المجلة الأردنية في العلوم التربوية، مجلد 5، عدد 1، 2009، 71- 87

أثر طريقة تصحيح فقرات الصواب والخطأ المتعدد في دقة تقدير قدرات الأفراد وصعوبة الفقرات باستخدام نموذج التقدير الجزئي

مروان الصمادي* ويوسف السوالمة**

تاريخ تسلم البحث 2007/3/19 تاريخ قبوله 2009/3/23

ملخص: هدفت هذه الدراسة إلى الكشف عن أثر طريقة تصحيح فقرات الصواب والخطأ المتعدد في دقة تقدير قدرات الأفراد وصعوبة الفقرات باستخدام نموذج التقدير الجزئي. ولتحقيق ذلك؛ تم بناء اختبار تحصيلي يتكون من 35 فقرة صواب وخطأ متعدد في وحدة الإحصاء المقررة لطلبة الصف الأول الثانوي في الفرعين الأدبي والإدارة المعلوماتية للعام الدراسي 2005/2004 في الأردن. وطبق الاختبار على 873 طالباً وطالبة بعد دراستهم للوحدة في عشر مدارس اختيرت عشوائياً من مدارس مديرية تربية اربد الأولى. وقد صححت أوراق الإجابة باستخدام أربع طرق للتصحيح متدرجة في مراعاتها للمعرفة الجزئية. وباستخدام برمجية (BigSteps) تم الحصول على إحصاءات المطابقة وتقديرات لقدرات الأفراد و صعوبة الفقرات والأخطاء المعيارية في تقديرها بحسب طريقة التصحيح بعد استبعاد الأفراد والفقرات غير المطابقين لنموذج التقدير الجزئي. أظهرت النتائج أن طريقة التصحيح الرابعة _الأكثر مراعاة للمعرفة الجزئية_ أكثر دقة في تقدير قدرات الأفراد و صعوبة الفقرات، تلتها طريقة التصحيح الثالثة، ثم الثانية، فالأولى. (**الكلمات المفتاحية:** طرق التصحيح، فقرات الصواب والخطأ المتعدد، النظرية الحديثة للقياس، نموذج التقدير الجزئي).

The Effect of Multiple True-False Test Scoring Method on the Accuracy of Estimating Ability and Item Difficulty Parameters Using Partial Credit Model

Marwan Al Smadi, *Faculty of Education, Irbid private Universityy, Irbid, Jordan.*
Yousef Al Sawalmeh, *Faculty of Education, Yarmouk University, Irbid, Jordan*

***Abstract**: The purpose of the study was to examine the effect of the multiple true-false test scoring method on the accuracy of estimating ability and difficulty parameters using Partial Credit Model. To achieve this goal, a multiple true- false mathematics test composed of (35) multiple true-false items was developed and administered to (873) male and female 11th graders of the literary stream and the information technology stream in Irbid First Directorate of Education. The test was scored by using four different methods that varied in the degree of considering partial knowledge . By using BIGSTEPS , estimates of individuals' abilities, items difficulty values, standard errors ,and person and item fit statistics were obtained for each scoring method. Data was re-analyzed after excluding misfitting subjects and items. The findings of the study indicated that the fourth method in which partial knowledge was most considered is the most accurate method for estimating ability and item difficulty parameters.*
***(Keywords** :Scoring Method, Multiple True-False Tests, Modern Test Theory, Partial Credit Model).*

المقدمة:

تعد الاختبارات من أدوات القياس التربوي المهمة فهي متعددة الأغراض والأهداف، وتؤدي إلى تحسين عملية اتخاذ القرارات التدريسية، وتحسّن دافعية التلاميذ، وتزيد مستوى الاحتفاظ بالمعلومات، وانتقال أثر التعلم، وتوفر التغذية الراجعة عن فعالية العملية التدريسية. لذلك ينبغي أن تتسق الاختبارات التحصيلية مع الأهداف التدريسية، وأن تمثل المادة الدراسية ، وأن تتمتع بأكبر درجة من الثبات، وأن تصاغ بطريقة تخدم الغرض من خلال انتقاء النوع المناسب من الأسئلة. من هنا فإن قياس التحصيل يتأثر بشكل فعال باختيار نوعية الأسئلة التي من شأنها أن تستثير الإجابة المطلوبة وتستبعد الإجابات الأخرى التي ليس لها صلة بالإجابات الصحيحة (عدس، 1988).

وللاختبارات التحصيلية أشكال متعددة، منها فئة الاختبارات الكتابية، وفئة الاختبارات غير الكتابية(ثورندايك وهيجن،1986)، وتتضمن الاختبارات الكتابية، الاختبارات ذات الإجابة القصيرة (المقيدة) والاختبارات ذات الإجابة الحرة الطويلة (المفتوحة) واختبارات المزاوجة Matching، واختبارات الصواب والخطأ (ص خ) True False، واختبارات الاختيار من متعدد (خ م) Multiple Choice، واختبارات الصواب والخطأ المتعدد (ص خ م) Multiple True False.

ويتكون اختبار الصواب والخطأ المتعدد-موضوع الدراسة الحالية- من مجموعة من الفقرات التي تشبه فقرات الاختيار من متعدد من حيث الشكل، ولكن بدلاً من اختيار إجابة واحدة صحيحة من بين عدة بدائل معطاة، يقوم المفحوص بوضع إشارة (صح) أو (خطأ) أمام كل بديل من البدائل في الفقرة الواحدة، ويمكن ألا يكون هناك أي بديل صحيح أو أن تكون جميع البدائل صحيحة، أو أن بعض البدائل صحيحة وبعضها خطأ ضمن الفقرة الواحدة

* كلية العلوم التربوية، جامعة إربد الأهلية، إربد، الأردن
** كلية التربية، جامعة اليرموك، إربد، الأردن.
© حقوق الطبع محفوظة لجامعة اليرموك، 2009، إربد، الأردن.

(Ebel and Frisbie, 1986). وبذلك فهو أسلوب اختباري يجمع بين خصائص فقرات الاختيار من متعدد وخصائص الصواب والخطأ وخصائص أخرى لا تتوفر في كليهما، ويمكن أن تعالج الفقرة فيه فكرة واحدة أو عدة أفكار مترابطة (Frisbie and Sweeny, 1982).

وتمتاز أسئلة الصواب والخطأ المتعدد بأنها سهلة الإعداد، وتقلل من درجة غموض الفقرة وفرص الحصول على علامات مرتفعة بالتخمين مما يؤثر إيجاباً على صدق الاختبار وثباته، وهناك إمكانية للتنويع في عدد البدائل الصحيحة في الفقرة، مما يمكّن من التغلب على بعض عيوب فقرات الاختيار من متعدد (Ebel and Frisbie, 1986). كما يختلف التوزيع الاحتمالي للإجابات العشوائية على فقرة صواب وخطأ متعدد تتكون من أربعة بدائل، (مصححة بطريقة MTF)، عن التوزيع الاحتمالي لفقرة اختيار من متعدد بنفس العدد من البدائل (سوا لمة، 1994).ويلخص الجدول (1) بعض الفروق بين فقرات الاختيار من متعدد وفقرات الصواب والخطأ المتعدد التي تتكون من أربعة بدائل (مصححة بطريقة MTF)

الجدول (1): الفروق بين فقرات (ص خ م) وفقرات (خ م) لكل منها أربعة بدائل من حيث العلامات الممكنة والمتوقعة واحتمال الحصول على العلامة الكاملة

	فقرات خ م	فقرات ص خ م
العلامة الممكنة	0 ، 1	0 ، 0.25، 0.50، 0.75، 1
علامة الصدفة المتوقعة	0.25	0.50
احتمال* العلامة الكاملة (1)	0.25	0.0625

* بازدياد عدد فقرات الاختبار تتناقص الفجوة بين الاحتمالين.

وقد زاد الاهتمام باختبارات الصواب-الخطأ المتعدد في السنوات العشر الأخيرة من القرن الماضي وبالأخص في المجالات الطبية.ويعود ذلك إلى أن أغلب الطلبة يفضلون فقرات الصواب والخطأ المتعدد كونها تراعي المعرفة الجزئية ويمكنهم أن يحصلوا من خلالها على علامات أعلى، ولتفوقها على فقرات الاختيار من متعدد في الخصائص السيكومترية (Frisbie, 1992). وقد أشار فرزبي وسويني (Frisbie and Sweeney, 1982) إلى أن هذا النوع من الاختبارات _على الأقل من الناحية النظرية_ يمتلك من الخصائص الجيدة ما تمتلكه اختبارات الاختيار من متعدد، كما أن لها مزايا أخرى هامة تفوق ما هو موجود في اختبارات الاختيار من متعدد، حيث يمكن بواسطتها الحصول على استجابات متعددة في وقت محدد، كما تتسم هذه الاختبارات بثبات أكبر من ثبات اختبارات الاختيار من متعدد، كما يمكن استخدام هذا النوع في تغطية محتوى أوسع من المادة، وتحقيق صدق أعلى.

وهدفت دراسة فرزبي وسويني (Frisbie and Sweeney; 1982) إلى مقارنة نوعين من الاختبارات (اختبار الاختيار من متعدد واختبار الصواب والخطأ المتعدد) المتوازية في المحتوى من حيث الزمن اللازم للإجابة، والثبات والصدق التلازمي، والصعوبة، وتفضيل الطالب. وقد شارك في الدراسة (574) طالباً مسجلين لمساق في الأحياء. وقد تم بناء نموذجين متوازيين في المحتوى يتكون الأول من (50) فقرة صواب وخطأ متعدد ويتكون الثاني من (50) فقرة اختيار من متعدد. وأظهرت النتائج أن الزمن الذي يكفي لإجابة أربع فقرات اختيار من متعدد يكفي فقط لإجابة ثلاث فقرات صواب وخطأ متعدد ، كما أن استخدام فقرات الصواب والخطأ المتعدد يسمح بتمثيل أكبر للمحتوى من فقرات الاختيار من متعدد وكان لها صدق تلازمي مرتفع. وقد عدّ معظم الطلبة (78%) أن فقرات الصواب والخطأ المتعدد أسهل وقد فضلها (66%) منهم على فقرات الاختيار من متعدد.

وفي دراسة أجراها فرزبي ودروفا (Frisbie and Druva, 1987) قارنا فيها بين اختبار الاختيار من متعدد واختبار الصواب و الخطأ المتعدد واختبار الاختيار من متعدد المركب، من حيث الثبات، حيث أجريت الدراسة على عينة من (279) طالباً في مادة الأحياء لطلبة الطب، وأشارت النتائج إلى عدم وجود فروق في معامل الثبات بين هذه الاختبارات.

وقارن إريكسون (Erickson, 1988) بين اختبارات الاختيار من متعدد والإجابة القصيرة، والصواب و الخطأ المتعدد، من حيث الصدق والثبات ومتوسط الصعوبة ومتوسط التمييز. وتكونت عينة دراسته من (89) طالب تمريض في المرحلة العليا. وأشارت النتائج إلى وجود فروق ذات دلالة إحصائية بين الأنواع الثلاثة من الاختبارات في كل من قيم معاملات ثبات الاتساق الداخلي ومتوسط صعوبة الفقرات وتمييزها ، بينما لم يكن هناك فروق ذات دلالة إحصائية بين معاملات الصدق. فقد تبين أن اختبار الصواب والخطأ المتعدد أعلى ثباتاً من الاختبارات الأخرى وأن تلك الاختبارات أكثر صعوبة منه.

كما قارن فرزبي وكريتر (Kreiter and Frisbie, 1989) بين فقرات الاختيار من متعدد، والصواب والخطأ المتعدد، باستخدام نماذج متوازية المحتوى لاختبار في مساق علم النفس الاجتماعي، من حيث الصدق والثبات، والصعوبة. و تبين من الدراسة أن اختبارات الصواب والخطأ المتعدد تتمتع بدرجة ثبات أعلى وأن أداء الطلبة عليها يكون أفضل مقارنة بفقرات الاختيار من متعدد. وقد كانت العلاقات الارتباطية بين الاختبارين قوية حيث كانت قيمة معامل الارتباط بينهما بعد تصحيحها قريبة جداً من واحد. وقد وفرت نتائج هذه الدراسة دعماً إضافياً لاستخدام اختبارات الصواب والخطأ المتعدد كبديل لاختبارات الاختيار من متعدد في قياس التحصيل.

