامة على الاختبار. ولاختبار قدرة الاختبار على المحافظة على ترتيب الطلبة وفق قدرتهم؛ طبق الاختبار المكون من ١٣ فقرة على العينة الرئيسة (٢٨٤ طالباً وطالبة) بعد ثلاثة أسابيع من تطبيق الاختبار كاملا عليهم. وحسبت تقديرات قدراتهم على الاختبارين وفق النموذج الثنائي البراميتر باستخدام برمجية BILOG-MG، وتم إيجاد معامل ارتباط بيرسون بين تقديرات القدرة وفق الاختبارين، وكانت قيمته تساوي ٠,٦٥. وعلى الرغم من كونها دالة عند مستوى دلالة (α=٠,٠٥)، إلا أنه بتأمل هذه النتيجة يلاحظ أن نسبة التباين المشترك بين تقديري القدرة يساوي ٤٢,٢٥٪، فقط، بمعنى أن هناك ٥٧,٧٥٪ من التباين غير مشترك بين التقديرين، يعزى إلى متغيرات أخرى. ومفاد هذه النتيجة أنه لا يمكن القول أنّ الاختبارين يرتبان الطلبة بالطريقة ذاتها. وفي ضوء ما سبق، يمكن التوصل إلى أن الاختبار الذي يحقق معظم افتراضات النظرية (يتكون من ١٣ فقرة، فقط) يفتقر إلى الصدق من منظوري التمثيل للمحتوى والمترتبات على تفسير العلامة (Messick, 1995).

مناقشة النتائج

هدفت هذه الدراسة إلى استقصاء إمكان استخدام نظرية الاستجابة للمفردة في تطوير اختبار تحصيلي –عدد فقراته قليل؛ أي محاولة وضع النظرية موضع التطبيق. وقد تم تحقيق هدف الدراسة على مرحلتين: الأولى، بناء الاختبار. وقد تم تحديد ٢٣ هدفاً تقيس النتاجات التعلمية المتعلقة بالعمليات الأربع على الكسور العادية. ووضع فقرتين من نمط الاختيار من أربعة بدائل على كل هدف. ومن خلال عمليات التحكيم والتجريب وتحليل الفقرات باستخدام النظرية الكلاسيكية تم اختيار أفضل ٢٣ فقرة تقيس الأهداف المخططة، لتشكل اختبارا تحصيلياً. وتم التحقق من فعالية فقرات هذا الاختبار وثباته باستخدام النظرية الكلاسيكية.

وتطلبت المرحلة الثانية تطوير الاختبار باستخدام نظرية الاستجابة للمفردة. فقد استخدم اختبار مطابقة النموذج وتم التوصل إلى أن أنسب نموذج للبيانات الناتجة من استجابة الفقرات هو النموذج الثنائي البراميتر. وتم اختبار مطابقة كل من الفقرات لهذا النموذج، وتم التوصل إلى أن أربعا من فقراته لم تطابق النموذج وتم حذفها. ومن ثم تم التحقق من افتراضات النموذج، فقد بين التحليل العاملي الاستكشافي والتوكيدي أن الاختبار أحادي البعد، في حين كشف اختبارQ3 أن هناك ثلاث فقرات ليست مستقلة موضعياً، وقد تم حذفها. وعليه، فقد بقيت ست عشرة فقرة من أصل ثلاث وعشرين فقرة. وقد استخدم الدليل -2LLH للتحقق من عدم اختلاف تقديرات القدرة، وتبين أن النموذج أحادي البراميتر هو الذي يؤدي إلى تحقيق هذا الافتراض وليس النموذج الثنائي. ولدى استخدام الدليل نفسه للتحقق من عدم اختلاف تقديرات برامترات الفقرة تبين أن هذا الافتراض يتحقق بعد حذف فقرتين اختلف ترتيب صعوبتيهما عند تطبيقهما على مجموعتين إحداهما ذات قدرة مرتفعة، والأخرى ذات قدرة متدنية. وتم إيجاد أكبر معلومات لكل فقرة، ووجد أن هناك فقرة تزود بمعلومات قليلة جداً وتم حذفها، ومن ثم تم إيجاد دالة معلومات الاختبار الذي يتكون من ثلاث عشرة فقرة، ووجد أنه يزود بمعلومات مناسبة في فترة ضيقة يكون الخطأ المعياري فيها قليل.

