كلية التربية

كلية معتمدة من الهيئة القومية لضمان جودة التعليم

إدارة: البحوث والنشر العلمي ( المجلة العلمية)

=======

# أثر اختلاف طريقة المعادلة وطرق تقدير الدرجات وقواعد صياغة الفقرات على دقة تقدير معالم الفقرات وقدرات الأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي ثلاثي البارامتر

إعـــداد

**د/ياسر عبدالله حفني حسن**

أستاذ علم النفس التربوي المساعد

كلية التربية بقنا – جامعة جنوب الوادي

﴿ المجلد الخامس والثلاثون – العدد السابع – يوليو ٢٠١٩م ﴾

http://www.aun.edu.eg/faculty_education/arabic

## ملخص الدراسة

هدفت الدراسة إلى بحث أثر اختلاف طريقة المعادلة (المتوسط/المتوسط، المتوسط / الانحراف المعياري) وطرق تقدير الدرجات (التقليدية، التجريبية، وطريقة الاحتمال المقترح للإجابة الصحيحة) وقواعد صياغة فقرات الاختبار (المحكم، المخالف) على دقة تقدير معالم الفقرات وقدرات الأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي ثلاثي البارامتر، وتكونت عينة الدراسة من ١٥٠٠ طالباً وطالبة تراوحت أعمارهم بين (٢٠.٤ – ٢١,٧) سنة، من طلاب كلية التربية جامعة أم القرى بمكة المكرمة، تم اختيارهم بالطريقة العشوائية الطبقية، ولتحقيق أهداف الدراسة والإجابة عن تساؤلاتها قام الباحث بإعداد نموذجي اختبار لمقرر الاختبارات والمقاييس من نوع الاختيار من متعدد ذو الأربعة بدائل، وتم معالجة النتائج وتحليلها باستخدام البرامج الإحصائية IRTEQ – XCalibre (4.1.7) – SPSS(22) ، وتوصل الباحث إلى النتائج التالية: اختلاف التقديرات لكل من النظرية الكلاسيكية والنموذج اللوجستي ثلاثي البارامتر، فمن منظور القياس الكلاسيكي: كان متوسط الصعوبة والتمييز لفقرات الاختبار المحكم البناء أعلى من متوسط صعوبة وتمييز فقرات الاختبار المخالف لقواعد الصياغة، ومن منظور النموذج اللوجستي ثلاثي البارامتر: أظهرت النتائج أن الاختبار المحكم أكثر كفاءة وفاعلية من الاختبار المخالف عند مستويات القدرة المختلفة، وأن فقرات الاختبار المحكم كانت أكثر دقة في تقدير قدرة الأفراد من الاختبار المخالف، وأن تحليل الفقرة في ضوء نظرية الاستجابة للفقرة كان أكثر دقة من النظرية الكلاسيكية في تقدير معلمة الصعوبة والتمييز والتخمين، وكانت أكثر طرق تقدير الدرجات الكلاسيكية ارتباطاً بالنموذج اللوجستي ثلاثي البارامتر في تقدير قدرات الطلاب وصعوبة وتمييز الفقرات، الطريقة التقليدية ثم الطريقة التجريبية ثم طريقة الاحتمال المقترح للإجابة الصحيحة، وأشارت النتائج إلى أن قيم التحيز وجذر متوسط مربع الخطأ، تقل مع ازدياد حجم العينة وطول الاختبار، فكلما زاد حجم العينة، وطول الاختبار زادت دقة المعادلة، وفي ضوء محكي التحيز وجذر متوسط مربع الخطأ، تعتبر طريقة (المتوسط/المتوسط) أكثر دقة في معادلة درجات الاختبارات من طريقة (المتوسط/الانحراف المعياري) وفق النموذج اللوجستي ثلاثي البارامتر.


**الكلمات المفتاحية:** طريقة المعادلة، طرق تقدير الدرجات، قواعد صياغة الفقرات، معالم الفقرات وقدرات الأفراد، القياس الكلاسيكي، النموذج اللوجستي ثلاثي البارامتر.

The study aimed at investigating the effect of different functioning Method (Mean & Mean Method, Mean & Sigma Method), Methods of Scoring (the conventional method, the experimental Method and the method of probability assigned to the correct answer), and the rules of crafting items (the well-structured test, the ill structured test) on the accuracy of estimating the parameters of items and the abilities of individuals in the light of Classical Measurement and the three-Parameter Logistic Model. The sample of the study consisted of (1500) male and female students aging from (20.4-21.7) years, from the faculty of Education, at Umm Al-Qura University, who have been chosen stratified randomly. In order to achieve the aims of this study and to answer its questions, the researcher prepared two test modules for the course of tests and measurements of multiple choices type with four alternatives. Data were analyzed through using (SPSS 22, XCalibre 4.1.7, IRTEQ). The results indicated differences between classical theory and three parameters logistic model. The Classical perspective: The difficulty and discrimination mean of well-structured test items was higher than the difficulty and discrimination mean of ill structured test items. Three parameters logistic model perspective: The well-structured test is more efficient and effective than the violated test at different ability levels. The well-structured test is more accurate in estimating the parameters of individuals than the violated test, and the item analysis in the light of item response theory is better than classical theory of the test regarding parameter difficulty, discrimination and guessing. The Conventional method was the most related method to the three parameters logistic model among the other classical methods in estimating the abilities of the students, the difficulty and discrimination of items, then the experimental method, followed by the method of

proposed probability of the answer. The results showed that bias values and root mean square errors decreased with the increase of sample size and test duration. The bigger sample size and the longer test, the more accurate the equation becomes. for the effect of three parameter model, in light of bias simulation and Root Mean Square Error, Mean & Mean Method is considered better than Mean & Sigma method in equating test score.