وقدمت مراجعة فرزبي (Frisbie, 1992) للأدب التربوي المرتبط بفقرات الصواب والخطأ المتعدد تلخيصاً لخصائص هذه الفقرات، كما وفرت إجابات لعدد من الأسئلة المتعلقة بهذا النوع من الفقرات مقارنة بفقرات الاختيار من متعدد، من حيث الزمن اللازم لإجابتها، والطرق المناسبة لتقدير الثبات، ومؤشرات صدقها، وصعوبتها، وإدراك التلاميذ وتقبلهم لها. وخلصت المراجعة إلى أن الزمن اللازم لإجابة فقرة صواب وخطأ متعدد أطول من الزمن اللازم لإجابة فقرة اختيار من متعدد مكافئة لها في المحتوى، فالزمن الذي يكفي لإجابة ثلاث فقرات (ص خ م) قد يكفي لأكثر من أربع

فقرات (خ م). وفيما يتعلق بالثبات فقد تفوقت فقرات (ص خ م) على فقرات (خ م) بغض النظر عن طريقة التصحيح، وبالنسبة للصدق التلازمي؛ كشفت مراجعة معاملات الارتباط بعد التصحيح لأثر الانخفاض أنه يمكن استخدام فقرات (ص خ م) وفقرات(خ م) على نحو تبادلي حيث تراوحت معاملات الارتباط المصححة بين (91.) و(1.0). وبخصوص صعوبة الفقرات كانت النتائج متفاوتة وغير مستقرة.

وهدفت دراسة أمريزيق (1992) إلى مقارنة اختبار الاختيار من متعدد واختبار الصواب والخطأ واختبار الصواب و الخطأ المتعدد من حيث الثبات ومتوسط الصعوبة للاختبار ومتوسط تمييزه وشكل منحنى توزيع أداء المفحوصين على الاختبار. وقد استخدم طريقة الاستجابة المتعددة لتصحيح فقرات الصواب والخطأ المتعدد، أي يُعطى المفحوص العلامة الكاملة (1) على الفقرة عندما يجيب بصورة صحيحة عن جميع بدائلها، ويعطى العلامة صفر عندما يخفق في إجابة أي بديل من تلك البدائل. وقد أشارت النتائج إلى وجود فروق ذات دلالة إحصائية بين معاملات الثبات لصالح اختبار الاختيار من متعدد، كما لم توجد فروق بين معاملي ثبات اختبار الصواب والخطأ، واختبار الصواب

و الخطأ المتعدد. كما أشارت النتائج إلى عدم وجود فروق ذات دلالة إحصائية في معامل الصدق. أما من حيث الصعوبة فقد كان الفرق دالاً إحصائياً بين معاملي صعوبة اختبار الاختيار من متعدد واختبار الصواب والخطأ المتعدد ولم يكن هناك فرق دال بين معاملي الصعوبة لاختبار الصواب والخطأ واختبار الاختيار من متعدد.وبالنسبة للتمييز، لم يكن هناك فروق بين متوسطات تمييز الاختبارات الثلاثة. أما من حيث معاملات الالتواء، فقد كانت جميع معاملات الالتواء موجبة وكان معامل الالتواء لاختبار الاختيار من متعدد هو الأقرب إلى الصفر مقارنة بالاختبارين الآخرين .

وأجرى سوا لمة (1994) دراسة هدفت إلى مقارنة فقرات الصواب والخطأ المتعدد مع فقرات الاختيار من متعدد الموازية لها في المحتوى وعدد البدائل من حيث الثبات والصدق التلازمي والصعوبة. وأخذت الدراسة متغير عدد البدائل الصحيحة في فقرات الصواب والخطأ المتعدد بالاعتبار. وأشارت نتائج الدراسة إلى تمتع اختبارات الصواب والخطأ المتعدد بفعالية عالية في قياس تحصيل الطلبة للمهارات والقدرات التي تقيسها اختبارات الاختيار من متعدد الموازية لها في المحتوى.وقد كانت فقرات الاختيار من متعدد أسهل من فقرات الصواب والخطأ المتعدد وكانت معاملات الارتباط بين الاختبارين عالية نسبياً، وتميل العلامات على اختبارات الصواب والخطأ المتعدد لتكون أعلى ثباتاً من العلامات على اختبارات الاختيار من متعدد ويكون الفرق في الثبات له دلالته الإحصائية وبشكل خاص عندما يكون هناك أكثر من بديل صحيح في فقرة الصواب والخطأ المتعدد.

وفي المجالات الطبية، تم مقارنة الخصائص السيكومترية لفقرات الصواب والخطأ المتعدد وفقرات الاختيار من متعدد(Downing, Baranowski, Gross,and Norcini, 1995).

وقد بينت النتائج أن فقرات الصواب والخطأ المتعدد أكثر ثباتاً من فقرات الاختيار من متعدد. بينما تحقق لفقرات الاختيار من متعدد درجة أعلى من الصدق المرتبط بمحك حيث كانت أكثر ارتباطاً بمقياس مستقل للأداء .

وقد أجرى المومني (1996) دراسة هدفت إلى مقارنة اختبارات الاختيار من متعدد مع اختبار الصواب والخطأ المتعدد من حيث تأثيرها في احتفاظ طلاب الصف الثاني الثانوي الأكاديمي بمادة قواعد اللغة الإنجليزية في مستويي المعرفة والفهم. وأظهرت نتائج الدراسة أن متوسط علامات مجموعة الصواب والخطأ المتعدد كانت أعلى رقمياً من متوسط مجموعة الاختيار من متعدد إلا أن هذا الفرق لا يرقى لمستوى الدلالة الإحصائية.

وقد كانت أسئلة الصواب والخطأ المتعدد مداراً للاهتمام في مجال الاختبارات المتعلقة باللغة الثانية.فقد قارن دودلي(Dudley,2006) بين اختبارات الصواب والخطأ المتعدد واختبارات الاختيار من متعدد في مجالي المفردات والقراءة باللغة الإنجليزية.وقد حللت البيانات باستخدام كل من النظرية التقليدية في القياس ونموذج راش. وقد بينت النتائج تحسناً في الأداء عند استخدام فقرات الصواب والخطأ المتعدد. وقد استجاب المشاركون لعدد أكبر من فقرات الصواب والخطأ المتعدد مقارنة بفقرات الاختيار من متعدد. كما كشف التحليل عن استقلالية الفقرات وتحسن الثبات في مجال القراءة.

أما من حيث طريقة التصحيح فيؤخذ على فقرات الصواب والخطأ المتعدد أنه لا يوجد اتفاق على طريقة تصحيحها، إذ يمكن تصحيح فقرات الصواب والخطأ المتعدد بعدة طرق ربما يكون لها أثار مختلفة على خصائص الفقرات نفسها، وعلى دقة قياسها لقدرات الأفراد. وقد درس ألبانيس وسابرس (Albanese and Sabers, 1988) أربع طرق يمكن استخدامها في تصحيح فقرات الصواب والخطأ المتعدد. بينت النتائج أن أساليب التصحيح التي تراعي المعرفة الجزئية أكثر ثباتاً وأكثر صدقاً تلازمياً وتنبؤياً من تلك الأساليب التي لا تراعي المعرفة الجزئية. وقد حدد فرزبي (Frisbie, 1992) نتيجة لمراجعته الشاملة للأدب التربوي المتعلق بفقرات الصواب والخطأ المتعدد بعض الأسئلة المهمة التي يجب أن تكون على جدول أعمال البحث الخاص بهذا النمط من الفقرات، ومن هذه الأسئلة ما يلي:

1. ما هو أسلوب التصحيح المطلوب للمعايرة وفق نماذج النظرية الحديثة للقياس؟
2. كيف تؤثر بدائل التصحيح لفقرات الصواب و الخطأ المتعدد وإجراءات المعايرة على استخدام هذه الفقرات في الاختبارات المكيفة؟

وقد قارن تيساي وسوين (Tsai and Suen,1993) ستة أساليب لتصحيح فقرات الصواب والخطأ المتعدد تعتمد على عدد البدائل k ، وعدد الإجابات الصحيحة i ، وعدد البدائل المتروكة بدون إجابة j ، وهذه الأساليب هي:

1. **Multiple-response (MR) method** : ويتمثل هذا

الأسلوب بإعطاء علامة كاملة لكل فقرة إذا أجاب الطالب بشكل صحيح عن جميع البدائل، وصفر إذا لم يُجب بشكل صحيح عن أي من البدائل، ويتم ذلك وفق القاعدة:

$$f(i)=\begin{cases}1 & \text{if } i=k;\\ 0 & \text{otherwise}\end{cases}$$

2. **Count for 3 options correct (Count 3) method**: ويتمثل هذا الأسلوب بإعطاء نصف علامة إذا أجاب بشكل صحيح عن ثلاثة بدائل وعلامة كاملة إذا أجاب عن البدائل الأربعة بشكل صحيح، وصفر فيما عدا ذلك. ويتم ذلك وفق القاعدة:

$$f(i)=\begin{cases}0 & \text{if } i=0 \text{ or } 1 \text{ or } 2;\\ \dfrac{i-2}{k-2} & \text{otherwise}\end{cases}$$

3. **Count for 2 options correct (Count 2) method**: ويتمثل هذا الأسلوب بإعطاء ثلث علامة إذا أجاب عن بديلين بشكل صحيح، وثلثَي علامة إذا أجاب عن ثلاثة بدائل بشكل صحيح، وعلامة كاملة إذا أجاب عن البدائل الأربعة بشكل صحيح، وصفر فيما عدا ذلك. ويتم ذلك وفق القاعدة:

$$f(i)=\begin{cases}0 & \text{if } i=0;\\ \dfrac{i-1}{k-1} & \text{otherwise}\end{cases}$$

4. **Multiple true - false (MTF) method**: ويتمثل هذا الأسلوب بإعطاء رُبع علامة لكل بديل يتم الإجابة عنه بشكل صحيح، ويتم ذلك وفق القاعدة التالية:

$$f(i)=\frac{i}{k} \qquad (k\geq i\geq 0)$$

5. **Correction for guessing (CFG) method:** ويتمثل هذا الأسلوب باستخدام طريقة التصحيح لأثر التخمين، ويتم ذلك وفق القاعدة التالية:

$$f(i)=\frac{i-(k-i)}{k} \qquad (k\geq i\geq 0)$$

6. **Let - Omit (LO) method:** وهنا يحاسب الممتحن على الإجابة الصحيحة والإجابة الخاطئة فقط، ويتم ذلك وفق القاعدة التالية:

$$f(i)=\frac{i-(k-i-j)}{k} \qquad (k\geq i\ ,\ j\geq 0)$$

وقد أظهرت نتائج الدراسة أنه لا توجد فروق ذات دلالة إحصائية($\alpha = 0.05$) في قيم الثبات والتمييز تعزى لطريقة التصحيح . أما من حيث الصعوبة، فقد تبين أن كلاً من أسلوب الصواب والخطأ المتعدد (MTF) وأسلوب تصحيح التخمين (CFG) وأسلوب دعها ـ احذفها (Let-Omit) تعطي أوساطاً حسابية أعلى للفقرات.

ويلاحظ هنا أن الأساليب الأربعة الأولى هي تجزئة منتظمة للعلامة الواقعة بين(0) و(1)، أما الأسلوبان الخامس والسادس فهما تجزئة منتظمة للعلامة الواقعة بين(-1) و(1)، ويمثل الأسلوب الأول الإتقان التام ، وهو متشدد لكونه يعتمد المعرفة الكاملة ولا يراعي المعرفة الجزئية بتاتاً في حين أن بقية الطرق تراعي المعرفة الجزئية بدرجات متفاوتة.