وبالتأمل في نتائج الدراسة، يمكن ملاحظة ما يأتي: هناك تناقض بين عدم اختلاف تقديرات القدرة وبين مطابقة البيانات الناتجة عن الاستجابات للفقرات للنموذج. فمع تبيّن أن النموذج ثنائي البراميتر هو الأنسب لبيانات هذه الدراسة، إلا أن عدم اختلاف تقديرات القدرة ظهر أنها تتحقق عندما تطابق البيانات النموذج أحادي البراميتر. وقد يكون سبب التناقض الناتج صغر حجم العينة التي طبق عليها الاختبار (٢٨٤ فردا/٢٣ فقرة)، فقد تكون هناك حاجة إلى التطبيق على عينة أكبر من ذلك. أو قد يكون السبب قلة عدد الفقرات؛ فقد أشار كين (Kane, 2008) إلى أنه من الضروري أن يكون عدد الفقرات كبيراً بدرجة كافية كي يمكن الحصول على تقديرات مقبولة لعلامة الامتحان المتوقعة. ذلك أن الهدف في أثناء عملية التطوير كان بناء اختبار عدد فقراته قليل –شأنه شأن أي اختبار تحصيلي من إعداد المعلم.

وتتفق نتيجة هذه الدراسة المتعلقة بعدم تحقق افتراض عدم اختلاف تقديرات القدرة عند استخدام النموذج الأنسب مع نتائج دراسة كستر وآخرين (Custer, et al., 2008)، مع أن الأخيرة استخدمت بيانات ناتجة عن اختبارات عالمية تتضمن عددا كبيرا من الفقرات. ولكن هذه النتيجة تناقض نتائج دراسة كيلكار وآخرين (Kelkar, Wightman & Leucht,

2000) في أن إضافة برامترات إضافية للنموذج تحسن مطابقته للبيانات الناتجة عن الاستجابة لفقرات الاختبار. ويبدو أن هذه النتائج تشير إلى أنه عندما تسلك الفقرات بشكل غير متوقع، بمعنى أن تعزى استجابة الأفراد إلى عوامل أخرى غير القدرة من مثل: الغش، أو اللامبالاة، أو عدم الاهتمام، فسيحصل هناك اختلاف كبير في ترتيب رتب الأفراد أو الفقرات.

كما أنها تناقض نتائج دراسة ويلس وآخرين (Wells, Subkoviak & Serlin, 2002)، الذين استخدموا بيانات غير حقيقية (ناتجة عن المحاكاة simulation) بعكس هذه الدراسة. وقد يعزى هذا التناقض إلى أن بعض الملمحات أو التفاصيل المرتبطة بالبيانات الحقيقية يصعب إيجادها وتضمينها في البيانات المحاكاة. فمن الجدير بالذكر أن استخدام نظرية الاستجابة للمفردة في بيانات محاكاة سهل نسبياً، ذلك أن عمليات حذف الفقرات تتم دون مرجعية لما تقيسه الفقرة، وإنما تتم في ضوء عدم تحقيقها افتراضاً أو أكثر من افتراضات النموذج. ومن ثمّ، لا يتم النظر إلى صدق الاختبار، فلا توجد بنية نفسية أو أهداف تربوية محددة يتم قياسها. ولكن الوضع مختلف عند التعامل مع بيانات حقيقية تقيس بنىً أو أهدافاً محددة.

