فاعلية الاختبار التكيفي المحوسب في تقدير القدرة العقلية باستخدام مصفوفات رافن

*أحمد سليمان عودة، وعمر سليمان عبيدات**

ملخص

هدفت هذه الدراسة إلى فحص فاعلية الاختبار التكيفي المحوسب في دقة تقدير القدرة العقلية باستخدام مصفوفات رافن باختلاف طرق تقدير القدرة (طريقة الارجحية العظمى MLE، طريقة التقدير البعدي الاعظم MAP) وقواعد إنهاء الاختبار (قاعدة الإنهاء عدد محدد من الفقرات، قاعدة الإنهاء أدنى خطأ معياري)، ولتحقيق أهداف الدراسة تم تكوين بنك أسئلة مكون من ١٠٥ فقرات من تلك المصفوفات. وإجراء خمسة تطبيقات محوسبة، بواقع اختبارين لكل تطبيق، على عينات بلغ عددها ٦٣٨ طالباً وطالبة. أظهرت النتائج أن قاعدة إنهاء الاختبار بعدد محدد من الفقرات توفر تقديرات للقدرة أدق، ودالة معلومات أعلى من قاعدة أدنى خطأ معياري باختلاف طريقتي تقدير القدرة، كما توفر قاعدة أدنى خطأ معياري في عدد الفقرات المطبقة بنسبة تصل إلى ٥٠% من قاعدة عدد محدد من الفقرات. كما أن الاختبار التكيفي يوفر تقديرات للقدرة أدق، ويوفر في عدد الفقرات المطبقة بنسبة تصل إلى ٧٠%، وله دالة معلومات أعلى من الاختبار الخطي باختلاف طريقتي تقدير القدرة. وتعطي كل من طريقتي تقدير القدرة (MLE, MAP) تقديرات قدرة متساوية ومؤشرات دقة متساوية ولطريقةMLE دالة معلومات للاختبار أعلى من MAP. ومن ابرز التوصيات استخدام الاختبارات التكيفية المحوسبة في الاختبارات العامة كاختبارات القبول واختبارات القدرات العقلية لفاعليتها من حيث الدقة، وفعاليتها الاقتصادية من زمن التطبيق، وتقليل الضغوطات على المفحوصين. واستخدام طريقة الأرجحية العظمى في تقدير القدرة بالاختبارات التكيفية لأن المؤشرات تعطي أفضلية لهذه الطريقة مقارنة بطريقة البعدي الأعظم. وإجراء دراسات مقارنة بين طريقتي التقدير MLE,MAP عند اختلاف عدد الفقرات الأصلية في بنك الأسئلة، وعند اختلاف قيم أدنى خطأ معياري غير التي استخدمت في الدراسة.

الكلمات الدالة: الإختبار التكيفي المحوسب، مفوفات رافن، نظرية استجابة الفقرة، بنك الأسئلة، قاعدة انهاء الإختبار التكيفي، دقة تقدير القدرة.

المقدمة

يتزايد الإهتمام ببناء وتطوير المقاييس والاختبارات العقلية والادائيه لقياس وتقييم قدرات الأفراد، وتحقيق معايير القياس العلمي الذي يضمن تقديرا كميا صادقا وبدرجة مقبولة من الدقة والموضوعية للدرجات التي تم اعتمادها كتقدير كمي لهذه القدرات، وقد واكب هذا الإهتمام تطور في البرمجيات ذات الصلة بتطبيقات النظرية الحديثة في القياس، والذي أدى الى ظهور الاختبارات المحوسبة (Murphy and Davidshofer, 1994) التي توفر امكانية التخزين والتطبيق للاختبارات.

وتختلف الاختبارات المحوسبة في طريقة تطبيقها، فقد يتم تقديم جميع الفقرات لجميع المفحوصين بنفس الترتيب، ويسمى ذلك الاختبار "الاختبار الخطي المحوسب"، أو أن تقدم للمفحوص الواحد الفقرات التي تتناسب مع مستواه فقط، ويطلق عليه "الاختبار التكيفي المحوسب" "Computerized adaptive testing (CAT)، حيث يُعرض على المفحوص بعض الفقرات لتحديد قدرته المبدئية، وبناءً على قدرته المبدئية المقدرة من خلال أدائه على تلك الفقرات تُقدم له فقرات أخرى لاحقة من بنك الأسئلة بما يتناسب مع التقدير المستمر لمستواه، ويعتمد اختيار الفقرة اللاحقة على استجابات المفحوص على الفقرات السابقة، وخصائص تلك الفقرات (Cisar, Radosav, Markoski, Pinter and Cisar, 2010)

وغالباً ما يتم اختيار الفقرة الأولى التي تقدم للمفحوص حول متوسط القدرة صفر (θ=0)، وتختار الفقرة التالية باستخدام قاعدة القفز Step size، بحيث إذا أجاب المفحوص على الفقرة إجابة صحيحة يتم اختيار الفقرة اللاحقة عند قدرة أعلى مثل (θ=١) أو (θ=0.5)، أما إذا أجاب إجابة خاطئة فيتم اختيار الفقرة عند مستوى قدرة أقل مثل (θ=-1) أو

* قسم علم النفس والإرشاد التربوي، كلية التربية، جامعة اليرموك، إربد، الأردن. تاريخ استلام البحث ٢٠٠٩/٥/٤، وتاريخ قبوله ٢٠١٣/٥/٩.

© ٢٠١٣ عمادة البحث العلمي/ الجامعة الأردنية. جميع الحقوق محفوظة.

(θ =-0.5) وهكذا (Cisar, et al., 2010; Jacobusse and Buuren, 2007; Weiss,2004). كما يعرف الاختبار التكيفي بأنه الاختبار المُفصّل أو المحبوك Tailored testing المتمثل بتقديم فقرات ذات صعوبة تتناسب مع قدرات المفحوصين (Jain-quan , Dan-min xia and, Jing-jing, 2007) على النحو الآتي:

١) بعد تقدير القدرة الأولى يقوم الحاسب باختيار فقرة جديدة تكون ذات قيمة في عملية تقدير القدرة الحقيقية، وهذا مرتبط بطريقة اختيار الفقرة.
٢) يتم تقديم الفقرة للمفحوص والإجابة عنها، ثم تصحيحها.
٣) يتم إعادة تقدير القدرة بناءً على جميع الفقرات التي تم تقديمها للمفحوص باستخدام إحدى طرق التقدير.
٤) تحديد فيما اذا استوجب التطبيق تقديم فقرة أخرى بناء على المحك المستخدم في قاعدة إنهاء الاختبار.
٥) إذا لم يتحقق محك إنهاء الاختبار يتم الرجوع للخطوة الأولى، وإذا تحقق يتم إنهاء الاختبار.

ومن فوائد الاختبار التكيفي انهُ اكثر مرونة من الخطي، ويوفر عدة صور للاختبار (Cisar, et al., 2010)، فهو يقلص عدد الفقرات الضرورية للوصول إلى مستوى معين من الدقة في القياس، ويقل الوقت المطلوب للاختبار (Stone and Davey, 2011)، كما يقلل من احتمالية معاناة المفحوص من التعب والملل وتراجع مستوى الدافعية (Huo, 2009; Magis and Raiche, 2011) كما أنه يقدم معلومات أكثر عند أطراف متصل القدرة أكثر من أي اختبار آخر، ويوفر معلومات كافية عن مستويات متوسطة من الصعوبة القدرة، ويتطلب عدداً أقل من الفقرات للحصول على المستوى المطلوب من الدقة والثبات مقارنة بالإختبارات الخطية (Vispoel, 1993).

وتكمن أهمية استخدام الاختبارات التكيفية المحوسبة في مزاياها على الاختبارات التقليدية، حيث ان خوارزميات الحاسب تقوم بدور أخصائي الاختبارات والقياس النفسي، والاختبار يتكيف مع قدرة المفحوص؛ فهي تختلف تماماً عن الاختبارات الخطية والعشوائية ليس بسبب اختلاف عدد وخصائص الفقرات التي يتلقاها كل مفحوص عن الآخر فقط، بل يمكن ان يتم تقدير القدرة عن طريق النظرية الحديثة في القياس ونماذجها المختلفة بعد الاستجابة عن كل فقرة إما بطريقة الارجحية العظمى Maximum Likelihood، أو بالطرق البييزية Bayes، وتعتبر عملية اختيار طريقة التقدير الملائمة للقدرة ذات أهمية بالغة لأنها لا تؤثر فقط في النتيجة النهائية للاختبار، بل تؤثر في طريقة اختيار الفقرات التي يتم تطبيقها، وفي قاعدة إنهاء الاختبار.

يتطلب الاختبار التكيفي المحوسب بناء بنك للأسئلة كخطوة اولى، لأنه يعتمد على مجموعة كبيرة نسبيا من الفقرات ذات معالمParameters معروفة. وترى أمبرتسون ورايس (Embretson and Reiase, 2000) أن القياس الدقيق يتطلب احتواء هذا البنك على عدد كافٍ من الفقرات، وذات صعوبة موزعة بشكل جيد على متصل السمة. ويُعرّف كل من ميلمان وآرتر (Millman and Arter, 1984) ذلك البنك بأنه "مجموعة كبيرة من الأسئلة التي يسهل استخدامها، وأن المعلومات عن كل فقرة مفهرسة وموصوفة بصورة كاملة، ويسهل انتقائها في أي اختبار بمواصفات محددة ، ومن تلك المعلومات مثلاً: الإجابة الصحيحة، ومعالم الصعوبة، والتمييز، والتخمين. وأن ٣٠ فقرة بهذه المواصفات يمكن أن تكون كافية وملائمة لإعداد اختبار تكيفي (Dodd, Koch and De Ayla, 1989; Chen, Hou and Dodd, 1998)

أما الخطوة التالية فهي تحديد أسلوب اختيار الفقرة، ومنها أسلوب مطابقة الصعوبة (b_i) مع القدرة المقدّرة ($\hat{\theta}$). ففي هذا الأسلوب يتم اختيار الفقرة ذات الصعوبة الأقرب للقدرة المقدّرة، وبعد أن يستجيب المفحوص على الفقرة أو الفقرات الأولية لتقدير القدرة يتم اختيار الفقرة ذات المعلومات الأكبر عند مستوى القدرة المقدّرة للمفحوص، ثم تقدر هذه القدرة من جديد وهكذا...، حتى يتم إنهاء الاختبار. ويتمتع هذا الأسلوب بأنه يوفر أقصى معلومات بأقل خطأ في القياس عند مستوى القدرة المستهدفة (Kingsbury and Zara, 1989). (Vispoel,1993).

وهناك عدة طرق لتقدير القدرة Ability estimation methods ابرزها طريقة الأرجحية العظمى Maximum Likelihood Estimation (MLE) والطرق البييزية Bayesian Methods التي تتضمن طريقة التقدير البعدي المتوقع Expected A Posterior (EAP) وطريقة التقدير البعدي المُعظم Maximum A Posterior (MAP). الا انها تختلف في دقة التقدير الذي يؤثر بالتالي في كيفية إنهاء الاختبار.وتُعد الأرجحية العظمى من أكثرها استخداماً، وتعتمد في تقدير القدرة على نمط إجابة المفحوص لمجموعة فقرات تأخذ واحداً أو صفراً. أما طريقة التقدير البعدي المُعظم (MAP) فهي تستخدم معلومات سابقة عن توزيع القدرة، ويتم افتراض هذا التوزيع بناء على معلومات سابقة، والتوزيع الأكثر استخداماً هو التوزيع الطبيعي المعياري. والمشكلة في هذا الأسلوب أن التقدير يكون متحيزا عندما يكون عدد الفقرات أقل من ٢٠ (Embreston and Reiase, 2000).