والمتفحص لدراسة تيساي وسوين (Tsai and Suen, 1993) يرى أنها لاتستجيب للتساؤلات التي طرحها فرزبي (Frisbie, 1992). و تعد دراسة براد (Brad, 2003) المتعلقة بتصحيح فقرات الصواب والخطأ المتعدد ضمن سياق الاستجابة لهذه التساؤلات إلى حد ما؛ إذ كانت أساليب التصحيح الأربعة الأولى _المشار لها سابقاً في دراسة تيساي وسوين_ مدار اهتمام دراسته. فقد اهتمت دراسته بمقارنة هذه الأساليب من خلال فحص العلاقة بين العلامات الخام الناتجة عن استخدام كل منها وتقديرات القدرة المقدرة بحسب النموذج الثلاثي من نماذج نظرية استجابة الفقرة(Item response theory). وتبين من النتائج أنه توجد علاقة موجبة شبه تامة بين العلامات الخام وتقديرات القدرة في الأساليب الأربعة . فقد تراوحت قيم معاملات ارتباط بيرسون بين 0.97 و0.99 على مستوى الاختبار المتكون من خمس فقرات، وبين 0.78 و 0.98 على مستوى الفقرة الواحدة . كذلك أشارت النتائج إلى عدم اختلاف الثبات للعلامات باختلاف أسلوب التصحيح.

وهكذا يبدو واضحاً أن معظم الدراسات السابقة قد اهتمت بدراسة الخصائص السيكومترية النسبية لفقرات الصواب والخطأ مقارنة مع أشكال أخرى للفقرات، وأن هناك عدداً قليلاً جداً- في حدود ما توفر للباحثان- من الدراسات التي اهتمت بطرق تصحيح فقرات الصواب والخطأ المتعدد. إذ اهتمت دراسة تيساي وسوين (Tsai and Suen,1993) بأثر طريقة التصحيح في كل من الثبات للاختبار وصعوبة الفقرات وتمييزها بحسب النظرية التقليدية للقياس، بينما درس براد (Brad,2003) أثر طريقة التصحيح لفقرات الصواب والخطأ على الصدق والثبات بحسب النظرية الحديثة للقياس. وتتميز الدراسة الحالية بأنها تستجيب للتساؤل الأول الذي طرحه فرزبي (Frisbie, 1992) عن أسلوب التصحيح المناسب لمعايرة فقرات الصواب و الخطأ المتعدد وفق نماذج النظرية الحديثة في القياس التربوي من خلال دراسة أثر طريقة تصحيح فقرات الصواب والخطأ المتعدد في دقة تقدير قدرات الأفراد وصعوبة الفقرات باستخدام نموذج التقدير الجزئي. فهي أول دراسة تهتم بمعيار دقة التقديرات لمعلمتي القدرة وصعوبة الفقرة باختلاف طريقة التصحيح ،وسيتم ذلك من خلال فحص مطابقة البيانات الخاصة باستجابات المفحوصين على فقرات الاختبار لنموذج التقدير الجزئي وفحص الاختلافات في دقة تقدير قدرات الأفراد وصعوبة الفقرات. وقد اقتصرت الدراسة الحالية على استخدام أساليب التصحيح الأربعة الأولى فقط، حيث تم استثناء

الأسلوبين الخامس والسادس، كونهما يحتاجان إلى إعطاء المفحوصين تعليمات خاصة للإجابة مختلفة عن التعليمات المقدمة للأساليب الأربعة الأولى. وكذلك فإن جروس(Gross) (المشار إليه في سوا لمة، 1994) يؤيد استعمال فقرات الصواب والخطأ المتعدد، و يرى في الوقت نفسه أنه لا داعي لتصحيح علاماتها من أثر التخمين.

ويستخدم نموذج التقدير الجزئي (Partial Credit Model) (Masters, 1982) مع الفقرات التي يمكن تصحيحها في عدة مستويات من الأداء تقع بين الإجابة الخطأ كلياً والإجابة الصحيحة التامة، أي عندما تتعدد فئات(مستويات) الاستجابة للفقرة . وتجدر الإشارة إلى أن تحويل الفقرة متعددة المستويات للاستجابة إلى فقرة ثنائية غير ملائم في معظم الحالات حيث يغير ذلك في طبيعة التدريج للمقياس وبالتالي قد يهدد الصدق للمقياس بصورة خطيرة.(و Han,n.d.2008) ويُنمذج هذا النموذج الفهم الجزئي للمسائل متعددة الخطوات وهو تعميم لنموذج راش. وبدلاً من وجود معلمة صعوبة واحدة لكل فقرة كما هو الحال في الاستجابة الثنائية فإن نموذج التقدير الجزئي يتطلب وجود معلمة صعوبة واحدة لكل عتبة فئة (category threshold) من الممكن أن يتجاوزها الفرد. فمثلاً إذا كان هناك خمس فئات(مستويات) للاستجابة فإنه سيكون هناك أربعة معالم للصعوبة لوجود أربع عتبات تفصل هذه المستويات. ويمثل كل معلم منها صعوبة الانتقال من مستوى إلى المستوى الذي يليه. وقد أطلق ماستر على هذه الحركات ومعالم صعوبتها مسمى "صعوبة الخطوات" . والنموذج لا يضع أي قيود على معالم هذه العتبات.لذلك يسمح لعتبات أي فقرة بالاختلاف بدون أي اعتبار لعلاقتها بالفقرات الأخرى. ويتطلب النموذج أن تكون المستويات مرتبة بحيث تمثل الأعداد الصحيحة (0،1،2،3) مستويات الأداء المختلفة بحيث يعبر العدد الأقل عن قيمة أقل للسمة المقاسة. وتجدر الإشارة إلى أنه لا يمكن للفرد أن يصل المستوى الثالث إلا إذا تجاوز المستويات التي تسبقه، لكنه لا توجد متطلبات تتعلق بالصعوبة المرتبطة بالتقدم من مستوى إلى آخر. فمثلاً قد يكون من الصعب على الفرد الوصول إلى المستوى الثاني ، لكن بعد وصوله إلى هذا المستوى قد يكون من السهل عليه تجاوزه للمستوى الثالث. ويفترض نموذج التقدير الجزئي أن جميع الفقرات تميز بالتساوي عند جميع مستويات القدرة.(Lustina, 2004). ويعد النموذج مناسباً للتطبيق عندما تشكل مجموعة من الفقرات الثنائية المرتبطة فقرة عنقودية.(De Gruijter and Van der Kamp,2003) وهذا هو مجال اهتمام الدراسة الحالية حيث فقرة الصواب والخطأ المتعدد هي فقرة عنقودية تتكون من عدة فقرات من نوع الصواب والخطأ التي تشترك في نفس المتن.

ويتطلب هذا النموذج تحديداً مسبقاً لعدد مستويات الأداء في ضوء عدد الخطوات المطلوب تنفيذها بنجاح. ويبين الجدول (2) مستويات الأداء بدلالة عدد خطوات حل المسألة الرياضية التالية:

$$\sqrt{\frac{7.5}{0.3}-16}=?$$

حيث عدد الخطوات المطلوب إنجازها هو ثلاث خطوات.

جدول (2): العلاقة بين مستويات الأداء وخطوات الاستجابة لمسألة رياضية من ثلاث خطوات

الخطوات	مستويات الأداء			
	0	1	2	3
الأولى	الخطوة الأولى 0 ← 1			
الثانية		الخطوة الثانية 1 ← 2		
الثالثة			الخطوة الثالثة 2 ← 3	

حيث يُعطى المفحوص العلامة (1) بدلاً من الصفر إذا قام بحل الخطوة الأولى، والعلامة (2) بدلاً من (1) إذا قام بحل الخطوة الثانية، والعلامة (3) بدلاً من (2) إذا قام بحل الخطوة الثالثة. ويفسر ماسترز (Masters, 1982) علامات المستويات على أنها تمثل عدد المهمات الجزئية أو الخطوات في السؤال التي ينجزها الفرد بنجاح.

وقد بُني نموذج التقدير الجزئي اعتماداً على فرضية أن درجة صعوبة الخطوة (k) تدل على صعوبة تخطي الفرد للخطوة (k) بدلاً من بقائه في المستوى (k-1)، وعند اجتياز الخطوة k فإن الاختيار يقع على بديل الاستجابة k بدلاً من (k-1) في الاستجابة على الفقرة حيث تؤخذ الخطوات بشكل مرتب متسلسل، ويمكن التعبير عن احتمالية وصول الشخص (n) لمستوى الأداء (k) للفقرة (i) بالعلاقة الرياضية التالية(Masters, 1982):

$$P_{kni}(\theta)=\frac{\exp(\theta_n-b_{ik})}{1+\exp(\theta_n-b_{ik})}=\frac{\pi_{kni}}{\pi_{(k-1)ni}+\pi_{kni}}$$

حيث تعني π_{kni} : احتمالية استجابة الفرد n على الخطوة k في الفقرة i. وتعطى قيمة π_{xni} على الصورة الرياضية التالية:

$$\pi_{xni}=\frac{\exp\sum_{j=0}^{x}(\theta_n-b_{ij})}{\sum_{j=0}^{mi}\exp\sum_{k=0}^{j}(\theta_n-b_{ij})}$$

حيث θ_n: قدرة الفرد (n)، b_{ik} صعوبة الخطوة (k) في الفقرة (i) ، m_i عدد الخطوات(البدائل)، k عدد المستويات $k=0, 1,2 \ldots\ldots m_i$.

وتعرف هذه المعادلة بدالة استجابة فئة العلامة (score category response function) وتستخدم على نطاق واسع في برنامج Bigstep (المتوفر للباحثين) لتقدير معلمة الصعوبة للفقرة ولكل عتبة من عتباتها التي تفصل فئات الاستجابة بعضها عن بعض. ويوفر البرنامج كذلك إحصاءات المطابقة الداخلية والخارجية وتقديرات لمعلمة القدرة ومنحنيات خاصية الفقرات التي تحدد احتمال كل مستوى من مستويات الإجابة (Steps probability curves) بدلالة الفرق بين موقع المفحوص وموقع الفقرة على متصل السمة المقاسة. (أنظر الملحق 1)

وقد تم استخدام نموذج التقدير الجزئي في العديد من الدراسات العربية والأجنبية (التقي،1992؛ شاهين، 1993؛ الشريفين، 2003؛ العلي، 2004؛ Dungohue, 1994; Lee, Kolen, Frisbie and Ankennan, 2001; Randall and Ching-Fung, 2004; Wright and Masters, 1982)

ومن خلال مراجعة نتائج هذه الدراسات التي استخدمت نموذج التقدير الجزئي، يمكن استخلاص ما يلي:

1. إن حجم المعلومات التي تعطيها الأسئلة متعددة التدريج تزداد بزيادة عدد نقاط التدريج.
2. إن نموذج التقدير الجزئي فاعل في تدريج الأفراد والأسئلة ويوصي الباحثون باستخدامه عند تقنين الاختبارات التحصيلية ذات الفقرات متعددة النقاط .
3. إن الخطأ المعياري في قياس قدرات الأفراد وتقدير قيم صعوبة الأسئلة، يتناقص بزيادة عدد نقاط التدريج.
4. تزداد الكفاءة النسبية للاختبار بزيادة عدد خطوات الحل أو عدد مستويات الأداء.

وتحاول الدراسة الحالية التحقق من هذه الاستنتاجات في فقرات الصواب والخطأ المتعدد، من خلال استخدام نموذج التقدير الجزئي، كونه النموذج المناسب لتحليل هذه الفقرات حسب طرق التصحيح المختلفة المستخدمة في هذه الدراسة.