لقد كان الهدف من هذه الدراسة استقصاء إمكان استخدام نماذج نظرية الاستجابة للمفردة لبناء اختبار تحصيلي صادق وموثوق–عدد فقراته قليل، وكانت النتيجة الحصول على اختبار لا يحقق جميع افتراضات النموذج ويفتقر إلى الصدق. ومن ثمّ، فإن أي قرار سيبنى على نتائجه سيكون غير صحيح أو غير عادل. وعلى الرغم من أن ما تم إجراؤه في هذه الدراسة قد يبدو تطبيقاً لما وجد في كتب القياس والدراسات السابقة–شأنه في ذلك شأن العديد من الدراسات، إلا أن الخوض في التجربة الحقيقية يلقي الضوء على إمكان استخدام هذه الإجراءات من الفئة المستهدفة وهم المعلمون، علماً بأن معظمهم غير متخصصين في القياس. فعملية بناء وتطوير الاختبار التحصيلي من المعلمين وفق نظرية الاستجابة للمفردة قد تبدو صعبة. فالإجراءات التي اتبعت في البناء، مع أنها نفسها التي تتبع وفق النظرية الكلاسيكية، تحتاج إلى جهد في التطبيق والتجريب الذي قد يتطلب أكثر من مرة، قبل أن يركن المعلم إلى أن الاختبار الذي تم بناؤه مناسب للغرض ويقيس الأهداف أو النتاجات التعلمية المستهدفة، وهذا أسهل جزء في عملية تطوير الاختبار، مع أنه يتطلب أموراً لوجستية كثيرة. وأما إجراءات التطوير وفق نماذج نظرية الاستجابة للمفردة، فهي مشكلة أخرى. فهي تفترض في المعلمين معرفة بافتراضاتها، وقدرة على استخدام البرمجيات المختلفة، ومعرفة بالاختبارات الإحصائية المناسبة لاختبار أي من الافتراضات التي تقوم عليها النظرية، وقدرة على قراءة النتائج والتوصل إلى قرارات في ضوئها. وفي ضوء ما سبق، قد يكون من المناسب طرح التساؤلات الآتية: هل تصلح نظرية الاستجابة للمفردة لبناء الاختبارات الصفية ذات

العدد القليل من الفقرات والمفحوصين؟ وهل بالإمكان الطلب من المعلمين استخدام هذه النظرية في بناء وتطوير اختباراتهم الصفية، أم لا بد من أن يوكل أمر إعداد الاختبارات إلى مؤسسات أو شركات متخصصة تقوم ببناء الاختبارات وتطويرها للأغراض المختلفة، وفق هذه النظرية ويكون دور المعلمين استخدام هذه الاختبارات الموثوقة.

وعليه، وفي ضوء نتائج هذه الدراسة أوصي بإجراء دراسات تحدد أقل عدد من الفقرات الواجب تضمينها في الاختبار الصفي، وأقل عدد لازم من المفحوصين للتطبيق كي يحقق الاختبار افتراضات نماذج نظرية الاستجابة للمفردة والأهداف المبتغاة منه.

المراجع

أبو لبدة، خطاب (٢٠٠٣). **الأخطاء الرياضية عند االطلبة الأردنيين في الدراسة الدولية الثالثة للرياضيات والعلوم–إعادة**. عمان: المركز الوطني لتنمية الموارد البشرية.

التقي، أحمد (٢٠٠٩). **النظرية الحديثة في القياس** (ط١). عمان، الأردن: دار المسيرة.

الثبيتي، علي (٢٠٠٢). واقع الاختبارات المدرسية ومدى ملاءمتها لقياس الأهداف التعليمية. مجلة جامعة الملك سعود، **العلوم التربوية والدراسات الإسلامية، ١٤**(٢)، ٣٩٧–٤٣٠.

حرز الله، علية (٢٠٠٤). **بناء بنك أسئلة في الرياضيات والتحقق من فاعليته في انتقاء فقرات اختبار محكي المرجع في مستوى امتحان شهادة الدراسة الثانوية العامة في الأردن**. أطروحة دكتوراه غير منشورة، جامعة عمان العربية للدراسات العليا، عمان: الأردن.