ومن حيث مؤشرات الدقة كوجه من أوجه المقارنة بين طرق تقدير القدرة فقد اشاروانج وفيزبول(Wang and Vispoel, 1998)

ووانج (Wang, 1995) أن لطريقة MLE أخطاء معيارية أعلى، وكان الجذر التربيعي لمتوسط مربعات الخطأ Root (RMSE) Mean Square Error أعلى، وتحيزاً أقل من الطرق البيزية، كما أن لطريقة (EAP) دقة أعلى، وأكد وانج ووانج (Wang and Wang, 2001) أنMLE لها تحيز أقل وخطأ معياري أكبر من طريقتي EAP وMAP، وان EAP أكثر دقة من MAP ، كما أشارا إلى أن لطريقتي EAP وMAP مزايا على طريقتي WLE وMLE من حيث كفاءة الاختبار، وأنه يوجد أثر لطريقة إنهاء الاختبار أكبر من أثر حجم بنك الأسئلة على دقة طرق تقدير القدرة. وحسب نتائج دراسة أخرى فقد كان طول الاختبار المثالي للطرق البيزية ١٤ فقرة، ولطريقة الأرجحية العظمى ١٢ فقرة، وان للطرق البيزية متوسط معلومات أعلى من طريقة الأرجحية العظمى. وأشار روسو وريكاس (Rosso and Reckase, 1981) إلى انه بالإمكان الحصول على تقديرات مقبولة للقدرة باستخدام ١٢-١٤ فقرة.

وأما الخطوة الأهم في الإختبارات التكيفية فهي تحديد قاعدة الإنهاء للاختبار، حيث ينتهي الاختبار التكيفي المحوسب عادةً عند تطبيق عدد محدد من الفقراتFixed Length بدرجة مقبولة من الدقة لكل مفحوص، ووفقاً لهذه القاعدة يأخذ كل مفحوص مجموعة فقرات مختلفة عن المفحوصين الآخرين وبنفس العدد. كما وينتهي تقديم الاختبار عند الوصول لأدنى خطأ معياري Minimum Standard error محدد مسبقا. وقد أشارت أمبرستون وريس (Emperston and Reiase, 2000) أنّهَ في الغالب يستخدم الخطأ المعياري ٠.٢٥ كحد اقصى. وينتهي عند الوصول لأدنى قيمة للمعلومات، حيث يستمر تطبيق الفقرات طالما ان هناك امكانية للحصول على مقدار أكبر من المعلومات عن المفحوص، وعندما يصل الاختبار إلى أدنى زيادة في المعلومات المصاحبة لتقدير القدرة يتوقف الاختبار. وأشار روكس وديجريس (Roex and Degryse, 2004) أن CAT يحتاج بالمتوسط إلى حوالي ٤٠ فقرة. وبالمقابل أكد فليج وبيكر والتر وبجورنر وكلّب وروز (Fliege, Becker, Walter, Bjorner, Klappand Rose, 2005) الى إمكانية تقدير السمة باستخدام ست فقرات تقريباً وبخطأ معياري محدد مسبقاً أقل أو يساوي (٠.٣٢).

ومن حيث عدد الفقرات الداخلة في الاختبار كوجه من أوجه المقارنة بين الاختبارات الخطية والاختبارات التكيفية فقد أشار وارد (Ward, 1984) أن هذه الاختبارات تحتاج بشكلٍ عام إلى عدد من الفقرات تقل بنسبة ٥٠% إلى ٦٠% من الاختبارات الخطية عند المستوى نفسه من الدقة. كما وأشار لنِدن وبشلي (Linden and Pushley,2003) إلى أن الاختبارات التكيفية المحوسبة ذات الطول المتغير التي تتوقف عند الوصول لخطأ معياري معين قد خفضت عدد الفقرات التي تطبق بمقدار يصل إلى حوالي ٩١%. والاختبارات ذات الطول الثابت التي تتوقف عند تطبيق عدد محدد من الفقرات قد خفضت عددها إلى حوالي ٨٩%. كما أشار كل من ستون وديفي (Stone and Davey, 2011) أن CAT ذات الطول المتغير يزيد من كفاءة الاختبار، ويوفر ما نسبتة ٤٠% الى ٥٠% في عدد الفقرات التي يحتاجها الاختبار الخطي. وخلصت دراسة الخضر وكلارك وأندرسون (Alkhder, Clark, Anerson, 1998) إلى إمكانية تخفيض عدد الفقرات في الاختبارات التكيفية المحوسبة لقياس الاستعداد الى أقل من ٥٠% من عدد فقراته الأصلية في قياس قدرة الأفراد بدرجة مكافئة للاختبار الأصلي. وأشار فيزبول ووانج وبلير (Vispoel and Wang and Bleier, 1997) إلى أنَّ الاختبار التكيفي للموسيقى يحتاج إلى فقرات تقل بنسبة ٥٠% إلى ٩٣% ليطابق الثبات والصدق التلازمي لاختيار الفقرات الثابتة العدد في الإختبار الخطي، كما أشار فيزبول ووانج (Vispoel,1988) عند تقييم كفاءة الاختبار التكيفي بتطبيقه على (٢٠) مفحوصاً، فكانت النتيجة ان هناك حاجة إلى ما متوسطة (٦، ٩، ١٢) فقرة للحصول على معاملات ثبات (٠.٨٠، ٠.٨٥، ٠.٩٠) على التوالي.

أهمية الدراسة

يتضح من الدراسات السابقة أنها استخدمت بيانات مولدة، ولم تقارن بين طرق تقدير القدرة وقواعد إنهاء الاختبار باستخدام النموذج ثلاثي المعلمة. كما انها لم تقارن بين الاختبارات الخطية المحوسبة والاختبارات التكيفية المحوسبة؛ ولذلك تشكل الدراسة الحالية إضافة تتمثل في التعامل أولاً مع بياناتٍ واقعية وذلك بتحويل اختبار قدرات عقلية كان قد بنيَ أصلاً وفق النظرية التقليدية إلى اختبار تكيفي محوسب وفق النظرية الحديثة (النموذج ثلاثي المعلمة)، ثم دراسة فاعلية الاختبار التكيفي المحوسب باستخدام طريقتين لتقدير القدرة هما الارجحية العظمى MLE والبعدي الاعظم MAP، وطريقتين لإنهاء الاختبار هما: عدد محدد من الفقرات والخطأ المعياري، وقد استخدم في هذه الدراسة برنامج حاسوبي في إدارة بنوك الأسئلة ويتعامل مع الاختبارات التكيفية والخطية والعشوائية المحوسبة، وهو برنامج فاست تست برو (Fasttest pro-2006) المعد اصلاً لتنفيذ الاختبارات التكيفية والاختبارات الخطية المحوسبة، وقد تم الحصول على هذا البرنامج من مؤسسة Assessment Software على الموقع www.assess.com وذلك

بدعم من عمادة البحث العلمي والدراسات العليا بجامعة اليرموك لتغطية تكاليف الحصول على هذا البرنامج.

مشكلة الدراسة وأسئلتها

في ضوء التصور السابق لأهمية الدراسة وادبياتها التي ساهمت في تحديد الإطار لمشكلة الدراسة، ولتأمين بيانات حقيقية من اختبارات شائعة الإستخدام في قياس القدرة العقلية العامة ومتحررة نسبيا من الثقافات، فقد تم اختيار اختبار مصفوفات رافن الذي يتكون من ثلاث مصفوفات لقياس الذكاء العام للأفراد (الملونة، المعيارية، المتقدمة)، والتعامل مع البيانات المستخلصة منه وفقا للنموذج ثلاثي المعلمة (Three parameters logistic model) كمرحلة أولى، ومن ثم تطويره إلى اختبار تكيفي محوسب في المرحلة الثانية وفقا للخطوات الآتية:

– تطوير اختبار مصفوفات رافن للقدرات العقلية الذي يقدم بالطريقة التقليدية (الورقة والقلم) إلى اختبار تكيفي محوسب.

– دراسة فاعلية الاختبار التكيفي المحوسب باستخدام طريقتين لإنهاء الاختبار (عدد محدد من الفقرات، أدنى خطأ معياري) باستخدام طريقتين لتقدير القدرة (MLE, MAP).

– دراسة فاعلية الاختبار التكيفي المحوسب الذي ينتهي باستخدام قاعدة أدنى خطأ معياري مقارنة بالاختبار الخطي المحوسب باستخدام طريقتين لتقدير القدرة (MLE, MAP).

– دراسة فاعلية الاختبار التكيفي المحوسب الذي ينتهي باستخدام قاعدة أدنى خطأ معياري باستخدام طريقتين لتقدير القدرة (MLE, MAP).

وفي ضوء هذا التصور أيضا، فإن المشكلة تتلخص في الإجابة عن الأسئلة المتضمنة في اجراءات تقصي فاعلية الاختبار التكيفي المحوسب في عدة اوضاع وتصاميم تجريبية تختلف بالتناوب في قاعدة انهاء الإختبار او/ و طريقة تقدير القدرة على النحو الآتي:

١. استخدام قاعدة إنهاء الاختبار بعدد محدد من الفقرات، مقارنة بالاختبار التكيفي المحوسب (باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري) واستخدام طريقة تقدير الأرجحية العظمى (MLE) للقدرة؟
٢. استخدام قاعدة إنهاء الاختبار بعدد محدد من الفقرات، مقارنة بالاختبار التكيفي المحوسب (باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري) واستخدام طريقة التقدير البعدي الأعظم (MAP) القدرة؟
٣. استخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري، مقارنة بالاختبار الخطي المحوسب باستخدام طريقة تقدير الأرجحية العظمى(MLE) للقدرة؟
٤. استخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري، مقارنة بالاختبار الخطي المحوسب باستخدام طريقة التقدير البعدي الأعظم (MAP) القدرة؟
٥. استخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري، وباستخدام طريقة تقدير الأرجحية العظمى(MLE) لتقدير القدرة، مقارنة بالاختبار التكيفي المحوسب (باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري) واستخدام طريقة التقدير البعدي الأعظم (MAP) للقدرة؟

تعريف المصطلحات

الاختبار التكيفي المحوسب: هو الاختبار الذي يُفصل لكل مفحوص على حدة، وذلك بعرض الفقرات التي تتناسب مع مستوى قدرته، مما يمكن من تقدير أدق للقدرة بأقل عدد من الفقرات.

الفاعلية: دقة قياس الذكاء العام المقدرة بمصفوفة رافن والمستخلصة في هذه الدراسة بالمؤشرات الإحصائية التالية: متوسط عدد الفقرات المطبقة في الاختبارات، ودالة المعلومات التي توفرها الاختبارات، والخطأ المعياري في تقدير القدرة، والخطأ المعياري في القياس، والجذر التربيعي لمتوسط مربعات الخطأ، والتحيز في تقديرات القدرة المقدرة.

الاختبار الخطي: هو الاختبار الذي يتم فيه تطبيق جميع فقراته على جميع المفحوصين بالعدد نفسه والترتيب نفسه.

عدد محدد من الفقرات (الطول الثابت للاختبار): هي قاعدة لإنهاء الاختبار التكيفي المحوسب، ويتم فيها تطبيق عدد محدد من الفقرات، وفي هذه القاعدة يأخذ كل مفحوص مجموعة فقرات مختلفة عن الآخرين بالعدد نفسه.