مشكلة الدراسة وأهميتها:

في العقدين الماضيين زاد الاهتمام بفقرات الصواب والخطأ المتعدد، واتسع مجال استخدامها، ورافق ذلك مشكلات تتعلق بأسلوب تصحيحها حيث توجد أكثر من طريقه لتصحيح فقرات الصواب والخطأ المتعدد. وفي ظل تعدد طرق التصحيح لفقرات الصواب والخطأ المتعدد، كان لا بد من مقارنة هذه الأساليب لمعرفة الأنسب منها للحصول على اختبار يتمتع بخصائص سيكومترية أفضل من حيث الصدق والثبات. ومع التوجه الواسع نحو استخدام النظرية الحديثة في القياس في بناء الاختبارات ومعايرتها فإنّ هناك حاجةً لمعرفة أسلوب التصحيح الأنسب لمعايرة فقرات الصواب والخطأ المتعدد وفق نماذج النظرية الحديثة للقياس. وبذلك تأتى هذه الدراسة لاستكمال ما توصلت إليه الدراسات السابقة واستجابة لبعض توصياتها. كما يمكن أن توفر نتائج الدراسة المعلومات الكافية، لمستخدمي هذا النمط من الاختبارات سواءً في حقل التدريس أو في غيره، حول أنسب طرق تصحيح فقرات الصواب والخطأ المتعدد الضرورية لمعايرة الفقرات وفق نماذج النظرية الحديثة في القياس.

هدف الدراسة وأسئلتها

تهدف الدارسة إلى بحث أثر طريقة تصحيح فقرات الصواب والخطأ المتعدد في دقة تقدير قدرات الأفراد وصعوبة الفقرات باستخدام نموذج التقدير الجزئي وتحديد انسب طرق تصحيح فقرات الصواب والخطأ لمعايرة فقرات الاختبار.

وبالتحديد ستتم الإجابة عن الأسئلة التالية:

1. ما مدى مطابقة البيانات الخاصة بفقرات الصواب و الخطأ المتعدد لنموذج التقدير الجزئي في كل طريقة من طرق التصحيح؟
2. هل تختلف دقة تقدير معلمة القدرة للفرد(مقدرة بالخطأ المعياري في تقديرها) باختلاف طريقة تصحيح فقرات الصواب والخطأ المتعدد؟
3. هل تختلف دقة تقدير معلمة الصعوبة للفقرة ككل(مقدرة بالخطأ المعياري في تقديرها) باختلاف طريقة تصحيح فقرات الصواب والخطأ المتعدد؟

الطريقة والإجراءات

عينة الدراسة

تكونت عينة الدراسة من 873 طالباً وطالبة من طلبة الصف الأول الثانوي(696 في الفرع الأدبي، 177في فرع الإدارة المعلوماتية) في مديرية تربية إربد الأولى. إذ تم اختيار 10 مدارس بالطريقة العشوائية البسيطة فيها 29 شعبة، من بين 46 مدرسة فيها 110 شعب للصف الأول الثانوي. و تم تطبيق أداة الدراسة على جميع طلبة الصف الأول ثانوي الذكور والإناث بفرعيه الأدبي والإدارة المعلوماتية في تلك المدارس، وتشكل عينة الدراسة حوالي 24% من إجمالي طلبة الأول الثانوي في الفرعين الأدبي والإدارة المعلوماتية للعام الدراسي 2005/2004 بمديرية تربية إربد الأولى.

أداة الدراسة:

من أجل تحقيق أهداف الدراسة، تم إعداد اختبار من 35 فقرة صواب وخطأ متعدد تقيس تحصيل طلبة الصف الأول الثانوي في وحدة الإحصاء. وعند بناء هذا الاختبار، تم الاسترشاد بالأسس العامة المتبعة في تصميم اختبارات التحصيل الصفية (ثورندايك وهيجن، 1986)؛ إذ تم صياغة الأهداف السلوكية للوحدة ، وتحديد المحتوى والموضوعات المراد قياسها، و إعداد جدول المواصفات ، وكتابة (50) فقرة من نوع الصواب والخطأ المتعدد جميعها بأربعة بدائل لقياس الأهداف الواردة في جدول المواصفات. وقد استقر الاختبار نتيجة لعملية التحكيم والتجريب الأولي على 35 فقرة ، وقد بلغ معامل ثبات الاتساق الداخلي له بطريقة كرونباخ ألفا (0.93) وهي قيمة عالية تدل على ثبات الاختبار. ويوضح المثال التالي إحدى فقرات الاختبار حيث يستجيب الطالب لكل بديل من البدائل بوضع دائرة أمام البديل حول كلمة صح إذا كان صحيحاً أو كلمة خطأ إذا كان غير صحيح.

مثال: فقرة من الاختبار

أي القيم التالية هي قيم ممكنة لمعامل الارتباط ؟

صح **خطأ** أ. 1.2

صح خطأ ب. -1

صح خطأ ج. 0.6

صح **خطأ** د. -2

الإجـراءات

بعد إعداد الاختبار بصورته النهائية، تم تطبيقه خلال الفترة الواقعة ما بين2005/5/15ولغاية2005/5/26 على أفراد عينة الدراسة كل في مدرسته وفق برنامج زمني بإشراف الباحث الأول. وقبل بدء الطلبة بالإجابة عن الأسئلة تم توضيح تعليمات الاختبار لهم ، و إبلاغهم بأن نتيجة هذا الاختبار سوف تحتسب ضمن العلامة النهائية للمادة، وذلك لاستثارة اهتمامهم بالاختبار وزيادة الجدية لديهم في الإستجابة. وقد أعطي الطلبة الوقت الكافي (حصتين صفيتين=100 دقيقة) لتمكينهم من إجابة جميع الأسئلة.

وبعد الانتهاء من تطبيق الاختبار، تم تصحيحه بتحديد مستوى الأداء لكل طالب على كل فقرة بدلالة عدد الاستجابات الصحيحة له في تلك الفقرة، والتي تعبر عن عدد المهمات الجزئية المنجزة لكل فقرة (0، 1، 2، 3، 4)، ومن ثم إدخال هذه االبيانات في الحاسوب كملف بيانات(Data File) معتمد لجميع طرق التصحيح المستخدمة في الدراسة.وقد تم تحويل هذه البيانات إلى مستويات بموجب القاعدة المتضمنة في كل طريقة من طرق التصحيح الأربعة المستخدمة في الدراسة . ونتيجة لذلك اعطيت استجابات الطلبة المستويات (0، 1) للطريقة الأولى، و (0، 1، 2) للطريقة الثانية،و(0، 1، 2، 3) للطريقة الثالثة، و(0، 1، 2، 3، 4) للطريقة الرابعة.

وتوضح الجداول (3-6) العلامات التي تعطى للفقرة في مستويات الأداء المختلفة بدلالة عدد المهمات الجزئية المنجزة(الخطوات)والمعبر عنها بعدد الاستجابات الصحيحة للفرد.

جدول(3) :تدريج علامة الفقرة بحسب الطريقة الأولى للتصحيح

عدد الاستجابات الصحيحة	مستويات الأداء	
	0	**1**
4	0 ← خطوة واحدة	1

جدول (4): تدريج علامة الفقرة بحسب الطريقة الثانية للتصحيح

عدد الاستجابات الصحيحة	مستويات الأداء		
	0	**1**	**2**
3	0 ← الخطوة الأولى 0.5		
4		الخطوة الثانية 0.5 → 1	

جدول (5): تدريج علامة الفقرة بحسب الطريقة الثالثة للتصحيح

عدد الاستجابات الصحيحة	مستويات الأداء			
	0	**1**	**2**	**3**
2	0 ← الخطوة الأولى 0.33			
3		0.33 → الخطوة الثانية 0.66		
4			0.66 → الخطوة الثالثة 1	

جدول (6): تدريج علامة الفقرة بحسب الطريقة الرابعة للتصحيح

عدد الاستجابات الصحيحة	مستويات الأداء				
	0	**1**	**2**	**3**	**4**
1	0 ← الخطوة الأولى 0.25				
2		0.25 ← الخطوة الثانية 0.5			
3			0.5 ← الخطوة الثالثة 0.75		
4				0.75 ← الخطوة الرابعة 1	

ونظراً لأن برمجية (BigSteps) لا تتقبل العلامات الكسرية، فقد تم تحويل العلامة على الفقرة وبحسب كل أسلوب من أساليب التصحيح الأربعة، بما يتلاءم مع نموذج التقدير الجزئي، بعد إجراء تحويل لعلامة الفقرة بحسب كل طريقة من طرق التصحيح لتصبح عددا صحيحا انسجاما مع متطلبات نموذج التقدير الجزئي المستخدم في الدراسة الحالية، وبذلك تبقى العلامة كما هي في طريقة التصحيح

الأولى، ويتم ضرب العلامة في (2) في طريقة التصحيح الثانية، وفي (3) في طريقة التصحيح الثالثة، وفي (4) في طريقة التصحيح الرابعة. ويبين الجدول (7) العلامات المستحقة وفق كل طريقة من طرق التصحيح بعد تحويلها إلى أعداد صحيحة.

جـدول (7): العلامة المستحقة للمفحوص بحسب طريقة التصحيح(ط) وعدد الاستجابات الصحيحة

عدد الاستجابات الصحيحة	العلامة/ط1	العلامة/ط2	العلامة/ط3	العلامة/ط4
0	0	0	0	0
1	0	0	0	1
2	0	0	1	2
3	0	1	2	3
4	1	2	3	4

المعالجات الإحصائية:

قبل معالجة البيانات إحصائياً كان لا بد من فحص الافتراضات الأساسية التي يجب توفرها في البيانات عند تطبيق نموذج التقدير الجزئي، وأهم هذه الافتراضات أحادية البعد للسمة المقيسة. وقد تم ذلك من خلال التحليل العاملي للبيانات وحساب معامل الارتباط بين الأداء على الفقرة والعلامة الكلية لجميع فقرات الاختبار في كل طريقة من طرق التصحيح الأربعة.

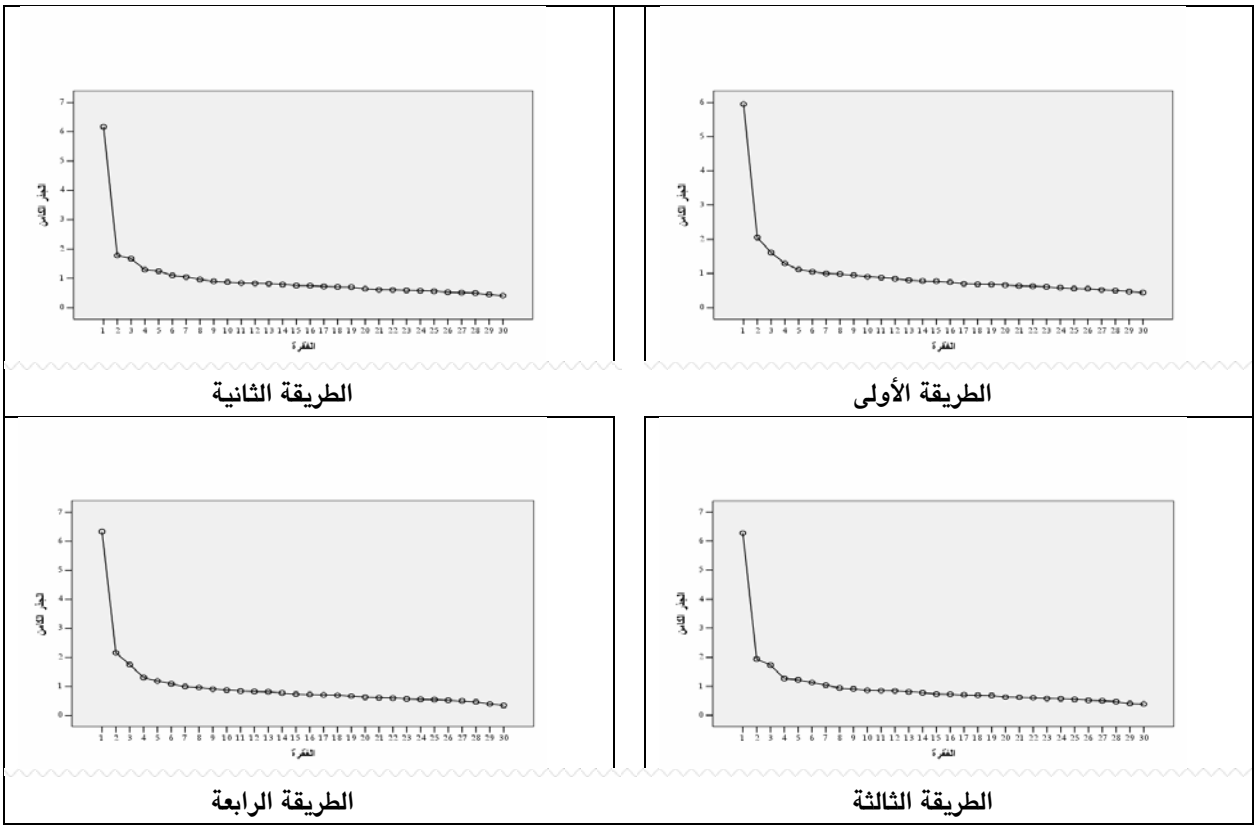
وقد تم إجراء تحليل عاملي للبيانات باستخدام طريقة المكونات الأساسية (Principal Components Analysis). وفي التحليل العاملي تعتمد أحادية البعد على أن نسبة الجذر الكامن للعامل الأول إلى الجذر الكامن للعامل الثاني تكون نسبة كبيرة لا تقل عن (2)، وأن ما يفسره العامل الأول من التباين أكثر من 20% تقريبا (Hambleton and Swaminathan, 1985).