دعنا، زينات (٢٠٠٢). **بناء اختبار المفاهيم الرياضية الأساسية لطلبة الصفوف الأساسية في الأردن على وفق الستراتيجية ثنائية المرحلة في نظرية السمات الكامنة**. أطروحة دكتوراه غير منشورة، كلية التربية–ابن رشد، جامعة بغداد، بغداد، العراق.

دعنا، زينات (٢٠٠٥). بناء اختبار محبوك هرمي في الرياضيات للصف الثامن الأساسي وفق أنموذج راش في نظرية السمة الكامنة. **دراسات، العلوم التربوية، ٣٥**(١)، ٤٢–٦١.

الرفيع، أحمد وسكاف، أنطون وأبو لبدة، خطاب والخضري، سليمان وساسي، محمد ومطر، محمد (٢٠٠٧). **نتائج الدول العربية المشاركة في الدراسة الدولية لتوجهات مستويات التحصيل في الرياضيات والعلوم «TIMSS ٢٠٠٣»**. عمان: النول الدولية للدعاية والإعلان.

الشريفين، نضال (٢٠٠٦). الخصائص السيكومترية لاختبار محكي المرجع في القياس والتقويم التربوي وفق النظرية الحديثة في القياس النفسي والتربوي. **مجلة العلوم التربوية والنفسية**، جامعة البحرين، ٧(٤)، ٧٩–١٠٩.

الصيداوي، أحمد (٢٠٠٤). **التقويم التربوي المستقبلي: من التشخيصي إلى التكويني إلى الأدائي إلى الحقيقي**. بيروت: مكتب اليونسكو الإقليمي للتربية في الدول العربية.

العطيوي، إيمان (٢٠٠٦)، **تطوير بنك فقرات في العلوم العامة باستخدام أساليب المعادلة الأفقية المستندة إلى النظرية الحديثة في القياس**. أطروحة دكتوراه غير منشورة، ، جامعة عمان العربية للدراسات العليا، عمان، الأردن.

عثمان، علام (٢٠٠٦). **بناء بنك أسئلة في الرياضيات للصف الثاني الثانوي العلمي باستخدام نظرية الاستجابة للفقرة**. رسالة ماجستير غير منشورة، الجامعة الأردنية، عمان، الأردن.

علام، صلاح الدين (١٩٩٥). **الاختبارات التشخيصية مرجعية المحك في المجالات التربوية والنفسية والتدريبية** (ط١). القاهرة: دار الفكر العربي.

الفرجات، هشام (٢٠٠٤). **بناء بنك أسئلة لمبحث الكيمياء للصف الثاني الثانوي العلمي**. رسالة ماجستير غير منشورة، جامعة مؤتة، الكرك، الأردن.

كاظم، أمينة (١٩٨٨). **استخدام نموذج راش في بناء اختبار تحصيلي في علم النفس وتحقيق التفسير الموضوعي للنتائج**. الكويت: مطبوعات جامعة الكويت.

مهيدات، عبد الحكيم (٢٠٠٥). **بناء بنك أسئلة للمهارات الرياضية في نهاية المرحلة الأساسية «نموذج مقترح»** .أطروحة دكنوراه غير منشورة ،جامعة اليرموك، إربد، الأردن.

النجار، نبيل .(٢٠٠٦). **بناء بنك أسئلة في مهارات الحاسوب للمرحلة الثانوية في الأردن باستخدام نماذج نظرية استجابة الفقرة «دراسة مقارنة بمعلمة ومعلمتين»**. أطروحة دكتوراه غير منشورة، جامعة اليرموك، إربد، الأردن.

وزارة التربية والتعليم (٢٠٠٥). **الخطة الاستراتيجية لوزارة التربية والتعليم خلال الفترة ٢٠٠٦/٢٠١٠**. المملكة الأردنية الهاشمية: منشورات إدارة البحث والتطوير التربوي.