أدنى خطأ معياري: هي قاعدة لإنهاء الاختبار التكيفي المحوسب، وفيها يستمر تقدير القدرة بعد تطبيق كل فقرة، وعندما يصل الخطأ المعياري لتقدير القدرة عند قيمة محددة سلفاً ينتهي الاختبار.

الأرجحية العظمى (MLE):هي إحدى طرق تقدير القدرة في نظرية استجابة الفقرة، ويتم إيجاد المعالم من خلال إجراءات تعظيم الاحتمالية للمعالم المراد تقديرها، وتعتمد هذه الطريقة في تقدير القدرة على نمط إجابة المفحوص لمجموعة فقرات حيث تأخذ الإجابة إما واحد أو صفر.

التقدير البعدي الأعظم (MAP): هو إحدى الطرق البيزية في نظرية استجابة الفقرة لتقدير القدرة، وتستخدم فيه معلومات سابقة عن توزيع القدرة، ويتم افتراض شكل هذا التوزيع بناء على معلومات سابقة، والتوزيع الأكثر استخداما هو التوزيع الطبيعي المعياري.

محددات الدراسة

تتلخص المحددات التي تحد من تعميم النتائج في هذه الدراسة بما يلي:

١. لا يتعامل برنامج فاست تست برو Fasttest Pro مع صعوبة الفقرات التي تزيد قيمتها على ٥ وتقل عن -٥. وعلى الرغم من اتساع هذا المدى، الا ان الفروق بين الأفراد وتوزيع القدرة قد يتعدى هذا المدى. وقد تم التعامل مع الفئات المفتوحة خارج هذا المدى بناء على استشارة معد هذا البرنامج.

٢. اقتصرت قواعد إنهاء الاختبارات التكيفية على ٠.٢٥ لقاعدة أدنى خطأ معياري و٣٠ فقرة لقاعدة عدد محدد من الفقرات، ويمكن أن تكون أي من القاعدتين متغيراً من المتغيرات في دراسات لاحقة.

٣. اقتصرت عينات الدراسة على طلبة البكالوريوس في كلية التربية بجامعة اليرموك، حيث يحتاج التطبيق إلى تسهيلات بحثية، خاصة ما يتعلق بتوفير مختبرات ومشرفين وفق شروط استخدام البرنامج من حيث عدد الأجهزة.

الطريقة والإجراءات

تتطلب الإجراءات في هذه الدراسة إعداد بنك الأسئلة من مصفوفات رافن، والتعريف بالبرنامج المستخدم في بناء البنك وما يتطلبه من بيانات ومعالم تصف خصائص فقراته، ثم تطبيقها على عينة المفحوصين للوصول إلى اختبارات تكيفية بخصائص محددة. وفيما يلي تعريف بالمصفوفة والعينات واجراءات بناء البنك، قبل البدء بتطبيق إجراءات جمع بيانات الاختبارات التكيفية والخطية.

التعريف باختبار مصفوفات رافن

يتكون اختبار رافن من ثلاث مصفوفات (الملونة والمعيارية والمتقدمة وبعدد فقرات (٣٦) (٦٠) (٤٨) فقرة على التوالي، وهي متدرجة في الصعوبة من السهل إلى الصعب، وتغطي متصل القدرة العقلية العامة أو الذكاء العام في المراحل العمرية المختلفة، وتكفي لبناء بنك الفقرات اللازم للاختبار التكيفي المحوسب، ومن المعلوم أن هذه المصفوفات، كان قد تم إعدادها أصلاً لقياس الذكاء العام وفق النظرية التقليدية، وقسمت فقرات المصفوفات الثلاث في الدراسة الحالية إلى اختبارين، يتكون كل منهما من (٧٠) فقرة، حيث احتوى الاختبار الأول على جميع فقرات المصفوفة المعيارية وعشر فقرات مشتركة من فقرات المصفوفة المتقدمة تم اختيارها بالطريقة العشوائية البسيطة، أما الاختبار الثاني فقد احتوى على جميع فقرات المصفوفة المتقدمة وعددها ٤٨ فقرة و(١٢) فقرة من المصفوفة الملونة وعشر فقرات مشتركة اختيرت بشكل عشوائي من فقرات المصفوفة المعيارية، هذا مع ملاحظة أن (٢٤) فقرة من فقرات المصفوفة الملونة هي الفقرات ذاتها الـ (٢٤) الأولى من المصفوفة المعيارية.

عينات الدراسة

تكونت عينة الدراسة الأولى من ٢٦٩٥ طالباً وطالبة من طلبة كلية التربية في جامعة اليرموك، وذلك لأغراض بناء بنك الأسئلة، وتم اختيار الشُعب بطريقة عشوائية بسيطة، وتقدم ١٣٤٠ طالباً وطالبة للاختبار الأول، و١٣٥٥ طالباً وطالبة للاختبار الثاني.

وتكونت عينة الدراسة الثانية من ٦٣٨ طالباً وطالبة من طلبة كلية التربية بطريقة عشوائية بسيطة من الشعب بواقع (٩٢، ١٥٣، ١٦٣، ١٤٩، ٨١) لكل سؤال من أسئلة الدراسة على الترتيب من مختلف التخصصات بغرض تطبيق الاختبارات التكيفية المحوسبة.

إجراءات جمع البيانات لبناء بنك الأسئلة

تم تجهيز الاختبارين بصورتهما النهائية، وصممت أوراق إجابة خاصة تُصَّحح أليا باستخدام جهاز الماسح الضوئي Optical Mark Reader (OMR) المتوفر في الجامعة. وتم التطبيق على الطلبة في القاعات الدراسية في الوقت المخصص للمحاضرات بموافقة رسمية من الجامعة، وقراءة استجابات الطلبة باستخدام القارئ الضوئي، ومن ثم نقلها إلى برنامج اكسل EXCEL، وصُحِحت الاستجابات أليا بإعطاء العلامة (١) للإجابة الصحيحة و(صفر) للإجابة الخطأ.

كانت مصفوفات رافن قد أُعدت أصلاً لقياس الذكاء العام للأفراد، وهي بالتعريف تقيس سمة واحدة في أساسها النظري، إلا انه تم التحقق من افتراض احادية البعد (Unidimensionality) باستخدام برمجية نوهارم NOHARM-2003 لفحص احادية البعد لكل من الاختبارين، حيث تقوم هذه البرمجية بفحص احادية البعد بحساب مؤشر تاناكا Tanaka لحسن المطابقة، فإذا كانت قيمة هذا المؤشر أكبر من علامة القطع ٠.٨ فإنه مؤشر على احادية البعد (Clarke, Mackinnon, Mckenzie and Herrman, 2000). وبحساب مؤشر تاناكا للاختبارين على البيانات من عينة الدراسة، كانت قيمته للاختبار الأول تساوي ٠.٩٤ و٠.٨٣ للاختبار الثاني.

كما تم التحقق من افتراض الاستقلال الموضعي (local independence) باستخدام الإحصائي (Q3) وهو معامل الارتباط للبواقي لزوج من الفقرات بعد ضبط السمة المقدرة، وقد كانت قيمتة للاختبارين (٠.٠٠٣، -٠.٠١٢) على الترتيب، وكانت جميع معاملات الارتباط للبواقي بين أزواج الفقرات صغيرة (قريبة من الصفر) وهذا مؤشر على تحقق هذا الإفتراض أيضا.

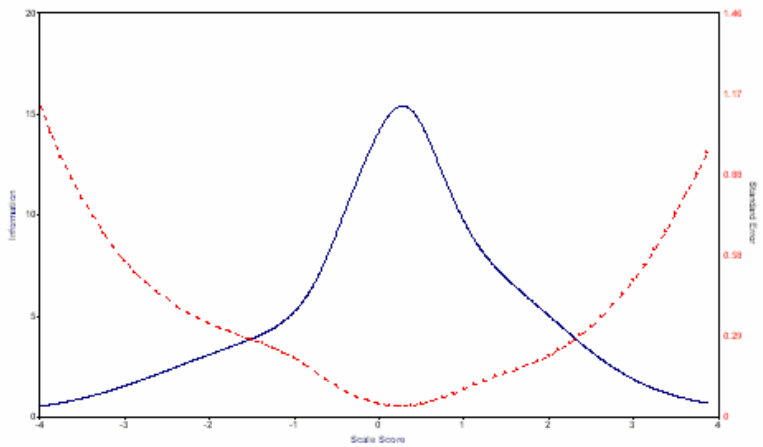
كما تم التحقق من مدى مطابقة الأفراد للنموذج ثلاثي المعلمة باستخدام برنامج Bilog-mg لتحليل بيانات الاختبارين بشكل مستقل، فقد تبين عدم مطابقة ٢٣ فرداً للاختبار الأول و٢٧ فرداً للاختبار الثاني وحذفت استجابات هؤلاء الأفراد. ثم تم إعادة التحليل لاختبار مدى مطابقة الفقرات لكل من الاختبارين بشكل مستقل وفق النموذج ثلاثي المعلمة، وتبين عدم مطابقة ١١ فقرة، منها فقرتان مشتركتان للاختبار الأول، و٧ فقرات منها ٥ فقرات مشتركة للاختبار الثاني، حيث كانت القيمة الاحتمالية للمطابقة أقل من ٠.٠١ لهذه الفقرات، واستقرعدد فقرات الاختبار الأول على ٥٩ فقرة و٦٣ فقرة في الاختبار الثاني بواقع ١٤ فقرة مشتركة بين الاختبارين، وبلغ معامل ثبات الاتساق الداخلي المحسوب بمعادلة كودر-ريتشاردسون ٢٠ (KR-20) 0.88 للاختبار الأول و0.87 للاختبار الثاني.

وتم تقدير معالم الفقرات بطريقة الارجحية العظمى الهامشية، كما تم حساب دوال المعلومات لفقرات الاختبارين باستخدام برمجية Bilog-mg. ويبين الجدول (١) وصفا احصائيا لمعالم فقرات الاختبارين.