ويبين الجدول (8) قيم الجذر الكامن، ونسب التباين المفسر للعامل الأول والعامل الثاني، ونواتج قسمة قيمة الجذر الكامن للعامل الأول على العامل الثاني لكل طريقة من طرق التصحيح الأربعة.

جدول (8): قيم الجذر الكامن ونسب التباين المفسر للعامل الأول والعامل الثاني، ونواتج قسمة الجذر الكامن للعامل الأول على الجذر الكامن للعامل الثاني في كل من طرق التصحيح الأربعة.

طريقة التصحيح		العامل 1	العامل 2	ناتج
الأولى	الجذر الكامن	5.95	2.055	2.895
	التباين المفسر	%19.82	%6.852	
الثانية	الجذر الكامن	6.166	1.781	3.462
	التباين المفسر	%20.554	%5.937	
الثالثة	الجذر الكامن	6.278	1.940	3.236
	التباين المفسر	%20.928	%6.467	
الرابعة	الجذر الكامن	6.330	2.154	2.939
	التباين المفسر	%20.099	%7.181	

ويتبين من جدول (8) أن الاختبار التحصيلي كان أحادي البعد في طرق التصحيح الأربعة، حيث أن نسبة الجذر الكامن للعامل الأول إلى العامل الثاني كانت كبيرة وتزيد على (2)، وأن نسبة التباين المفسر من العامل الأول في طرق التصحيح الأربعة هي نسبة عالية(أكبر أو تساوي 20% تقريباً). ويدعم ذلك الاستنتاج التمثيل البياني لقيم الجذور الكامنة للعوامل المكونة للاختبار في كل طريقة من طرق التصحيح المعتمدة في الدراسة (أنظر الشكل 1).



الشكل (1): رسم بياني لقيم الجذور الكامنة للعوامل المختلفة بحسب طرق التصحيح المعتمدة

كما تم حساب معامل الارتباط بين الأداء على كل فقرة والعلامة الكلية للاختبار في كل طريقة من طرق التصحيح ، حيث تشير الارتباطات المرتفعة بين الفقرة والعلامة الكلية_في حالة تحققها لمعظم الفقرات_ إلى أن فقرات الاختبار تقيس بعدا واحدا فقط تعبر عنه العلامة الكلية. كما يمثل معامل الارتباط بين الأداء على الفقرة والأداء على الاختبار القدرة التمييزية للفقرة. ويبين الجدول (9) معاملات الارتباط بين العلامات المستحقة على الفقرات والعلامة الكلية في كل طريقة من طرق التصحيح الأربعة. وقد تبين أن جميع معاملات الارتباط ذات دلالة إحصائية ($\alpha = 0.05$)، وأن معظمها يتخذ قيماً متوسطة لكنها تعد عالية عندما يتعلق الأمر بالفقرات.

جدول(9): معاملات الارتباط بين العلامة المستحقة على الفقرة والعلامة الكلية على الاختبار في طرق التصحيح المعتمدة

رقم الفقرة	طريقة التصحيح			
	1	2	3	4
1	0.31	0.30	0.29	0.28
2	0.23	0.25	0.30	0.32
3	0.47	0.49	0.47	0.46
4	0.42	0.45	0.47	0.46
5	0.41	0.40	0.41	0.40
6	0.43	0.44	0.47	0.47
7	0.48	0.50	0.52	0.52
8	0.38	0.36	0.28	0.29
9	0.46	0.45	0.41	0.38
10	0.54	0.53	0.51	0.51
11	0.39	0.48	0.49	0.47
12	0.42	0.43	0.41	0.41
13	0.49	0.52	0.51	0.50
14	0.41	0.48	0.49	0.49
15	0.43	0.45	0.50	0.51
16	0.50	0.51	0.53	0.54
17	0.36	0.39	0.43	0.45
18	0.33	0.30	0.36	0.38
19	0.51	0.49	0.44	0.42
20	0.31	0.42	0.43	0.43
21	0.47	0.54	0.57	0.57
22	0.56	0.55	0.55	0.53
23	0.55	0.55	0.56	0.57
24	0.58	0.59	0.56	0.55
25	0.47	0.46	0.50	0.51
26	0.39	0.37	0.36	0.35
27	0.56	0.55	0.53	0.52
28	0.33	0.23	0.25	0.27
29	0.40	0.39	0.41	0.45
30	0.41	0.40	0.42	0.45

ونظراً لأن افتراض الاستقلال الموضعي (Local Independence) يكافئ افتراض أحادية البعد (Hambleton and Swaminathan, 1985)، فإن تحقق افتراض أحادية البعد في الاختبار يعني تحقق افتراض الاستقلال الموضعي. وبما أن معظم مؤشرات التمييز للفقرات تتخذ قيماً متوسطة، فإنه يمكن اعتبار الفقرات متكافئة في قدرتها التمييزية. هذا وقد لوحظ بأن ما يزيد على 95% من الطلبة قد أنهوا الامتحان قبل الوقت المحدد له مما يعني أن السرعة في الإجابة لم تلعب دوراً في تحديد العلامة للمفحوص. وبذلك تتحقق افتراضات نموذج التقدير الجزئي.

وعند البحث في معامل الفصل (التباعد) للأفراد، وثبات هذا المعامل وعدد الطبقات الإحصائية المتمايزة للأفراد ، فإن الجدول (10) يبين قيم هذه المؤشرات في كل طريقة من طرق التصحيح.

جدول (10): معامل التباعد بين الأفراد وثبات التباعد لتقديرات قدرات الأفراد وعدد الطبقات الإحصائية المتمايزة في طرق التصحيح الأربعة

طريقة التصحيح	مؤشر الفصل بين الأفراد	ثبات تباعد الأفراد	عدد الطبقات الإحصائية للأفراد
ط1	1.89	0.78	2.853
ط2	2.14	0.82	3.186
ط3	2.17	0.82	3.226
ط4	2.16	0.82	3.213

ويتضح هنا أن قيم معامل التباعد للأفراد وقيم ثبات هذا المعامل كانت متقاربة في طرق التصحيح الثانية والثالثة والرابعة، لكنها تشير إلى تحسن طفيف على طريقة التصحيح الأولى،وهذا يعني أن اختلاف طريقة التصحيح لفقرات الصواب والخطأ المتعدد، لا يؤثر تأثيرا كبيرا في معامل الثبات ولا في عدد الطبقات الإحصائية المتمايزة التي تنتج عن تمايز الأفراد.

ولفحص مدى مطابقة الاستجابات على فقرات الاختبار مع نموذج التقدير الجزئي بحسب طرق التصحيح المختلفة، وباستخدام برمجية (BIGSTEPS)، تم تفحص قيم إحصائيات المطابقة الداخلية والخارجية (INFIT and OUTFIT Values) للأفراد والفقرات، للتأكد من وقوعها داخل الحدود المقبولة للمطابقة (0.7 إلى 1.3).

وتم معالجة البيانات بعد حذف الأفراد غير المطابقين والفقرات غير المطابقة لنموذج التقدير الجزئي بهدف الحصول على تقديرات القدرة والاخطاء المعيارية لتلك التقديرات، وتقديرات الصعوبة والاخطاء المعيارية لتلك التقديرات. وتم إجراء تحليل التباين وفق تصميم القياسات المتكررة للأخطاء المعيارية في تقدير قدرات الأفراد ،وللأخطاء المعيارية في تقدير صعوبة الفقرات حسب طريقة التصحيح. كما استخدم اختبار Least significant differences(LSD) للمقارنات الثنائية للكشف عن مصدر الفروق في أوساط الأخطاء المعيارية بين طرق التصحيح المختلفة.

النتـائـج

أولاً: مـدى مطابقـة البيانـات لنمـوذج التقـدير الجزئـي فـي كـل طريقة من طرق التصحيح:

لتحديد مؤشرات المطابقة الخاصة بالأفراد، تم تقدير القدرة لكل فرد و الخطأ المعياري في قياس هذه القدرة، بالإضافة إلى إحصائيات المطابقة الداخلية و الخارجية لكل تقدير من هذه التقديرات. ولدى تفحص قيم إحصائيات المطابقة الداخلية والخارجية للأفراد، تبين وجود 95، 98، 146، 147 فردا لم يتلاءموا مع نموذج التقدير الجزئي (غير مطابقين) في طرق التصحيح الأربع على الترتيب، أي أن قيمة إحصائي المطابقة الخارجية أو الداخلية المقابلة لتقديرات قدرات هؤلاء الأفراد تقع خارج حدود المطابقة الملاءمة (0.7 إلى1.3). وهذا يعني أن الاستجابات الملاحظة لهؤلاء الأفراد تختلف عن الاستجابات المتوقعة وفقاً لنموذج التقدير الجزئي، كأن يجيب الفرد عن فقرة ما إجابة خاطئة رغم أن مستوى صعوبتها دون مستوى قدرته، أو أن يجيب عن فقرة ما إجابة صحيحة رغم أن مستوى صعوبتها يفوق مستوى قدرته.فمثلاً:الفرد(860) رغم تدني مستوى قدرته (-3.34 لوجت)، أجاب إجابة صحيحة عن الفقرة 21 التي مستوى صعوبتها (- 0.3 لوجت) وهي أعلى من مستوى قدرته.

ولمعرفة مؤشرات المطابقة الخاصة بالفقرات، فقد تم تقدير معلمة الصعوبة لكل فقرة و الخطأ المعياري في تقدير هذه المعلمة، بالإضافة إلى إحصائيات المطابقة الداخلية و الخارجية لكل تقدير من هذه التقديرات. ولدى تفحص قيم إحصائيات المطابقة الداخلية والخارجية للفقرات، تبين وجود 4، 3، 2، 1 فقرة لم تتلاءم مع نموذج التقدير الجزئي (غير مطابقة) في طرق التصحيح الأربع على الترتيب، أي أن قيمة إحصائي المطابقة الخارجية أو الداخلية المقابلة لتقديرات معالم الصعوبة لهذه الفقرات تقع خارج حدود المطابقة الملاءمة (0.7 إلى1.3). بمعنى أن الفقرات أصبحت أكثر ملاءمة لتوقعات النموذج في طريقة التصحيح الرابعة.