يعقوب، منصور والمطرمي، عمر والعجارمة، أحمد والبرصان، إسماعيل وداهود، أنور والجراح، بندر والكردي، زياد والرقب، سعيد والفقهاء، عبد الرحمن والشرمان، محمود والهزايمة، محمود وشبانة، محمود والهندي، ميسر والعالم، ميسون. (٢٠٠٨). **نتائج الاختبار الوطني لضبط نوعية التعليم للعام الدراسي ٢٠٠٧-٢٠٠٨**. إدارة الامتحانات والاختبارات، مديرية الاختبارات. المملكة الأردنية الهاشمية: وزارة التربية والتعليم.

Adedoyin, O., Nenty, H. & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. **Educational Research and Review. 3**(2), 083-093,

Bond, T. & Fox, C. (2007). **Applying the Rasch model: Fundamental measurement in the human sciences,** (2ed ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Burket, G. (1984). Response to Hoover, **Educational Measurement: Issues and Practice, 3**(4), 15-16.

Crocker, L. & Algina, J. (1986). **Introduction to classical and modern test theory.** New York: Holt, Rinehart and Winston, Inc.

Custer, M., Sharairi, S., Yamazaki, K., Signatur, D., Swift, D., & Frey, S. (2008). **A paradox between IRT invariance and model-data fit when utilizing the one-, two- and three-parameter models.** Paper Presented at the Annual Meeting of the American educational research association, New York City, New York, March 28.

Doran, H. (2005). The information function for the one-parameter logistic model: Is it reliability?. **Educational and Psychological Measurement, 65,** 759-769.

Embretson, S. & Reise, S. (2000). **Item Response Theory for Psychologists**. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fan, X. & Ping, Y. (1999). **Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates.** Paper presented at the annual meeting of the American educational research association (Montreal, Quebec, Canada, April 19-23).

Green, D., Yen, W. & Burket, G. (1989). Experiences in the application of item response theory in test construction. **Applied Measurement in Education, 2**(4), 297-312.

Hambleton, R. (2004). **Personal communication.** Umeå University. Umeå: Department of Educational measurement.

Hambleton, R. & Jones, R. (1993). Comparison of classical test theory and item response theory and their applications to test development. **Educational Measurement: Issues and practice, 12**(3), 535-556.

Hambleton, R. & Swaminathan. H. (1985). **Item response theory: Principles applications.** Boston, Kluwer: Nijhoff Publishing.

Hays, R., Morales, L. & Reise, S. (2000). Item response theory and health outcomes measurement in the 21st century. **Medical Care, 38** (9 Supplement), 28-42.

Jiao, H. & Lau, A. (2003). **The effects of model misfit in computerized classification test.** Paper presented at the annual meeting of the National Council of Educational Measurement, Chicago, IL, April.

Jöreskog, K. & Sörbom, D. (1993a). **LISREL 8: Structural equation modeling with the SIMPLIS command language.** Hillsdale, NJ: Lawrence Erlbaum Associates.

Jöreskog, K. & Sörbom, D. (1993b). **LISREL 8 user's reference guide.** Chicago, IL: Scientific Software International.

Kane, M. (2008). Terminology, emphasis, and utility in validation, **Educational Researcher, 37**(2),76-82.

Kelkar, V., Wightman, L. & Leucht, R. (2000). Evaluation of the IRT Parameter Invariance Property for the MCAT. Paper presented at the annual meeting of the **National Council on Measurement in Education,** New Orleans, LA, April 25 - 27.

Kingstone, N., Leary, L., & Wightman, L. (1985). **An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test.** Educational Testing Service, Princeton, N.J. ED 268 141.

Lau, C. (1996). **Robustness of a unidimensional computerized mastery testing procedure with multidimensional testing data.** Unpublished doctoral dissertation. University of Iowa, Iowa City, IA.