الجدول (١): الإحصائيات الوصفية لمعالم الفقرات ودالة المعلومات للاختبارين

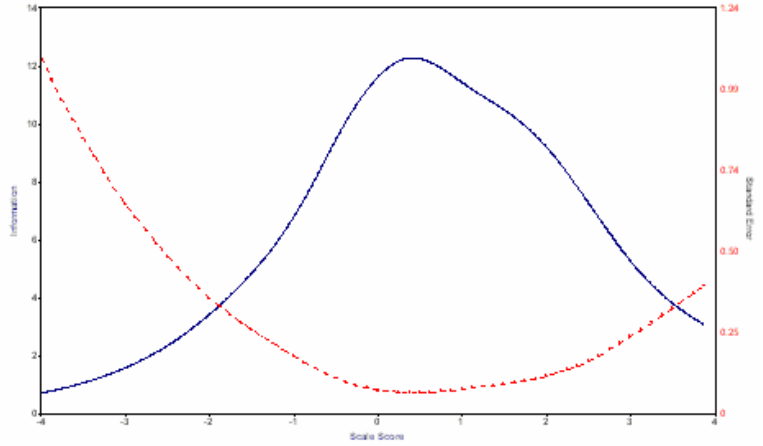
الاختبار	الإحصائي	الصعوبة	الخطأ المعياري لمعلمة الصعوبة	التمييز	الخطأ المعياري لمعلمة التمييز	التخمين	الخطأ المعياري لمعلمة التخمين	دالة المعلومات	الخطأ المعياري لدالة المعلومات
الأول	القيمة الصغرى	-17.125	0.038	0.075	0.021	0.024	0.010	0.0015	0.0006
	القيمة العظمى	12.020	5.130	2.091	0.260	0.500	0.096	2.6059	0.4579
	المتوسط الحسابي	-2.759	1.124	0.697	0.096	0.238	0.058	0.3768	0.0822
	الوسيط	-1.020	0.235	0.613	0.077	0.208	0.064	0.1738	0.0402
	الانحراف المعياري	6.573	1.770	0.513	0.065	0.143	0.027	0.5140	0.1049
الثاني	القيمة الصغرى	-16.204	0.053	0.080	0.022	0.038	0.007	0.0016	0.0007
	القيمة العظمى	5.818	4.636	1.569	0.398	0.500	0.092	1.2184	0.6923
	المتوسط الحسابي	0.046	0.409	0.761	0.129	0.136	0.044	0.3812	0.1118
	الوسيط	-0.157	0.200	0.743	0.113	0.123	0.040	0.3194	0.0852
	الانحراف المعياري	3.019	0.658	0.325	0.077	0.069	0.026	0.2980	0.1111

ويبين الشكل١ (أ، ب) العلاقة بين قيم دالة المعلومات للاختبارين والخطأ المعياري للتقديرات عند كل مستوى من مستويات القدرة، والذي يتضح فيه أن الاختبارين يقدمان معلومات عند القدرات المتوسطة التي عندها يكون الخطأ المعياري أقل ما يمكن مقابل أعلى دالة معلومات.



* المنحنى المتصل يمثل دالة المعلومات * المنحنى المتقطع يمثل الخطأ المعياري

الشكل ١-أ: دالة معلومات الاختبار والخطأ المعياري للتقدير لفقرات الاختبار الأول



* المنحنى المتصل يمثل دالة المعلومات * المنحنى المتقطع يمثل الخطأ المعياري

الشكل ١-ب: دالة معلومات الاختبار والخطأ المعياري للتقدير لفقرات الاختبار الثاني

وتم تدريج الاختبارين على مقياس واحد ليصبحا اختباراً واحداً، حيث تم التطبيق لصورتي الإختبار على مجموعتين مختلفتين من الأفراد (تصميم المجموعات غير المتكافئة مع التدوير) بوجود فقرات مشتركة بين الصورتين، وباستخدام برمجية Bilog-mg تم إعادة تدريج معالم الفقرات للاختبارين معاً على تدريج واحد، مكونا من ١٠٥ فقرات. ويبين الجدول ٢ الوسط الحسابي والوسيط والقيم الكبرى والصغرى والانحراف المعياري لمعالم الفقرات ودوال المعلومات للفقرات بعد المعايرة.

وبذلك تم تكوين بنك أسئلة من ١٠٥ فقرات معالمها معروفة ومفهرسة لاستخدامه في تطبيق الاختبارات المحوسبة التكيفية والخطية.

الجدول (٢): الإحصائيات الوصفية لمعالم الفقرات ودالة المعلومات بعد المعايرة

الإحصائي	الصعوبة	الخطأ المعياري لمعلمة الصعوبة	التمييز	الخطأ المعياري لمعلمة التمييز	التخمين	الخطأ المعياري لمعلمة التخمين	دالة المعلومات	الخطأ المعياري لدالة المعلومات
القيمة الصغرى	-19.283	0.023	0.023	0.005	0.032	0.007	0.00012	.0٠
القيمة العظمى	11.134	4.918	2.828	0.465	0.500	0.099	3.283	1.109
المتوسط الحسابي	-1.615	0.731	0.979	0.151	0.228	0.051	.60912	.1547
الوسيط	-0.258	0.189	0.760	0.133	0.206	0.044	0.269	0.072
الانحراف المعياري	4.880	1.263	0.666	0.108	0.130	0.029	.727913	.20088

إجراءات جمع بيانات الاختبارات التكيفية والخطية

تم إدخال فقرات بنك الأسئلة ومعالمها في برنامج Fasttest pro-2006، وبناءعلى المحدِّد الخاص بهذا البرنامج، فقد تم التعامل مع فئة مفتوحة من طرفي التوزيع لمعاملات الصعوبة، بحيث يتم التعامل مع جميع الفقرات بما في ذلك الفقرات التي تقع خارج المدى –٥، ٥ وعددها ٢١ فقرة، بناءً على مراسله عبر البريد الالكتروني (Email: dweiss@assess.com) مع مدير المؤسسة المعدة للبرنامج، الذي أشار بأن مثل هذه القيم المتطرفة تظهر في برنامج Bilog وبرنامج multilog (Wiess, 2004) وبتم التعامل معها على انها تحمل القيم القصوى من معاملات الصعوبة.

تم تجهيز المختبر لعقد الجلسات ألاختباريه باستخدام البرنامج، على أن يأخذ كل طالب اختبارين على جهاز الحاسب نفسه مع التدوير، أي ان يأخذ طالب الاختبار الأول ثم الثاني ويأخذ الطالب المجاور له الاختبار الثاني ثم الأول وهكذا. كما تم التجهيز لعقد خمس جلسات اختبارية تتناسب مع اسئلة الدراسة، كل جلسة باختبارين، واحتوت الجلسات المتعلقة بالسؤال الأول والثاني والخامس على اختبارين تكيفيين، أما السؤالان الثالث والرابع فاحتويا على اختبار تكيفي واختبار خطي عدد فقراته ٤٩ فقرة، والتي تشكل الفقرات المتبقية من المصفوفة المعيارية بعد حذف ١١ فقرة غير مطابقة للنموذج.

المعالجة الإحصائية

بعد الحصول على البيانات الخاصة لكل تطبيق على أفراد العينة، ولتحديد فاعلية الاختبار تم إجراء المعالجات الإحصائية التالية لكل سؤال من أسئلة الدراسة:

١–الخطأ المعياري في تقدير القدرة لكل مفحوص (SEE) Standard Error of Estimation (الاختبار الأكثر فعالية الذي يقدم أقل قيمة بالمتوسط من الخطأ المعياري لتقدير القدرة) ويتم الحصول علية من مخرجات برنامجFasttest Pro.

٢–الكفاءة المقاسة بعدد الفقرات المطبقة (أي أن الاختبار الأكثر فعالية هو الذي يستخدم أقل عدد من الفقرات في الوصول إلى قاعدة إنهاء الاختبار).

٣–الدلالة الإحصائية للفروق بين متوسطي كل اختبارين في كل تطبيق للمؤشرات الإحصائية المشار اليها في الخطوتين ١،٢.

٤–دالة المعلومات (أي أن الاختبار الأكثر فعالية هو الذي يقدم بالمتوسط أكبر قيمة من دالة المعلومات).

وقد تم حساب مؤشرات الدقة في القياس بمعادلات ثلاث مأخوذة من (Wang and Wang, 2002) وهي:

أ– الخطأ المعياري في القياس (SEM)

$$SE\ (\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \frac{\sum_{t=1}^{N}\hat{\theta}_t}{N}\right)^2}$$

ب– الجذر التربيعي لمتوسط مربعات الخطأ Root Mean Square Error (RMSE)، وكلما اقتربت قيمة هذا المؤشر من الصفر دل على دقة أعلى

$$RMSE\,(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\theta}_r - \theta)^2}$$

جـ– التحيز Bias، وكلما اقتربت قيمة هذا المؤشر من الصفر دل على دقة أعلى

$$Bias\,(\hat{\theta}) = \sum_{r=1}^{N}(\hat{\theta}_r - \theta)^2$$

حيث $\hat{\theta}_r$ القدرة المقدرة، θ القدرة الحقيقية المولدة باستخدام برنامج WinGen، الذي يقوم في مبدأ عمله على توليد القدرة الحقيقية لكل مفحوص بناءً على قيم معالم الفقرات.

النتائج والمناقشة

السؤال الأول: لفحص فاعلية الاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بعدد محدد من الفقرات (التطبيق الأول)، مقارنة بالاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري (التطبيق الثاني)، وفق طريقة الأرجحية العظمى لتقدير القدرة (MLE).

يبين الجدول (٣) نتائج اختبار t للفرق بين متوسطات تقديرات القدرة ومتوسطات الأخطاء المعيارية لتقديرات القدرة ومتوسطات أعداد الفقرات المطبقة في التطبيقين.

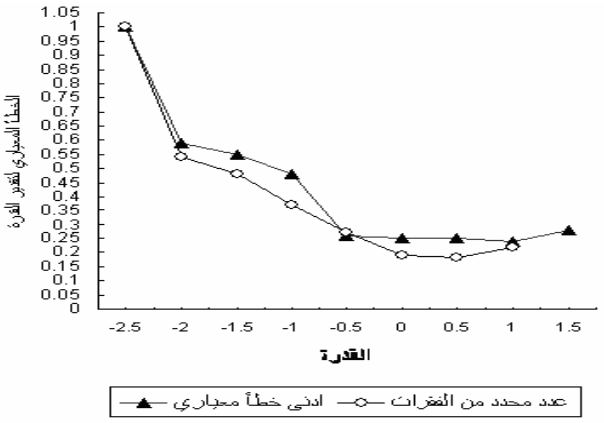
الجدول (٣): نتائج اختبار t للفرق بين متوسطات تقديرات القدرة والأخطاء المعيارية لتقديرات القدرة وأعداد الفقرات في التطبيقين تبعاً لقاعدتي الإنهاء باستخدام الارجحية العظمى

المؤشر	قاعدة الإنهاء	المتوسط الحسابي	الانحراف المعياري	ت	د ح	الدلالة
القدرة المقدرة	عدد محدد من الفقرات	0.22-	1.06	0.67	91	0.505
	أدنى خطأ معياري	0.29-	1.17			
الخطأ المعياري في التقدير	عدد محدد من الفقرات	0.28	0.21	0.639-	91	0.525
	أدنى خطأ معياري	0.29	0.17			
عدد الفقرات المطبقة	عدد محدد من الفقرات	30	0	16.163	91	*0.000
	أدنى خطأ معياري	١٦	8.36			

* تم اعتماد α=٠.٠٥ في هذه الدراسة. * دح: درجات الحرية

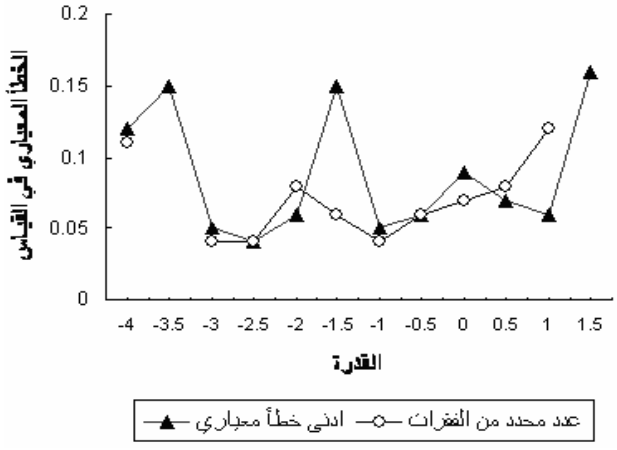
تظهر نتائج اختبار t بأن الفروق بين متوسطات القدرة المقدرة في التطبيقين غير دالة إحصائياً، وأنه لا يوجد فرق إحصائي بين متوسطات الأخطاء المعيارية في تقدير القدرة، ولكنها أظهرت وجود فرق ذي دلالة إحصائية بين متوسطات أعداد الفقرات المطبقة في التطبيقين لصالح التطبيق الذي ينتهي وفق قاعدة أدنى خطأ معياري. مما يشير إلى إمكانية تقدير القدرة بإستخدام ١٦ فقرة في طريقة تقدير الارجحية العظمى، وهذه النتيجة تتفق الى حد كبير مع النتائج التي حصل عليها كل من فيزبول (Vispoel, 1988)، وروسو وريكاس (Rosso and Reckase,1981) التي أشارت إلى أن الاختبارات التكيفية تكتفي بعدد من الفقرات يتراوح بين ١٢ إلى ١٤ فقرة. ومن الملاحظ أنَّ الزيادة في متوسط أعداد الفقرات المطبقة في قاعدة عدد محدد من الفقرات لم تحدث لزيادة في الدقة لتقدير القدرة، وهذا يبين ان الزيادة في اعداد الفقرات لا تزيد من دقة القياس وتقدير القدرة.