ثانياً:النتائج المتعلقة بالسؤال الثاني، "هل تختلف دقة تقدير معلمة القدرة باختلاف طريقة التصحيح؟"

للإجابة عن هذا السؤال تم استبعاد بيانات الأفراد والفقرات التي لم تتلاءم مع توقعات النموذج في أي طريقة من طرق التصحيح المعتمدة، وعددهم 231 فرداً ، و 5 فقرات . أي تم تحليل استجابات 642 فرداً عن 30 فقرة. وتم إيجاد تقدير لقدرة كل مفحوص في كل طريقة من طرق التصحيح المعتمدة، وكذلك الخطأ المعياري في تقدير هذه القدرة. و يبين جدول (11) الأوساط الحسابية والانحرافات المعيارية لتقديرات القدرة والاخطاء المعيارية في تقديرها لكل طريقة من طرق التصحيح الأربع.

ويتضح من جدول (11) أن الوسط الحسابي لتقديرات القدرة للأفراد يزداد بشكل مطرد عند الانتقال من طريقة التصحيح الأولى إلى الثانية إلى الثالثة إلى الرابعة، وقد بلغ الفرق في وسط التقدير لقدرات الأفراد بين طريقتي التصحيح الأولى والرابعة (1.54 لوجت). كما يتناقص الوسط الحسابي للخطأ المعياري في التقدير (بإعتبار تناقصه مؤشراً على دقة التقدير) بشكل مطرد بالانتقال من طريقة التصحيح الأولى إلى الثانية إلى الثالثة إلى الرابعة، حيث بلغ الفرق في الوسط الحسابي للخطأ المعياري بين طريقتي التصحيح الأولى والرابعة (0.292). ويلاحظ أن الخطأ المعياري في القياس بطريقة التصحيح الأولى يزيد على ضعفَي الخطأ المعياري في طريقة التصحيح الرابعة.

جدول (11): الأوساط الحسابية والانحرافات المعيارية لتقديرات القدرة والأخطاء المعيارية في تقديرها بحسب طريقه التصحيح المتبعة

طريقة التصحيح	الأولى		الثانية		الثالثة		الرابعة	
	القدرة	الخطأ المعياري	القدرة	الخطأ المعياري	القدرة	الخطأ المعياري	القدرة	الخطأ المعياري
الوسط الحسابي	0.90-	0.479	0.09-	0.257	0.32	0.203	0.64	0.187
الانحراف المعياري	1.09	0.087	0.64	0.029	0.53	0.037	0.49	0.044

ولفحص دلالة الفروق في دقة تقدير قدرات الأفراد، وفق نموذج التقدير الجزئي،بإختلاف طريقة التصحيح فقد تم استخدام تحليل التباين وفق تصميم القياسات المتكررة، للأخطاء المعيارية لهذه التقديرات، و يبين الجدول (12) نتائج هذا التحليل.

جدول (12): نتائج تحليل التباين لأثر طريقة التصحيح على مقدار خطأ التقدير لقدرات الأفراد

مصدر التباين		مجموع المربعات	درجة الحرية	متوسط المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
بين الأفراد		2.76367	641	0.00431		
داخل الأفراد	بين طرق التصحيح	**34.96665**	**3**	**11.6555498**	**4784.252**	**0.000**
	الخطأ	4.68488	1923	0.0024362		
	الكلي	39.65153	1926	0.02059		
الكلي		42.41519	2567	0.01652		

يتبين من الجدول (12) وجود فروق ذات دلالة إحصائية عند مستوى ($\alpha=0.01$) في الوسط الحسابي للأخطاء المعيارية لتقديرات قدرات الأفراد تعزى لطريقة التصحيح، وبهدف الكشف عن مواقع هذه الفروق بين طرق التصحيح، استخدم اختبار (LSD) للمقارنات الثنائية، ويبين الجدول (13) نتائج المقارنات الثنائية بين أوساط الأخطاء المعيارية لتقديرات قدرات الأفراد بحسب طريقة التصحيح .

الجدول (13): نتائج اختبار(LSD) للمقارنات الثنائية بين أوساط الأخطاء المعيارية لتقديرات قدرات الأفراد بحسب طريقة التصحيح

طريقة التصحيح		الرابعة	الثالثة	الثانية	الأولى
	المتوسط	0.187	0.203	0.257	0.479
الرابعة	0.187				
الثالثة	0.203	**0.016***			
الثانية	0.257	**0.070***	**0.054***		
الأولى	0.479	**0.291***	**0.275***	**0.222***	

* الفرق دال إحصائياً ($\alpha=0.05$)

ويلاحظ من الجدول (13) أن جميع الفروق الثنائية بين أوساط الأخطاء المعيارية دالة إحصائيا، عند مستوى ($\alpha=0.05$)، حيث جاءت هذه الفروق لصالح الطريقة الرابعة ثم الثالثة ثم الثانية.

ثالثاً: النتائج المتعلقة بالسؤال الثالث، "هل تختلف دقة تقدير معلمة الصعوبة باختلاف طريقة التصحيح؟"

للإجابة عن هذا السؤال ، تم الحصول على تقدير لمعلم الصعوبة لكل فقرة ككل ولكل خطوة من خطواتها ، في كل طريقة من طرق التصحيح الأربع، وكذلك الخطأ المعياري في تقدير هذا المعلم ويبين الملحق (1) عينة من مخرجات التحليل باستخدام برنامج Bigstep تخص الفقرة الأولى في كل طريقة من طرق التصحيح الأربعة.ويلاحظ هنا أن البرنامج يوفر قيمة لمعلمة صعوبة الفقرة ككل وهي بمثابة المتوسط الحسابي لمعالم خطواتها ، ونظراً لإختلاف عدد الخطوات من طريقة تصحيح إلى أخرى_ مما يجعل المقارنة بين الطرق غير ممكنة_ فقد تم الاهتمام بتقديرات الصعوبة للفقرات بصورتها الكلية(أي الاهتمام بمتوسطات معالم الفقرات ككل). و يبين الجدول (14) الأوساط الحسابية والانحرافات المعيارية لمعالم صعوبة الفقرات، والأخطاء المعيارية في تقديرها بحسب طريقة التصحيح .

جدول (14): الأوساط الحسابية والانحرافات المعيارية لمعالم صعوبة الفقرات، والأخطاء المعيارية في تقديرها بحسب طريقة التصحيح

طريقة التصحيح	الأولى		الثانية		الثالثة		الرابعة	
	الصعوبة	الخطأ المعياري	الصعوبة	الخطأ المعياري	الصعوبة	الخطأ المعياري	الصعوبة	الخطأ المعياري
الوسط الحسابي	0.00	0.105	0.00	0.057	0.00	0.042	0.00	0.039
الانحراف المعياري	1.19	0.019	0.57	0.008	0.45	0.006	0.45	0.006

يتضح من جدول (14) أن الوسط الحسابي لمعلمة صعوبة الفقرة في الاختبار ثابت ويساوي صفراً في جميع طرق التصحيح، لكن بإنحرافات معيارية مختلفة. كما أن الوسط الحسابي للخطأ المعياري في تقدير صعوبة الفقرة يقل بالانتقال من طريقة التصحيح الأولى إلى الثانية إلى الثالثة والرابعة.

ولفحص دلالة الفروق في دقة تقدير معالم الصعوبة، وفق نموذج التقدير الجزئي،بإختلاف طريقة التصحيح فقد تم استخدام تحليل التباين وفق تصميم القياسات المتكررة، للأخطاء المعيارية لهذه التقديرات، و يبين الجدول (15) نتائج هذا التحليل.

جدول (15): نتائج تحليل التباين للأخطاء المعيارية لمعلمة صعوبة الفقرة بحسب طريقة التصحيح

مصدر التباين		مجموع المربعات	درجة الحرية	متوسط المربعات	قيمة ف المحسوبة	الدلالة الإحصائية
بين الفقرات		0.00701	29	0.00024		
	بين طرق التصحيح	**0.08230**	**3**	**0.027431944**	**334.763**	**0.000**
داخل الفقرات	**الخطأ**	0.00713	87	0.000081944		
	الكلي	0.08943	90	0.00099		
الكلي		0.09643	119	0.00081		

ويتبين من نتيجة التحليل وجود فروق ذات دلالة إحصائية في الوسط الحسابي للأخطاء المعيارية في تقدير معالم صعوبة الفقرات تعزى لاختلاف طريقة التصحيح. وبهدف الكشف عن مواقع هذه الفروق بين طرق التصحيح، استخدم اختبار (LSD) للمقارنات الثنائية، ويبين الجدول (16) المقارنات الثنائية بين أوساط الأخطاء المعيارية لتقديرات معالم الصعوبة حسب طريقة التصحيح .

جدول (16): نتائج اختبار(LSD) للمقارنات الثنائية بين أوساط الأخطاء المعيارية لتقديرات صعوبة الفقرات بحسب طريقة التصحيح

طريقة التصحيح		الوسط	الرابعة	الثالثة	الثانية	الأولى
		0.039	0.042	0.057	0.105	
الرابعة	0.039					
الثالثة	0.042	**0.003***				
الثانية	0.057	**0.017***	**0.014***			
الأولى	0.105	**0.065***	**0.062***	**0.048***		

* الفرق دال إحصائياً ($\alpha=0.05$)

ويلاحظ من جدول (16) أن جميع الفروق الثنائية بين الأوساط الحسابية دالة إحصائيا عند مستوى ($\alpha=0.05$) ، حيث جاءت هذه الفروق لصالح الطريقة الرابعة ثم الثالثة ثم الثانية.

مناقشة النتائج

أولاً: مناقشة النتائج المتعلقة بمطابقة البيانات لنموذج التقدير الجزئي:

تشير النتائج إلى أن زيادة مراعاة طريقة التصحيح للمعرفة الجزئية كان له أثر في زيادة عدد الأفراد غير المطابقين للنموذج، فقد زاد عددهم من 95 فرداً في طريقة التصحيح الأولى، ليصبح 98 في الثانية، و 146 في الثالثة، و 147 فرداً في طريقة التصحيح الرابعة. ويمكن تفسير هذه النتيجة بدلالة تزايد عدد مستويات التصحيح للفقرة الواحدة بازدياد مراعاة طريقة التصحيح للمعرفة الجزئية؛إذ تشتمل الطريقة الأولى على مستويين، و الثانية على ثلاثة، والثالثة على أربعة، والرابعة على خمسة مستويات. فالأفراد في المستوى الأول(0) في طريقة التصحيح الأولى يتوزعون على المستويات الأربعة الأولى(0,1,2,3)في طريقة التصحيح الرابعة. ويتوقع مع تزايد عدد المستويات تزايد عدد العتبات ويرافق ذلك تباين في صعوبة هذه العتبات؛إذ قد تتضمن عتبات سهلة وأخرى صعبة وهذا قد يزيد من فرص أن يجيب الفرد عن فقرة ما إجابة خاطئة رغم أن مستوى صعوبتها دون مستوى قدرته، أو أن يجيب عن فقرة ما إجابة صحيحة رغم أن مستوى صعوبتها يفوق مستوى قدرته. وبالفعل فقد أجاب الأفراد ذوو الأرقام(865، 644، 149، 868، 862، 819)_ رغم تدني مستوى قدراتهم (-1.95، -1.95، -1.95، -2.22، -2.89، -2.89)_ عن الفقرة 14 إجابة صحيحة رغم ارتفاع مستوى صعوبتها (1.13 لوجت).

وتشير النتائج إلى أن زيادة مراعاة طريقة التصحيح للمعرفة الجزئية كان له أثر في نقصان عدد الفقرات غير المطابقة للنموذج، فقد نقص عدد الفقرات من 4 فقرات في طريقة التصحيح الأولى، ليصبح 3 في الثانية، و 2 في الثالثة، و فقرة واحدة في طريقة التصحيح الرابعة. بمعنى أن الفقرات أصبحت أكثر ملاءمة لتوقعات النموذج في طريقة التصحيح الرابعة. ويمكن تفسير هذه النتيجة بدلالة تزايد عدد مستويات التصحيح للفقرة الواحدة بازدياد مراعاة طريقة التصحيح للمعرفة الجزئية. ويتوقع مع تزايد عدد المستويات للفقرة أن تنزع الفقرات المتطرفة في موقعها على تدريج السمة نحو وسط التدريج وهذا يزيد من فرص ملائمتها للنموذج ؛فمثلاً الفقرة السهلة جداً في حالة وجود مستويين للإجابة تتوفر لها فرص وجود عتبات صعبة عندما يزداد عدد تلك المستويات والفقرة الصعبة جداً تتوفر لها فرص الحصول على عتبات سهلة.