Loehlin, J.C. (1987). **Latent variable models.** New Jersey: Lawrence Erlbaum Associates.

Meijer, R. & Sijtsma, K. (2008). **A review of methods for evaluating the fit of item score patterns on a test.** Research Report 99-01. (ERIC Reproduction Services ED 434 148).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. **American Psychologist, 50,** 741-749.

Partchev, I. (2004). **A visual guide to item response theory.** Friedrich-Schiller: Universität Jena.

Probst, T. (2003). Development and validation of the job security index and the job security satisfaction scale: A classical test theory and IRT approach. **Journal of Occupational and Organizational Psychology, 76,** 451-467.

Reid, C., Kolakowsky-Hayner, S., Lewis, A. & Armstrong, A. (2007). Modern psychometric methodology: Applications of item response theory. **Rehabilitation Counseling Bulletin, 50**(3), 177-188.

Rupp, A. & Zumbo, B. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. **Educational and Psychological Measurement, 65,** 588-599.

Santor, D. & Ramsay, J. (1998). Progress in the technology of measurement: Applications of item response models. **Psychological Assessment, 10,** 345-359.

Schilling, S. (2007).The role of psychometric modeling in test validation: An application of multidimensional item response theory. **Measurement: Interdisciplinary Research & Perspective, 5**(2). 93-106.

Shultz, K. & Whitney, D. (2005). **Measurement theory in action: Case studies and exercises.** California: Sage Publications Inc.

Sireci, S. (1991). Sample-independent item parameters? An investigation of the stability of IRT item parameters estimated from small data sets. Paper presented at the **Annual Meeting of the Northeastern Educational Research Association,** Ellenville, NY, October.

Spray, J., Abdel-fattah, A., Huang, C., & Lau, A. (1997). **Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional.** ACT Research Report Series 97-5. Iowa City, IA: American College Testing.

Stone, C., Weissman, A. & Lane, S. (2005). The consistency of student proficiency classifications under competing IRT models, **Educational Assessment, 10**(2), 125-146.

Thissen, D. & Orlando, M. (2001). Chapter 3-Item response theory for items scored in two categories. In Thissen D. & Wainer H. (Eds.), **Test Scoring.** Hillsdale, NJ: Erlbaum.

Weiss, D. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. J. Lubinski, and R. V. Dawis (Eds.), **Assessing individual differences in human behavior: New concepts, methods, and findings** (pp. 49-79). Palo Alto, CA: Davies-Black Publishing.

Wells, C., Subkoviak, M. & Serlin, R. (2002). The effect of item parameter drift on examinee ability estimates. **Applied Psychological Measurement, 26,** 77 - 87.

Wiberg, M. (2004). **Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test.** (EM No 50), Umea University Umea: Department of Educational measurement.

Yen, W. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. **Applied Psychological Measureurement, 8,** 125-145.

Yen, W. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), **Applications of item response theory** (pp. 123-141). Vancouver: Educational Research Institute of British Columbia.

Zagorsek, H., Stough, S. & Jaklic, M. (2006). Analysis of the reliability of the leadership practices inventory in the item response theory framework, **International Journal of Selection and Assessment, 14**(2), 180-191.

Zenisky, A., Hambleton, R. & Sireci, S. (2003). **Effects of local item dependent on the validity of IRT item, test and ability statistics.** Retrieved April 12, 2008, from: www.aamc.org/students/mcat/research/monograph5.pdf

Zenisky, A., Hambleton, R. & Sireci, S. (2006). Identification and evaluation of local item dependencies in the medical college admission test. **Journal of Educational Measurement, 39**(4), 291-309.

Zimowski, M., Muraki, E., Mislevy, R. & Bock, R. (2003). BILOG-MG 3. In Mathilda du Toit (Ed.). **IRT** from SSI: BILOG-MG, MULTILOG, PARSCALE, and TESTFACT. Chicago: Scientific Software International (SSI), Inc.