وفيما يتعلق بدالة المعلومات فقد تم حساب متوسط دوال المعلومات. ويتضح من هذه القيم أن الاختبارات التي تنتهي وفق قاعدة عدد محدد من الفقرات تقدم معلومات أكثر من الاختبارات التي تنتهي وفق قاعدة أدنى خطأ معياري، بمعنى أن زيادة عدد الفقرات أحدث فرقاً بدالة المعلومات، ولم يحدث فرقاً في تقدير القدرة والخطأ المعياري في تقديرها. ويبين الشكل (٢) توزيع قيم الأخطاء المعيارية لتقديرات القدرة.



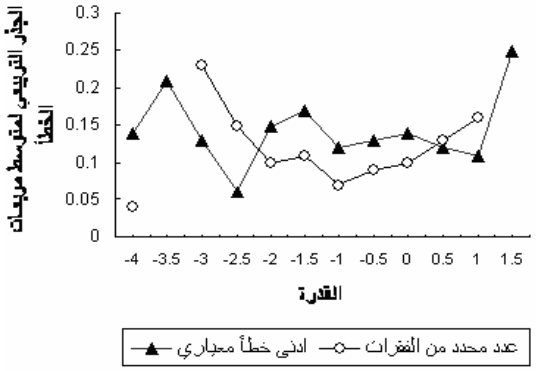
الشكل (٢): توزيع الأخطاء المعيارية لتقديرات القدرة تبعاً لقاعدتي إنهاء الاختبارات

حيث يلاحظ من الشكل أن دقة القياس تتناقص عند منخفضي القدرة لكلا التطبيقين، بينما تبرز فاعلية التطبيق الأول (الذي ينتهي وفق قاعدة عدد محدد من الفقرات) عند مستويات القدرة من –٢ إلى ١، حيث يقدم أخطاءً معيارية في التقدير أقل عند هذه المستويات من القدرة في التطبيق الثاني (الذي ينتهي وفق قاعدة أدنى خطأ معياري). كما وتتأكد هذه النتيجة من خلال الشكل (٣) الذي يشير إلى أن التطبيق الأول يقدم أخطاء معيارية في القياس أقل من التطبيق الثاني عند مستويات القدرة –٣، –١.٥، –١ وصفر.



الشكل (٣): توزيع قيم الأخطاء المعيارية في القياس تبعاً لقاعدتي إنهاء الاختبار

أما فيما يتعلق بالجذر التربيعي لمتوسط مربعات الخطأ (RMSE) فيبين الشكل (٤) أن التطبيق الأول يقدم RMSE بالمتوسط أقل عند مستويات القدرة من –٢ إلى ٠.٥ بينما هو أكبر عند مستويات القدرة –٢.٥ فاقل وعند مستوى القدرة ١، ويشير عدم الإستقرار او التوافق في المؤشرات الإحصائية المتعلقة بتقدير القدرة والخطأ المعياري في تقديرها الى أن أي زيادة في عدد الفقرات من بنك الأسئلة لا تقلل من اخطاء التقدير في القياس على متصل القدرة.



الشكل (٤): توزيع الجذر التربيعي لمتوسط مربعات الخطأ تبعاً لقاعدتي إنهاء الاختبار

السؤال الثاني: لفحص فاعلية الاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بعدد محدد من الفقرات (التطبيق الأول الذي ينتهي عند ٣٠ فقرة)، مقارنةً بالاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري (التطبيق الثاني والذي ينتهي عند أدنى خطأ معياري ٠.٢٥ أو أقل وبتطبيق ٣٠ فقرة على الأكثر) باستخدام طريقة التقدير البعدي الأعظم (MAP) لتقدير القدرة،

ويبين الجدول (٤) نتائج اختبار t للفرق بين متوسطات تقديرات القدرة والأخطاء المعيارية لتقديرات القدرة وأعداد الفقرات المطبقة.

© ٢٠١٣ عمادة البحث العلمي/ الجامعة الأردنية. جميع الحقوق محفوظة.

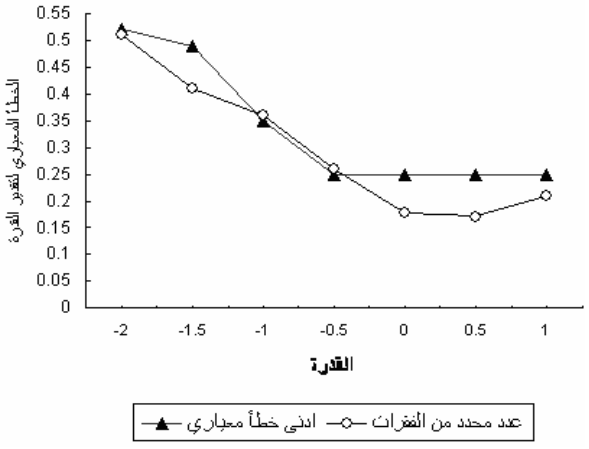
الجدول (٤): نتائج اختبار t للفرق بين متوسطات تقديرات القدرة والأخطاء المعيارية لتقديرات القدرة وأعداد الفقرات المطبقة للتطبيقين تبعا لقاعدتي الإنهاء باستخدام البعدي الأعظم

المؤشر	قاعدة الإنهاء	المتوسط الحسابي	الانحراف المعياري	ت	د ح	الدلالة
القدرة المقدرة	عدد محدد من الفقرات	0.21-	0.58	2.02-	152	*0.05
	أدنى خطأ معياري	0.14-	0.60			
الخطأ المعياري في التقدير	عدد محدد من الفقرات	0.23	0.09	8.38-	152	*0.00
	أدنى خطأ معياري	0.27	0.07			
عدد الفقرات المطبقة	عدد محدد من الفقرات	30		22.85	152	*0.00
	أدنى خطأ معياري	14.32	8.49			

* تم اعتماد α=٠.٠٥ في هذه الدراسة. * دح: درجات الحرية.

تظهر نتائج اختبار t للفرق بين متوسطات القدرة المقدرة في التطبيقين أنه يوجد فرق ذو دلالة إحصائية لصالح التطبيق الثاني، وأنه يوجد فرق ذو دلالة إحصائية بين متوسطات الأخطاء المعيارية في تقدير القدرة لصالح التطبيق الأول، كما وتظهر النتائج أنه يوجد فرق ذو دلالة إحصائية بين متوسطات أعداد الفقرات المطبقة في التطبيقين لصالح التطبيق الثاني. وهذه النتيجة تتفق مع الدراسات السابقة التي اشارت إلى أن الاختبارات التكيفية بحاجة إلى فقرات تتراوح ما بين ١٢- ١٤ فقرة في المتوسط. ويلاحظ من النتائج أنَّ نسبة متوسط أعداد الفقرات المستخدمة بطريقة أدنى خطأ معياري إلى متوسط أعداد الفقرات المستخدمة بطريقة عدد محدد من الفقرات ما يقارب ٥٠%. وأن الزيادة في متوسط أعداد الفقرات المطبقة في قاعدة عدد محدد من الفقرات لها اثر جوهري في زيادة دقة تقدير القدرة. بمعنى ان الزيادة في اعداد الفقرات لا توفر زيادة في دقة تقدير القدرة عندما تكون هناك فرصة للتحكم بقاعدة إنهاء الاختبارات التكيفية.

وفيما يتعلق بدالة المعلومات التي تقدمها جميع الاختبارات باعتماد قاعدة إنهاء الاختبار بعدد محدد من الفقرات، فقد كان ٣٣.٨٢، بينما بلغ ٢٥.٣٢ في حالة اعتماد قاعدة إنهاء الاختبار بأدنى خطأ معياري، وبنسبة زيادة حوالي ٢٥%. ويتضح من هذه القيم أن الاختبارات التي تنتهي عند تطبيق عدد محدد من الفقرات تقدم معلومات أكثر من الاختبارات التي تنتهي عند أدنى خطأ معياري باستخدام طريقة البعدي الأعظم لتقدير لقدرة. ويبين الشكل (٥) توزيع قيم الاخطاء المعيارية المستخلصة بالقاعدتين على متصل القدرة.



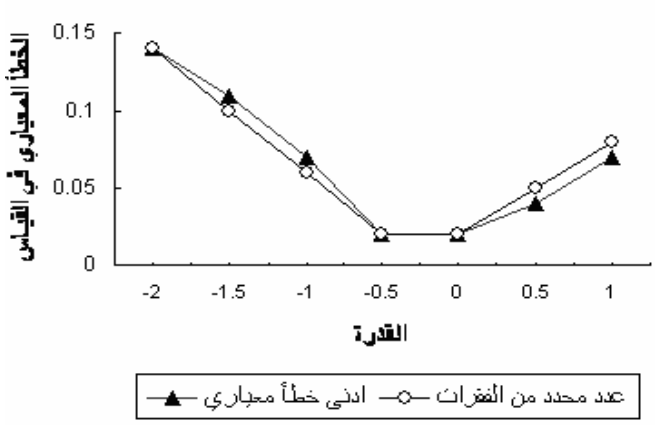
الشكل (٥): توزيع قيم الأخطاء المعيارية لتقديرات القدرة تبعاً لقاعدتي إنهاء الاختبارات

حيث يبين الشكل تقاطع خطي العلاقة بين القدرة والأخطاء المقدرة؛ لأن دقة القياس تتناقص عند منخفضي القدرة لكل من

© ٢٠١٣ عمادة البحث العلمي/ الجامعة الأردنية. جميع الحقوق محفوظة.

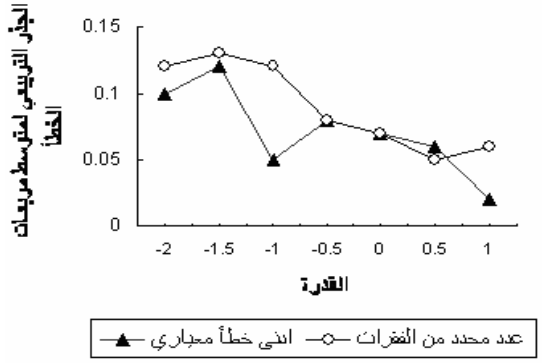
التطبيقين، وفاعلية اكبر للتطبيق الأول عند مستويات القدرة من -٠.٥ إلى ١ و-١ إلى -٢.