ثانياً: مناقشة النتائج المتعلقة باختلاف دقة تقدير معلمة القدرة باختلاف طريقة التصحيح:

تشير النتائج إلى وجود فروق ذات دلالة إحصائية عند مستوى ($\alpha=0.01$) في الوسط الحسابي للأخطاء المعيارية لتقديرات قدرات الأفراد تعزى لطريقة التصحيح، وكانت هذه الفروق لصالح الطريقة الرابعة ثم الثالثة ثم الثانية. ويمكن تفسير ذلك بأن مراعاة المعرفة الجزئية في تصحيح فقرات الصواب والخطأ المتعدد، يقلل من التباين، الذي بدوره يقلل من الأخطاء المعيارية في قياسها، وبالتالي فإن تقديرات قدرة الأفراد قد أصبحت أكثر دقة بالتدرج في طرق التصحيح من الأولى إلى الثانية إلى الثالثة إلى الرابعة.أي أن أخطاء القياس تقل وبشكل واضح بالتدرج من طريقة التصحيح الأولى إلى الثانية إلى الثالثة فالرابعة. وهذه النتيجة تتفق مع النتائج التي حصل عليها كل من التقي(1992)، وشاهين(1993)، والشريفين (2003)، و رايت وماسترز (Wright and Masters,1982).

ثالثاً: مناقشة النتائج المتعلقة باختلاف دقة تقدير معلمة الصعوبة باختلاف طريقة التصحيح:

تشير النتائج إلى وجود فروق ذات دلالة إحصائية في الوسط الحسابي للأخطاء المعيارية في تقدير معالم صعوبة الفقرات تعزى لاختلاف طريقة التصحيح. وكانت الفروق لصالح الطريقة الرابعة ثم الثالثة ثم الثانية.أي أن أخطاء القياس تقل بشكل واضح بالتدرج من طريقة التصحيح الأولى إلى الثانية إلى الثالثة فالرابعة أي بزيادة عدد مستويات الإجابة. وهذه النتيجة تتفق مع النتائج التي حصل عليها كل من التقي(1992)، وشاهين(1993)، والشريفين (2003)، و رايت وماسترز (Wright and Masters,1982). ويمكن تفسير ذلك بأن مراعاة المعرفة الجزئية في تصحيح فقرات الصواب والخطأ المتعدد، يقلل من التباين، الذي بدوره يقلل من الأخطاء المعيارية في قياسها، وبالتالي فإن تقديرات صعوبة الفقرات قد أصبحت أكثر دقة بالتدرج في طريقة التصحيح من الأولى إلى الثانية إلى الثالثة إلى الرابعة، أي أن الأسئلة تصبح أكثر إعتدالاً في موقعها على تدريج السمة المقاسة كلما تعددت مستويات الإجابة عنها من خلال استخدام طرق التصحيح التي تراعي المعرفة الجزئية.

من خلال هذه النتائج مجتمعة يمكننا أن نصل إلى خلاصة مفادها أن طريقة التصحيح الأكثر مراعاة للمعرفة الجزئية، هي الأكثر دقة في قياس قدرات الأفراد، وتقدير قيم صعوبة الفقرات. لذلك يُنصح باستخدامها وتفضيلها على الطرق الأخرى. ويتوقع أن يترك استخدام المدرسين لهذه الطريقة أثراً ايجابياً لدى الطلبة يتمثل بشعور الارتياح للحصول على علامات أعلى في الاختبار نتيجة

مراعاة المعرفة الجزئية وقد يزيد ذلك من دافعيتهم للتعلم ، كما يتيح استخدامها فرص زيادة دقة تقدير قدرات الأفراد مما ينعكس إيجاباً على صدق تلك التقديرات وثباتها ، وهذا يجعل كل بديل من بدائل فقرة الصواب والخطأ المتعدد بمثابة فقرة اختبارية مما يعني زيادة في عدد الفقرات وبالتالي تحسن كل من صدق نتائج الاختبار وثباتها.

المصادر والمراجع

أمريزيق، خالد. (1992). *المقارنة بين اختبار الاختيار من متعدد واختبار الصواب والخطأ المتعدد من حيث الخصائص السيكومترية*. رسالة ماجستير غير منشورة، الجامعة الأردنية، الأردن.

التقي، أحمد محمد عيسى. (1992). *اللاتغيير في معالم قدرات الأفراد ودرجات صعوبة أسئلة المقال من خلال نموذج التقدير الجزئي وسلم التقدير كحالتين خاصتين من نماذج راش*.رسالة دكتوراه غير منشورة، الجامعة الأردنية، الأردن.

ثورندايك، روبرت وهيجن، إليزابيث. (1986). *القياس والتقويم في علم النفس والتربية* (ترجمة عبد الله زيد الكيلاني وعبد الرحمن عدس). نيويورك: جون وايلي وأولاده.

سوالمة، يوسف. (1994). اختبارات الصواب والخطأ المتعدد، الخصائص النسبية وعدد البدائل الصحيحة. *أبحاث اليرموك، سلسلة العلوم الإنسانية والاجتماعية*،10 (3)، 573-595.

شاهين، محمد. (1993). *اثر عدد خطوات تصحيح الأسئلة المقالية في تقدير قدرات الأفراد وقيم صعوبة الأسئلة باستخدام نموذج التقدير الجزئي*. رسالة دكتوراه غير منشورة، الجامعة الأردنية، الأردن.

الشريفين، نضال. (2003). *مدى تحقيق معايير الفاعلية في معادلة اختبارين أحدهما ثنائي التدريج والآخر متعدد التدريج وفق نماذج النظرية الكلاسيكية والنظرية الحديثة في القياس*. رسالة دكتوراه غير منشورة، جامعة عمان العربية للدراسات العليا، الأردن.

عدس، عبد الرحمن. (1988). *دليل المعلم في بناء الاختبارات التحصيلية. المنظمة العربية للتربية والثقافة والعلوم*، تونس: إدارة البحوث التربوية.

العلي، محمد محمود البشير. (2004) . *فاعلية نموذج التقدير الجزئي في بناء بنك أسئلة من فقرات متعددة الخطوات في مادة الكيمياء في الصف الثاني الثانوي العلمي*. رسالة دكتوراه غير منشورة ، جامعة عمان العربية للدراسات العليا، الأردن.

المومني، محمد ضيف الله. (1996). *مقارنة أثر اختبارات الاختيار من متعدد، واختبارات الصواب والخطأ المتعدد في الاحتفاظ بقواعد اللغة الإنجليزية*. رسالة ماجستير غير منشورة، جامعة اليرموك، الأردن.

Albanese, M. A., and Sabers, D. L. (1988). Multiple true – false items: A study of interterm correlations scoring alternatives, and reliability estimation. *Journal of Educational Measurement*, *25* (2), 111 -123.

Brad, W. (2003). *Scoring multiple true false items: A comparison of summed scores and response pattern scores at item and test levels.*(ERIC Document Reproduction Service No. ED476 148).

Ebel, R. L., and Frisbie. D. A. (1986). *Essentials of educational measurement* (Forth edition). New Jersey: Prentice J Hall.

De Gruijter, D. and Van der Kamp, L.(2003). *Statistical test theory for education and psychology.*

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31* (4), 295 – 311.

Downing, S.M., Baranoski, R. A., Gross, L. J., and Norcni. J. J. (1995). Item Type and Cognitive Ability Measured The Validity Evidence for Multiple True-False Items in Medical Specialty Certification. *Applied Measurement in Education, 8* (2), 187-197.

Dudley, A. (2006). Multiple dichotomous scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing,23*(2),198-228.

Erickson, I. L. (1988). The effect of item type and test format on level of difficulty and mean item discrimination index in achievement test for nurses. *Dissertation Abstract International*, *49* (6), 1437.A.

Frisbie, D. A. (1992). The Multiple True False item formats A status Review. *Education Measurement: Issues and Practice, 11* (4), 21 – 26.

Frisbie, D.A., and Druva, C.A. (1987). Estimating the reliability of multiple true- false test. *Psychological Abstracts*,*74* (7), 2011.

Frisbie, D .A., and Sweeny, D. C. (1982). The Relative Merits of Multiple True – False Achievement Tests. *Journal of Educational Measurement, 19* (1), 29-35.

Hambleton. R.K., and Swaminathan, H. (1985). *Item response theory: Principles and application,* Boston: Kluwer Nijhoff publishing.

Han, K. (n. d.). *Item response models used within WinGen*. Retrieved September 18, 2008 from:

Masters, N.G.(1982). A Rasch model for partial credit scoring. *Psychometrika*. *47* (2), 149-174.

Randall, E. and Ching-Fung, Si. (2004). Ability Estimation Under Different Item Parameterization and Scoring Models*. International Journal of Testing. 4 (2), 137-181.* Retrieved 14/9/2004 from EBSCO host. (AN 14529295).

Tsai, Fu- Ju.and Suen, Hoik. (1993). A Brief report on a comparison of six scoring methods for multiples true false item. *Educational and Psychological Measurement*, *53,* 399-404.

Wright, B. D. and Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press

http:/www.umass.edu/remp/software/wingen/models.

Kreiter, C. D. and Frisbie, D. A. (1989). Effectiveness of multiple true-false item*, Applied Measurement in Education, 2,* 207-216.

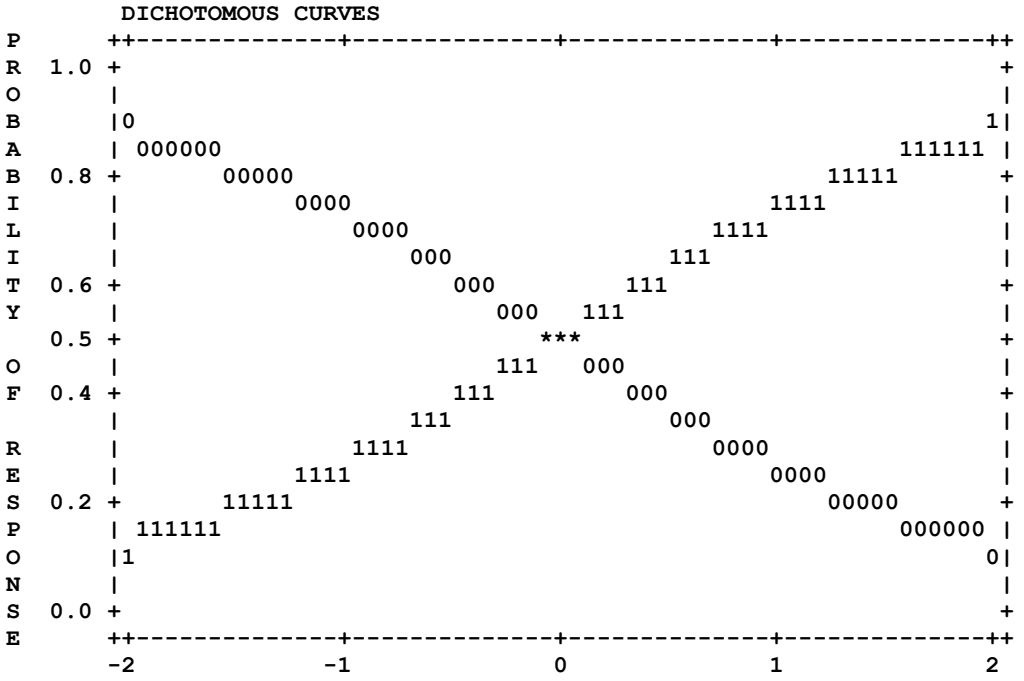
Lee ,G., Kolen, M.J., Frisbie, D,A. and Ankenmann ,R.D.(2001). Comparison of dichotomous and polytomous item response models in equating scores from testes composed of testlets*. Applied Psychological Measurement, 25* (4), 357 –372.