أما فيما يتعلق بالأخطاء المعيارية في القياس فيبين الشكل (٦) عدم التوافق الرتبي بين الأخطاء المعيارية في القياس وفق الطريقتين (التطبيقين) لإنهاء الإختبار على متصل القدرة. حيث أن التطبيق الأول يقدم أخطاء معيارية أقل بقليل عند مستويات القدرة الأقل من -٠.٥، كما أن التطبيقين لهما نفس الأخطاء المعيارية للقياس ما بين مستويي القدرة (-٠.٥- صفر) بينما يقدم التطبيق الأول أخطاء معيارية في القياس أعلى بقليل نسبياً من التطبيق الثاني عند مستويات القدرة من (صفر-١).



الشكل (٦): توزيع الأخطاء المعيارية في القياس تبعاً لقاعدتي إنهاء الاختبار

وفيما يتعلق بالجذر التربيعي لمتوسط مربعات الخطأ (RMSE) فيبين الشكل (٧) أن التطبيق الأول يقدم RMSE أعلى عند مستويات القدرة الأقل من -٠.٥، وأنهما يتساويان تقريبا عند مستويات قدرة تتراوح بين -٠.٥ و٠.٥. بمعنى ان هناك عدم انتظام، او عدم توافق كمي ورتبي في تقديرات الأخطاء بالطريقتين على متصل القدرة.



الشكل (٧): الجذر التربيعي لمتوسط مربعات الخطأ تبعاً لقاعدتي إنهاء الاختبارات

ومن نتائج السؤالين الأول والثاني يمكن الاستدلال أن قاعدة عدد محدد من الفقرات هي الخيار الأفضل باختلاف طرق تقدير القدرة (MLE, MAP)، لأنها توفر أقل أخطاء أقل في تقدير القدرة، وبدالة معلومات أكثر. أما إذا كان للوقت وأعداد الفقرات المطبقة أهمية، فإن طريقة أدنى خطأ معياري قد تكون الطريقة المفضلة على طريقة عدد محدد من الفقرات.

السؤال الثالث: المتعلق بفحص فاعلية الاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري مقارنة بالاختبار الخطي المحوسب باستخدام طريقة الارجحية العظمى MLE لتقدير القدرة، فقد تم استخدام تطبيقين للاختبارات المحوسبة، الأول تكيفي باستخدام قاعدة الإنهاء أدنى خطأ معياري وبخطأ ٠.٢٥ أو اقل وبتطبيق ٣٠ فقرة على الأكثر، والثاني خطي يتكون من ٤٩ فقرة، وتم تطبيقهما على العينة المكونة من ١٦٣ طالباً وطالبة، ويأخذ كل طالب اختبارين على نفس الجهاز، بحيث تأخذ نصف العينة الاختبار الخطي أولاً ثم التطبيق التكيفي والنصف الآخر أخذ التطبيق

التكيفي أولاً ثم الاختبار الخطي، وتم استخدام ٣٢٦ ملف، بواقع ملفين لكل طالب.

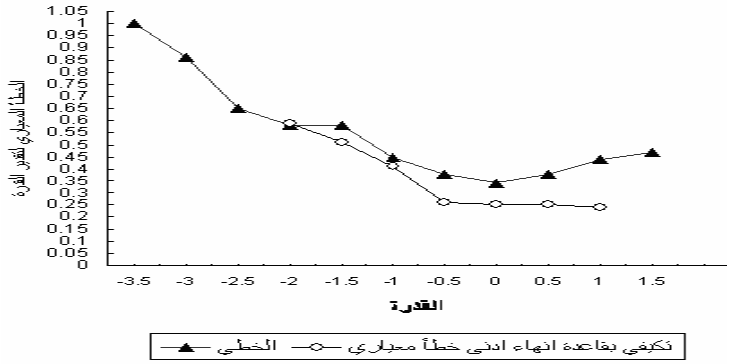
ويبين الجدول (٥) نتائج اختبار t للفرق بين متوسطات تقديرات القدرة ومتوسطات الأخطاء المعيارية لتقدير القدرة ومتوسطات أعداد الفقرات المطبقة في التطبيقين.

الجدول (٥): نتائج اختبار t للفرق بين متوسطات تقديرات القدرة والأخطاء المعيارية في التقدير وأعداد الفقرات المطبقة للتطبيقين تبعا لنمطي التطبيق باستخدام طريقة الارجحية العظمى لتقدير القدرة

المؤشر	الاختبار	المتوسط الحسابي	الانحراف المعياري	ت	د ح	الدلالة
القدرة المقدرة	الخطي	0.46-	1.04	5.59-	162	*0.00
	التكيفي	0.10-	0.66			
الخطأ المعياري في التقدير	الخطي	0.44	0.14	14.31	162	*0.00
	التكيفي	0.29	0.09			
عدد الفقرات المطبقة	الخطي	49	0	51.28	162	*0.00
	التكيفي	16.79	8.02			

* تم اعتماد α=٠.٠٥ في هذه الدراسة. * دح: درجات الحرية

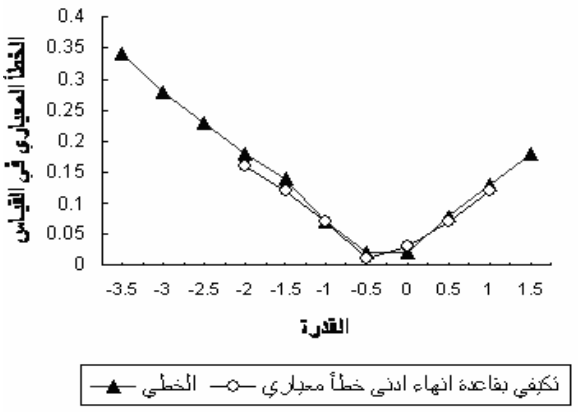
تظهر نتائج اختبار t أنه يوجد فرق ذو دلالة إحصائية بين متوسطات القدرة المقدرة لصالح التكيفي، وأنه يوجد فرق ذو دلالة إحصائية بين متوسطات قيم الأخطاء المعيارية لتقدير القدرة لصالح التكيفي، كما أظهرت أنه يوجد فرق ذو دلالة إحصائية بين متوسطات أعداد الفقرات المطبقة في التطبيقين لصالح التكيفي ايضاً. وهذا ما يشير إلى إمكانية تقدير القدرة باستخدام ١٧ فقرة تقريباً في الاختبارات التكيفية وفق طريقة الارجحية العظمى في التقدير. واتفقت هذه النتيجة مع نتائج دراسة فيزبول وآخرون (Vispoel, Wang and Bleier,1997) حيث أشارت نتائج هذه الدراسة إلى أن الاختبار التكيفي يحتاج إلى فقرات تقل بنسبة (٥٠% - ٩٣%) ليطابق ثبات الاختبارات الخطية. كما واتفقت مع دراسة الخضر وكلارك (Alkhader and Klarke, 1998) حيث أشارت إلى أنه يمكن تقدير السمة بعدد فقرات يقل بـ ٥٠% من عدد الفقرات الأصلي. ومما تجدر الإشارة إليه أن نسبة متوسط أعداد الفقرات المطبقة في الاختبارات التكيفية إلى عدد الفقرات المطبقة في الاختبار الخطي ٣٤%، وان الزيادة في أعداد الفقرات المطبقة في الاختبار الخطي لم تحدث أثراً لزيادة الدقة في تقدير القدرة. وقد كان متوسط دالة المعلومات ١٤.٧٦ للاختبار الخطي، وبمتوسط ٢٤.٤٤ للتطبيق التكيفي، ويتضح من هذه القيم أن التطبيق التكيفي يقدم معلومات أكثر من الاختبار الخطي باستخدام طريقة الارجحية العظمى لتقدير القدرة MLE؛ لأنه يتم اختيار الفقرات المناسبة للقدرة، حيث يتضح ذلك من قيم الاخطاء المعيارية لتقديرات القدرة في الشكل(٨). والذي يبين أن دقة القياس تتناقص في الاختبار الخطي عند منخفضي القدرة، كما يتضح أن التطبيق التكيفي يقدم أخطاء معيارية أقل من الاختبار الخطي عند جميع مستويات القدرة.



الشكل (٨): توزيع قيم الخطأ المعياري لتقدير القدرة تبعاً لنمط التطبيق

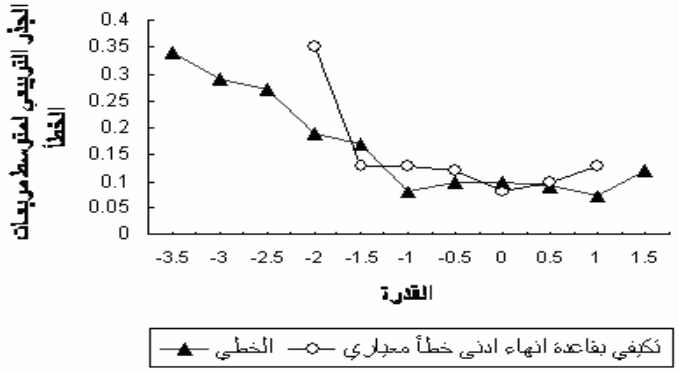
ومن توزيع الأخطاء المعيارية في الشكل(٩) يتبين أن التطبيق التكيفي يقدم أخطاء معيارية أقل من الاختبار الخطي،

لأنه يتم اختيار الفقرات التي تقدم اعلى معلومات عند مستوى القدرة المقدر في كل مرة.



الشكل (٩): توزيع الأخطاء المعيارية في القياس تبعاً لنمط التطبيق

أما فيما يتعلق بمؤشر الجذر التربيعي لمتوسط مربعات الخطأ RMSE فيبين الشكل (١٠) أن التطبيق الخطي والتكيفي يتعادلان نسبيا في RMSE عند معظم النقاط المشتركة على متصل القدرة.



الشكل (١٠): الجذر التربيعي لمتوسط مربعات الخطأ تبعاً لنمط التطبيق

السؤال الرابع: ما فاعلية الاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري مقارنة بالاختبار الخطي المحوسب باستخدام طريقة التقدير البعدي الأعظم (MAP) لتقدير القدرة؟ وتطلبت الإجابة عن هذا السؤال تصميم التجربة على النحو الآتي: اجراء تطبيقين للاختبارات المحوسبة الأول خطي يتكون من ٤٩ فقرة، والثاني تكيفي باستخدام قاعدة الإنهاء أدنى خطأ معياري وبخطأ ٠.٢٥ أو اقل وبتطبيق ٣٠ فقرة على الأكثر. وتم تطبيقهما على العينة المكونة من ١٤٩ طالباً وطالبة، بحيث يأخذ كل طالب اختبارين على نفس الجهاز، بحيث تأخذ نصف العينة الاختبار الخطي ثم التطبيق التكيفي والنصف الأخر يأخذ التطبيق التكيفي ثم الاختبار الخطي، وبذلك تم استخدام ٢٩٨ ملف، بواقع ملفين لكل طالب. ويبين الجدول (٦) نتائج اختبار t للفرق بين متوسطات تقديرات القدرة ومتوسطات الأخطاء المعيارية لتقدير القدرة ومتوسطات أعداد الفقرات المطبقة في التطبيقين.

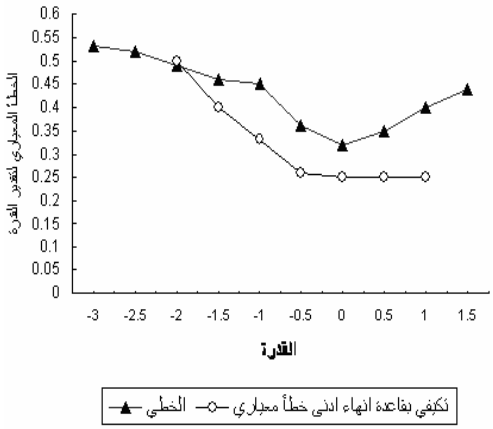
الجدول (٦): نتائج اختبار t للفرق بين متوسطات تقديرات القدرة والأخطاء المعيارية في التقدير وأعداد الفقرات المطبقة في

للتطبيقين تبعا لنمطي التطبيق باستخدام طريقة البعدي الاعظم لتقدير القدرة

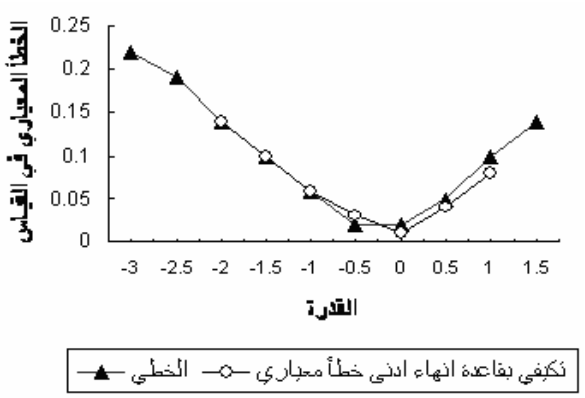
المؤشر	الاختبار	المتوسط الحسابي	الانحراف المعياري	ت	د ح	الدلالة
القدرة المقدرة	الخطي	0.43-	0.98	4.81-	148	*0.00
	التكيفي	0.14-	0.60			
الخطأ المعياري في التقدير	الخطي	0.39	0.08	17.43	148	*0.00
	التكيفي	0.27	0.06			
عدد الفقرات المطبقة	الخطي	49	0	48.26	148	*0.00
	التكيفي	14.59	8.70			

* تم اعتماد α=٠.٠٥ في هذه الدراسة. * دح: درجات الحرية

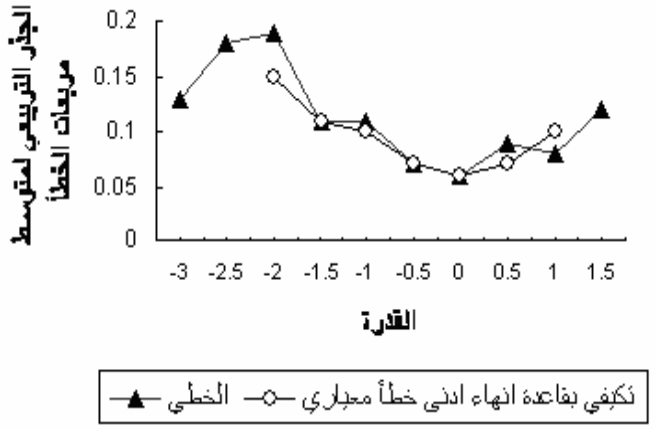
وتظهر نتائج اختبار t أنه يوجد فرق ذو دلالة إحصائية في متوسطات القدرة المقدرة، وكذلك في متوسطات الأخطاء المعيارية في تقدير القدرة، وبمتوسطات أعداد الفقرات المطبقة في التطبيقين لصالح التطبيق التكيفي. وهذا يدعم ماتم استخلاصه في التصاميم السابقة من أن الزيادة في أعداد الفقرات المطبقة في الاختبار الخطي لم تحدث أثراً لزيادة الدقة في تقدير القدرة، وبنفس التفسير القائم على ان طريقة اختيار الفقرات في الاختبارات التكيفية المحوسبة تتناسب مع كل مفحوص حسب قدرته. وينعكس ذلك على دالة المعلومات حيث بلغت ١٤.٧٦ للاختبار الخطي، بينما بلغت في المتوسط ٢٠.٦٩ للتطبيق التكيفي؛ مما يعني أن التطبيق التكيفي له متوسط دالة معلومات أعلى من الاختبار الخطي باستخدام طريقة تقدير البعدي الأعظم لتقدير القدرة MAP. ويمكن أن يعزى ذلك إلى اختلاف خصائص الفقرات التي تم تطبيقها في الاختبارات التكيفية، حيث تم اعتماد طريقة أقصى معلومات لاختيار الفقرات والتي تقدم للمفحوص الفقرات ذات المعلومات الأقصى عند مستوى القدرة المقدر بشكل تتابعي. كما يبين الشكل (١١) أن التطبيق التكيفي يقدم أخطاءً معيارية في تقدير القدرة أقل من الاختبار الخطي على متصل القدرة. بينما يبين الشكل (١٢) الذي يعرض توزيع الأخطاء المعيارية في القياس، أن لكل من الاختبارين أخطاء معيارية متساوية عند القدرات المنخفضة. كما يبين الشكل (١٣) أن كل من التطبيقين يقدمان قيما متقاربة للجذر التربيعي لمتوسط مربعات الخطأ RMSE عند معظم النقاط على متصل القدرة.



الشكل (١١): توزيع قيم الأخطاء المعيارية لتقدير القدرة تبعاً لنمط التطبيق



الشكل (١٢): توزيع الأخطاء المعيارية في القياس تبعاً لنمط التطبيق



الشكل (١٣): توزيع الجذر التربيعي لمتوسط مربعات الخطأ تبعاً لنط التطبيق

ومن نتائج السؤالين الثالث والرابع يمكن الاستدلال أن الاختبار التكيفي يقدم تقديرات للقدرة أفضل من الاختبار الخطي وبعدد أقل من الفقرات باختلاف طريقتي تقدير القدرة (MLE, MAP)، ومن الجدير بالملاحظة أن نسبة متوسط عدد فقرات الاختبارات التكيفية إلى متوسط عدد فقرات الاختبار الخطي باختلاف طريقتي تقدير القدرة (MLE, MAP) ٠.٣٠، ٠.٣٤ على الترتيب، وتدل هذه النسبة إلى أن الاختبار التكيفي يحتاج إلى ثلث عدد فقرات الاختبار الخطي تقريباً، ويمكن الاستدلال من ذلك إلى إمكانية تحديد عدد الفقرات اللازمة للوصول إلى الدقة المستهدفة في تقدير القدرة المقدرة. فلذلك يمكن القول أن ثلث عدد الفقرات المستخدمة في الاختبار الخطي تكفي لإعطاء تقديرات قدرة دقيقة فيما لو استخدمت في اختبار تكيفي مما يوفر الوقت والجهد والمحافظة على أمن فقرات الاختبار.

السؤال الخامس: لفحص فاعلية الاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري باستخدام طريقة الأرجحية العظمى MLE لتقدير القدرة مقارنة بالاختبار التكيفي المحوسب باستخدام قاعدة إنهاء الاختبار بأدنى خطأ معياري باستخدام طريقة التقدير البعدي الأعظم MAP لتقدير القدرة، فقد تم تطبيق الاختبارات التكيفية المحوسبة باستخدام قاعدة الإنهاء أدنى خطأ معياري وبخطأ ٠.٢٥ أو اقل وبتطبيق ٣٠ فقرة على الأكثر. بحيث تُقدّر القدرة في التطبيق الأول بطريقة الأرجحية العظمى MLE، بينما تُقدّر القدرة في التطبيق الثاني بطريقة التقدير البعدي الأعظم MAP، وطبق كل من التطبيقين على أفراد العينة المكونة من ٨١ طالباً وطالبة، بحيث يأخذ كل طالب اختبارين على نفس الجهاز، بحيث تأخذ نصف العينة التطبيق الذي يتم تقدير القدرة فيه باستخدام MLE (التطبيق الأول) ثم التطبيق الذي يتم تقدير القدرة فيه باستخدام MAP (التطبيق الثاني)، وأخذ النصف الآخر التطبيق الثاني ثم التطبيق الأول، وبذلك تم استخدام (١٦٢) ملف، بواقع ملفين لكل طالب. ويبين الجدول (٧) نتائج اختبار t للفرق بين متوسطات تقديرات القدرة ومتوسطات الأخطاء المعيارية لتقديرات القدرة، ومتوسطات أعداد الفقرات المطبقة.

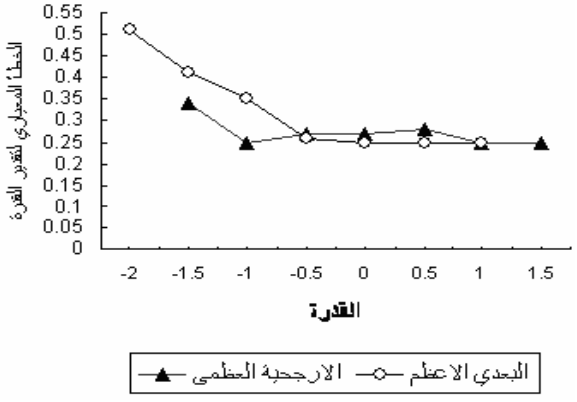
الجدول (٧): نتائج اختبار t للفرق بين متوسطات تقديرات القدرة ومتوسطات الأخطاء المعيارية لتقديرات القدرة ومتوسطات أعداد الفقرات المطبقة في للتطبيقين تبعا لطريقتي تقدير القدرة

المؤشر	الطريقة	المتوسط الحسابي	الانحراف المعياري	ت	د ح	الدلالة
القدرة المقدرة	MLE	0.03	0.70	1.44	80	0.153
	MAP	0.12-	0.60			
الخطأ المعياري في التقدير	MLE	0.27	0.07	1.19-	80	0.852
	MAP	0.27	0.06			
عدد الفقرات المطبقة	MLE	16.32	7.49	1.48	80	0.144
	MAP	14.47	8.63			

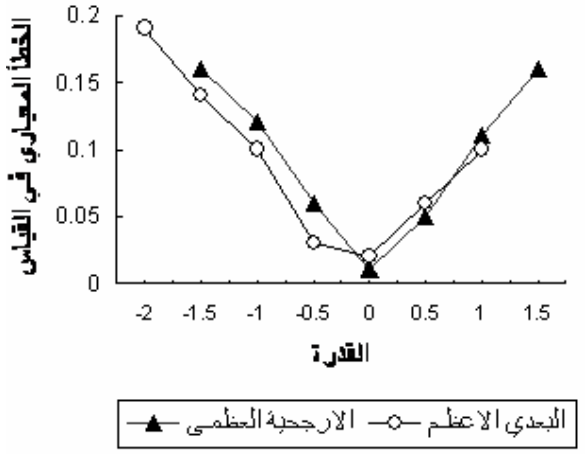
وتظهر نتائج اختبار t للفرق بين متوسطات القدرات المقدرة في التطبيقين أنه لا يوجد فرق ذو دلالة إحصائية بين متوسطات القدرة المقدرة، وأنه لا يوجد فرق ذو دلالة إحصائية بين متوسطات الأخطاء المعيارية في التقدير، كما وتظهر أيضاً أنه لا يوجد فرق ذو دلالة إحصائية بين متوسطات أعداد الفقرات المطبقة. وهذا ما يشير إلى إمكانية تقدير القدرة باستخدام عدد فقرات يتراوح بين ١٤-١٧ فقرة تقريباً في الاختبارات التكيفية باستخدام طريقتي التقدير الارجحية العظمى MLE والبعدي الاعظم MAP لتقدير القدرة، واختلفت هذه النتيجة مع نتائج دراسة وانق ووانق (Wang and Wang, 2001) حيث أشارت النتائج إلى أن طريقة MAP لهما مزايا على طريقة MLE من حيث كفاءة الاختبار (عدد الفقرات المطبقة).