Lustina, M. J.(2004). *A comparison of Andrich's rating scale model and Rast's successive intervals model*. Ph.D. Dissertation. The University of Texas at Austin. Retrieved October 1,2008 from www.lib.utexas.edu/etd/d/2004.

ملحق (1)

مخرجات التحليل بحسب طريقة التصحيح

```
TABLE 21.1 ANALYZED: 642 PERSONS, 30 ITEMS, Scoring Method 1
--------------------------------------------------------------------
SUMMARY OF MEASURED STEPS   ITEM  NUMBER:     1
ITEM  DIFFICULTY OF  -1.25 ADDED TO MEASURES
+-----------------------------------------------------+
|CATEGORY OBSERVED|  MEASURE   | COHERENCE|INFIT OUTFIT|
| LABEL    COUNT  |AVERGE  EXP.| M->C C->M|  MNSQ  MNSQ|
|-----------------+------------+----------+------------|
|   0         277 | -1.28 -1.43|  56%  56%|  1.15  1.21|
|   1         365 | -0.60 -0.49|  67%  66%|  1.13  1.15|
+-----------------------------------------------------+
AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.
M->C = Does Measure imply Category?   C->M = Does Category imply Measure?
```



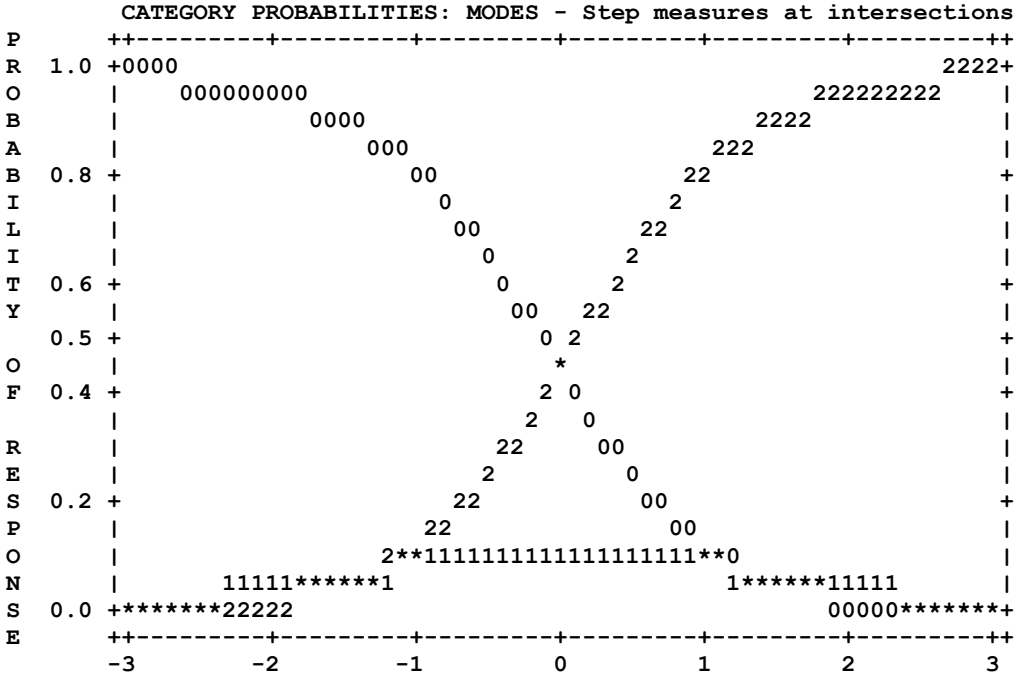
Scoring Method 2 ,SUMMARY OF MEASURED STEPS ITEM NUMBER: 1
ITEM DIFFICULTY OF -0.46 ADDED TO MEASURES

CATEGORY LABEL	OBSERVED COUNT	MEASURE AVERGE	MEASURE EXP.	COHERENCE M->C	COHERENCE C->M	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
0	211	-0.31	-0.45	58%	4%	1.20	1.32	NONE
1	66	-0.24	-0.20	11%	72%	1.16	1.34	1.29
2	365	0.06	0.14	71%	37%	1.16	1.18	-1.29*

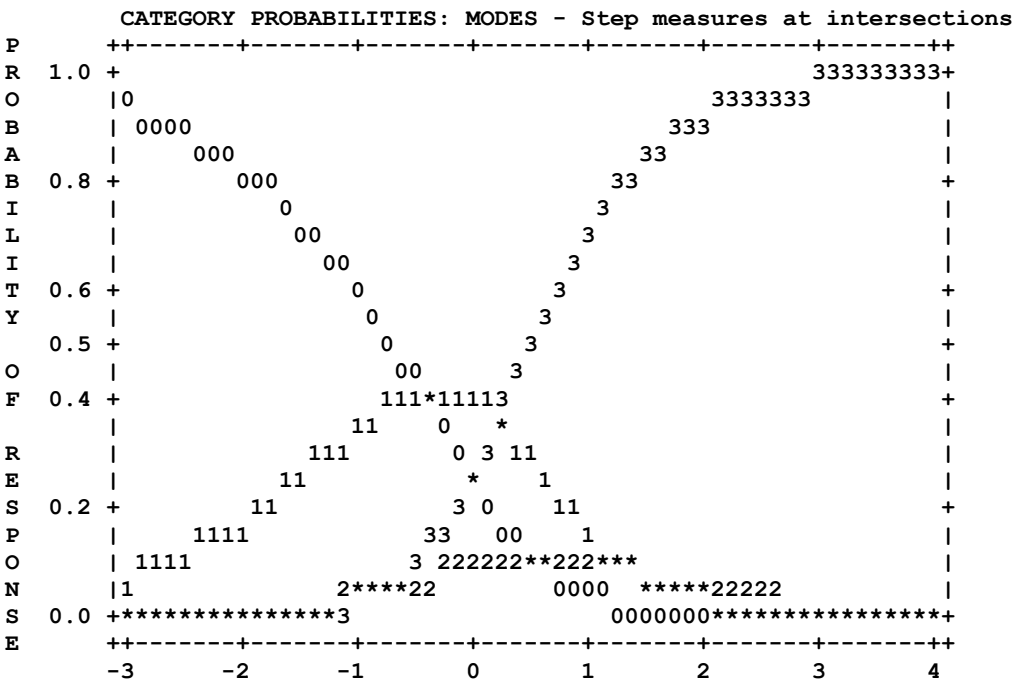
AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.
M->C = Does Measure imply Category? C->M = Does Category imply Measure?

CATEGORY LABEL	STEP MEASURE	STEP S.E.	SCORE-TO-MEASURE AT CAT.	----ZONE----		THURSTONE THRESHOLD
0	NONE		(-1.59)	-INF	-1.09	
1	0.83	0.09	-0.46	-1.09	0.17	-0.60
2	-1.75	0.09	(0.67)	0.17	+INF	-0.32

CATEGORY PROBABILITIES: MODES - Step measures at intersections



```
Scoring Method 3 ,SUMMARY OF MEASURED STEPS        ITEM  NUMBER:  1
ITEM  DIFFICULTY OF  -0.40 ADDED TO MEASURES
+-------------------------------------------------------------------+
|CATEGORY OBSERVED|  MEASURE    | COHERENCE|INFIT OUTFIT|  STEP    |
| LABEL    COUNT  |AVERGE  EXP.| M->C C->M|  MNSQ  MNSQ|CALIBRATN|
|-----------------+------------+----------+------------+---------|
|   0          62 |  0.13 -0.05|   0%   0%|  1.21  1.46|  NONE   |
|   1         149 |  0.17  0.09|  32%  10%|  1.14  1.20|   -0.45 |
|   2          66 |  0.21  0.26|  10%  68%|  1.34  1.41|    1.39 |
|   3         365 |  0.44  0.49|  72%  31%|  1.14  1.15|   -0.94*|
+-------------------------------------------------------------------+
+-----------------------------------------------------+
|CATEGORY   STEP    STEP |  SCORE-TO-MEASURE   |THURSTONE|
| LABEL    MEASURE  S.E. | AT CAT. ----ZONE----|THRESHOLD|
|----------------------+---------------------+---------|
|   0      NONE          |( -2.12) -INF   -1.40|         |
|   1       -0.86   0.14 |  -0.71  -1.40  -0.29|   -1.10 |
|   2        0.99   0.09 |   0.11  -0.29   0.61|   -0.11 |
|   3       -1.34   0.09 |(  1.06)  0.61  +INF |    0.11 |
+-----------------------------------------------------+
```

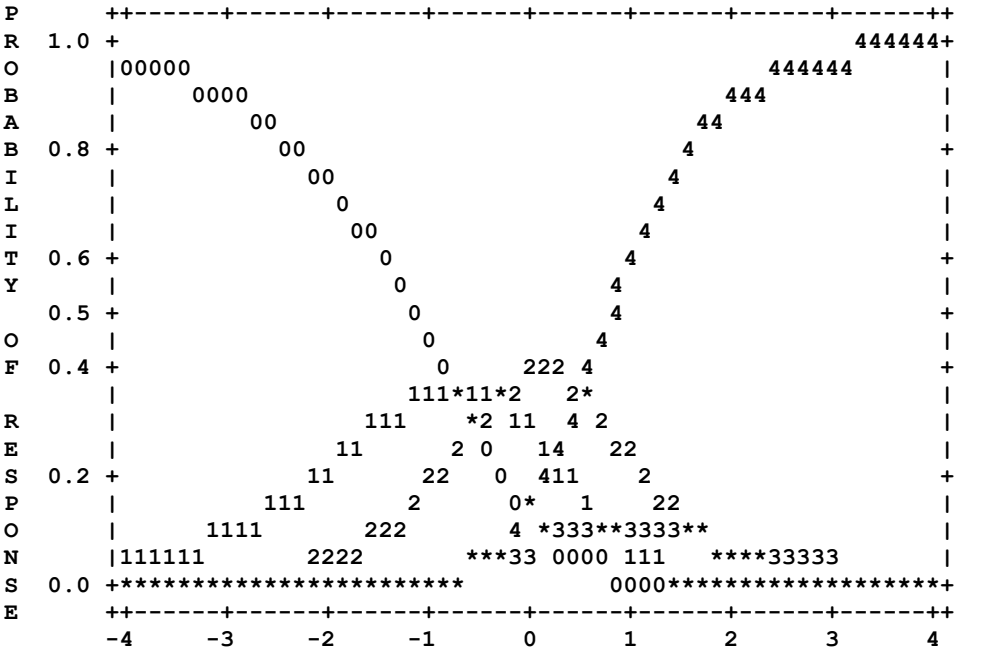


Scoring Method 4, SUMMARY OF MEASURED STEPS ITEM NUMBER: 1
ITEM DIFFICULTY OF -0.39 ADDED TO MEASURES

CATEGORY LABEL	OBSERVED COUNT	MEASURE AVERGE	MEASURE EXP.	COHERENCE M->C	COHERENCE C->M	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
0	12	0.37	NONE	0%	0%	1.15	1.34	NONE
1	50	0.49	0.33	0%	0%	1.20	1.47	-0.77
2	149	0.51	0.45	25%	8%	1.12	1.17	-0.31
3	66	0.55	0.58	10%	68%	1.45	1.36	1.72
4	365	0.75	0.79	71%	29%	1.14	1.12	-0.64*

CATEGORY LABEL	STEP MEASURE	STEP S.E.	SCORE-TO-MEASURE AT CAT.	----ZONE----		THURSTONE THRESHOLD
0	NONE		(-2.57)	-INF	-1.87	
1	-1.16	0.30	-1.11	-1.87	-0.62	-1.54
2	-0.70	0.14	-0.24	-0.62	0.10	-0.58
3	1.33	0.09	0.45	0.10	0.93	0.22
4	-1.03	0.09	(1.38)	0.93	+INF	0.44

CATEGORY PROBABILITIES: MODES - Step measures at intersections



PERSON [MINUS] ITEM MEASURE