وفيما يتعلق بدالة المعلومات فقد تم حساب متوسط دوال المعلومات التي تقدمها جميع الاختبارات حيث بلغت ٢٤.٢٥ للتطبيق الأول، و١٨.١٩ للتطبيق الثاني، ويتضح من هذه القيم أن الاختبارات التكيفية التي يتم فيها تقدير القدرة بطريقة MLE تقدم معلومات أكثر من الاختبارات التكيفية التي يتم فيها تقدير القدرة بطريقة MAP باستخدام قاعدة أدنى خطأ معياري لإنهاء الاختبار. واتفقت هذه النتيجة مع نتائج دراسة روسو وريكاس (Rosso and Reckase,1981)، حيث أشارت إلى أن إجراء الاختبار المفصل بالأرجحية العظمى أعطى معلومات اختبار كلية أعلى من إجراء بييز للاختبار المفصل.

أما بالنسبة لقيم الاخطاء المعيارية لتقديرات القدرة، فيبين الشكل(١٤) أن دقة القياس تتناقص عند منخفضي القدرة لكل من التطبيقين، إلا أنها أفضل في التطبيق الأول باستخدام طريقة الأرجحية العظمى لتقدير القدرة، كما يبين أن كل من التطبيقين لهما أخطاء معيارية في التقدير متساوية تقريباً عند مستويات القدرة من ٠.٥ الى١. ويبين الشكل (١٥) توزيع الأخطاء المعيارية في القياس أن التطبيق الأول له أخطاء معيارية أعلى من التطبيق الثاني عند مستويات القدرة الواقعة دون الوسط، كما أن كلاً من التطبيقين لهما أخطاء معيارية متساوية تقريباً عند المستوى المتوسط في القدرة كحد أدنى.

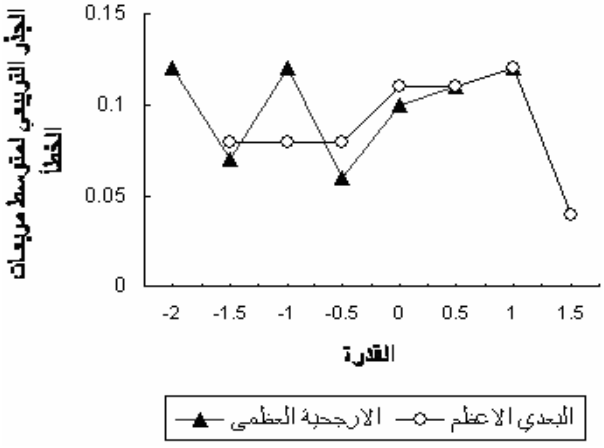


الشكل (١٤): توزيع قيم الأخطاء المعيارية لتقديرات القدرة تبعاً لطريقة تقدير القدرة



الشكل (١٥): توزيع الأخطاء المعيارية في القياس تبعاً لطريقة تقدير القدرة

وفيما يتعلق بالجذر التربيعي لمتوسط مربعات الخطأ RMSE فيبين الشكل (١٦) اختلافاً جذرياً في شكل التوزيع عند مستويات القدرة المختلفة، الا ان قيم هذه المتوسطات في التكيفي اكثر استقرارا من نظيراتها في الخطي. وتقل قيمة هذه المتوسطات عند تقدير القدرات العالية نسبيا.



الشكل (١٦): توزيع الجذر التربيعي لمتوسط مربعات الخطأ تبعاً لطريقة تقدير القدرة

ويظهر من مجمل النتائج أن قاعدة إنهاء الاختبار بعدد محدد من الفقرات توفر تقديرات أدق للقدرة، ودالة معلومات أعلى من قاعدة أدنى خطأ معياري باختلاف طريقتي تقدير القدرة. كما توفر قاعدة أدنى خطأ معياري في عدد الفقرات المطبقة بنسبة تصل إلى ٥٠% من قاعدة عدد محدد من الفقرات. كما أن الاختبار التكيفي يوفر تقديرات أدق للقدرة، ويوفر في عدد الفقرات المطبقة بنسبة تصل إلى ٧٠%. وله دالة معلومات أعلى من الاختبار الخطي باختلاف طريقتي تقدير القدرة. وتعطي كل من طريقتي تقدير القدرة (MLE, MAP) تقديرات قدرة متساوية، ومؤشرات دقة متساوية، وأن لطريقة MLE دالة معلومات للاختبار أعلى من MAP. ومن ابرز التوصيات المنطلقة من هذه النتائج الإشارة الى جدوى استخدام الاختبارات التكيفية المحوسبة في الاختبارات العامة كاختبارات القبول واختبارات القدرات العقلية لفاعليتها من حيث الدقة، وفعاليتها الاقتصادية من حيث زمن التطبيق، وتقليل الضغوطات على المفحوصين. وكذلك جدوى استخدام طريقة

الأرجحية العظمى في تقدير القدرة بالاختبارات التكيفية لأن المؤشرات تعطي أفضلية لهذه الطريقة مقارنة بطريقة البعدي الأعظم. كما توصي بإجراء دراسات مقارنة بين طريقتي التقدير MLE,MAP عند اختلاف عدد الفقرات الأصلية في بنك الأسئلة، وعند اختلاف قيم أدنى خطأ معياري غير التي استخدمت في الدراسة.

المراجع

Alkhader, O., Clark, D. and Anderson, N. 1998. Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the differential aptitude tests. *Journal of Occupational and Organizational Psychology,* 713, 205-218.

Chen, S., Hou, L. and Dodd, B. 1998. A Comparison of maximum likelihood estimation expected a posterior estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*, 569-595.

Cisar, D., Radosav, D., Markoski, B., Pinter, R. and Cisar, P. 2010. Computer adaptive testing of Student knowledge. *Acta Polytechnic Hungarica,* 7(4), 139-152.

Clarke, D. Mackinnon, F. Mckenzie, F. and Herrman, H. 2000. Dimensions of psychopathology in the medically ill: a latent trait analysis. *Psychosomatics,* 41(5), 420.

Dodd, B., Koch, W. and De Ayala, R. 1989. Operational characteristics of adaptive testing procedures using the graded response model, *Applied Psychological Measurement,* 13, 129-143.

Embretson, S. and Reiase, S. 2000. *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Fliege, H., Becker, J., Walter, O., Bjorner, J., Klapp, B. and Rose, M. 2005. Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research,* 14, 2277-2291.

Huo, Y. 2009. *Variable-length Computerized Adaptive Testing: adaptation of the a-stratified strategy in item selection with content balancing*. Dissertation of doctor of philosophy in psychology in the graduate college of the University of Illinois at Urbana-Champaign.

Jacobusse, G. and Buuren, S. 2007. Computerized adaptive testing for measuring development of young children. *Statistics in Medicine,* 26, 2630-2638.

Jain-quan, T., Dan-min, M., Xia, Z. and Jing-jing, G. 2007. An Introduction to the computerized adaptive testing. *US-china Education Review*, 4(1): 72-81.

Kingsbury, G. and Zara, A. 1989. Procedures for selecting item for computerized adaptive testing. *Applied Psychological Measurement,* 2, 359-375.

Linden, W. and Pushley, P. 2003. Item Selection and ability estimation in adaptive testing. In Linden, W. and Glass, C. (eds). *Computerized Adaptive Testing Theory and Practice*. Kluwer Academic Publishers.

Magis, D. and Raiche, G. 2011. CatR: An R package for computerized adaptive testing, *Applied Psychological Measurement*, 33(7): 576-577.

McKinley, R. and Reckase, M. 1981. *A Comparison of a Bayesian and a maximum likelihood tailored testing procedure*. Office of Naval Research, Arlington, VA. Personnel and Training Research Programs Office.

Millman, J. and Arter, J. 1984. Issues in item banking, *Journal of Educational Measurement*, 21, 315-330.

Murphy, K. and Davidshofer, C. O. 1994. *Psychological testing: Principles and applications.* 3ed. New Jersey: Prentice-Hall.

Roex, A. and Degryse, J. 2004. A Computerized adaptive knowledge test as an assessment tool in general practice: a pilot study, *Medical Teacher*, 26(2): 178-183.

Rosso, M. and Reckase, M. 1981. *A comparison of a maximum likelihood and a Bayesian ability estimation procedure for tailored testing*. Paper presented at the annual meeting of the national council on measurement in education, Missouri Univ., Columbia.

Stone, E. and Davey, T. 2011. *Computer-Adaptive Testing for Students with Disabilities: A Review of the Literature*. ETS, Princeton, New Jersey, Retrieved at online: http:// www. ets. org/ research/ contact. html.

Vispoel, W. 1988. An adaptive test of musical memory: An application of item response theory to the assessment of musical ability. *Dissertation Abstracts*, DAI-A 49101, P 79.

Vispoel, W. 1993. Computerized adaptive and fixed-item testing versions of the ITED vocabulary subtest. *Educational and Psychological Measurement,* 53, 779-788.

Vispoel, W., Wang, T. and Bleiler, T. 1997. Computerized adaptive and fixed-item testing of music listening skill: A

computerized adaptive testing (item response). *Dissertation Abstracts, DAI-A 56/ 06, P2212.*

Wang, T. and Vispoel, W. 1998. Properties of ability estimation methods and computerized adaptive testing, *Journal of Educational Measurement,* 35 (2): 109- 135.

Ward, W. 1984. Using microcomputers to administer measurement. *Issues and Practices*, 3, 16-20.

Weiss, D. 2004. Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37, 70-84.

comparison of efficiency, precision and concurrent validity. *Journal of Educational Measurement*, 34, 34-63.

Wang, S. and Wang, T. 2001. Precision of weighted likelihood estimates for a polytomous model in computerized adaptive testing, *Applied Psychological Measurement*, 25(4): 317-331.

Wang, S. and Wang, T. 2002. *Relative precision of ability estimation in polytomous CAT: A Comparison under the generalized partial credit model and graded response model*. ED 477926, shudon wang, 19500 Bulverde Road, San Antonio, TX 78259-3701.

Wang, T. 1995. The precision of ability estimation methods in

The Effectiveness of Computerized Adaptive Testing in Estimating Mental Ability Using Raven's Matrices

Ahmad S. Odeh, and Omar S. Obaidat*

ABSTRACT

The purpose of this study was to investigate the effectiveness of computerized adaptive testing on the accuracy of estimation of mental ability using Raven matrices based on two estimation methods (MLE versus MAP) and based on the test completion rule (limited number of items versus least standard error) . To achieve this purpose, item bank was built consisting of 105 items drawn from these matrices. Five computerized tests were used in the study, applied on a sample consisting of 638 students. Results of the study indicated that the rule of ending test with a limited number of items provide more accurate mental ability estimations, and more accurate information function than least standard error rule. Least standard error rule is 50% more accurate than ending tests with a limited number of items. Computerized adaptive testing provides more accurate mental ability estimations, and reduces the number of administered items by 70%. Furthermore, it has higher information function than linear testing depending on the used one of the two methods of estimation. Both MLE and MAP give equal ability estimations and accuracy indices, but MLE has a higher information function than MAP. Based on the results of the study it was recommended to: use adaptive testing for different practical purposes and to use maximum likelihood method in estimating ability based on adaptive testing, more studies were recommended to compare the two estimation methods with different numbers of items in the bank and with different lowest level standard error.

Keywords: Computerized Adaptive Testing, Raven Matrices, Item Response Theory, Item Bank, Adaptive Testing Completion Rule, Ability Estimation Accuracy.

* Faculty of Education, Yarmouk University, Irbid, Jordan. Received on 4/5/2009 and Accepted for Publication on 9/5/2013.