Key Words: Functioning Method, Methods of Scoring, the Rules of Crafting Items, The Parameters of Items and Individuals, Classical Measurement, Three-Parameter Logistic Model.

## مقدمة الدراسة:

يعد القياس والتقويم عنصراً أساسياً في العلمية التعليمية والتربوية، وأكثرها تأثيراً في تقدمها وتطوير مكوناتها ورفع كفاءتها، وتعتبر الاختبارات النفسية والتربوية من أهم أدوات القياس النفسي والتربوي، والتي تزود المؤسسات التعليمية والتربوية ببيانات كمية تتيح للتربويين فهم الظاهرة التربوية، وتزود القائمين على العملية التعليمية بمقدار التقدم في مستوى التحصيل الدراسي للطلاب، ومدى تحقيقهم للأهداف التعليمية، وبالتالي يمكن اتخاذ بعض القرارات، التي يعتمد سلامتها على نوع ودقة المعلومة والأساليب المستخدمة في تفسير وتحليل النتائج التي تزودنها بها تلك الاختبارات.

فالقياس والتقويم يتضمن إجراءات وطرقا منهجية لتقرير المدى الذي تعد فيه التفسيرات والإجراءات التي تتخذ ضمن إطار الميدان التربوي والنفسي مبرره وكافية، وفي هذا المجال هناك إطاران متنافسان في نظرية القياس، هما نظرية القياس الكلاسيكية ونظرية الاستجابة للمفردة الاختبارية، وتعد كلتا النظريتين غايةً في الأهمية في تقديم تقييمات مختلفة حول كل من فقرات الاختبار والاختبار ككل (Cappelleri, Jason & Hays, 2014; Coggins, Kim & Briggs, 2017).

وتعد النظرية الكلاسيكية للاختبارات Classical Test Theory (CTT) من أقدم النظريات، التي استخدمت في تطوير الاختبارات وبنائها لسنوات طويلة منذ أوائل القرن العشرين، والتي استخدمت فيها العديد من الدراسات في عملية بناء وتصميم مختلف الاختبارات النفسية والتربوية وتحليلها وتفسير البيانات المستمدة منها (Gregory, 2014; Hambelton & Swaminthan, 1985).

وعلى الرغم من أن نظرية القياس الكلاسيكية تعتمد على مسلمات بسيطة وتتطابق بسهولة مع بيانات الاختبارات الفعلية، فإن لديها جوانب قصور منها انعدام خطية القياس، وعدم وجود وحدة ثابتة للقياس بالإضافة إلى القياس في أكثر من بعد، كما أن من أهم مشكلات القياس الكلاسيكي أن معالم المفردات والأفراد تصبح محكومة بعينة المفحوصين التي طبق عليها الاختبار، فعندما تكون عينة المفحوصين مرتفعة في مستوى القدرة نحصل على صعوبة منخفضة للمفردات، وإذا كانت عينة المفحوصين منخفضة في مستوى القدرة نحصل على صعوبة مرتفعة للمفردات، وبالتالي لا يمكن التنبؤ بأداء المفحوصين على مفردة اختبارية معينة (Hambleton & Jones, 1993; Ojerinde, 2013).

وقد أجريت العديد من الدراسات والبحوث من أجل التغلب على جوانب القصور في النظرية الكلاسيكية، والوصول إلى قياس موضوعي يماثل القياس الفيزيائي، وقد أسفرت هذه الجهود في ظهور نظرية السمات الكامنة، والتي عرفت فيما بعد بنظرية استجابة المفردة، إذ أنه يمكن التنبؤ بأداء المفحوصين على اختبار نفسي أو تربوي بواسطة سمة أو قدرة تميز هؤلاء المفحوصين، والتي أطلق عليها السمات الكامنة، وتتميز هذه النظرية في أن تقدير معالم المفردات من صعوبة وتمييز، وتخمين مستقل عن قدرة المفحوصين التي استخدمت في تقدير هذه المعالم، وأن تقدير قدرات المفحوصين يكون مستقلاً عن عينة المفردات المستخدمة في عملية التقدير(Bond & Fox, 2015; Natarajan, 2009).

وتقوم الفكرة الأساسية لنظرية الاستجابة للمفردة Item Response Theory (IRT) على اشتقاق قيم تقديرية للسمات التي تنطوي عليها مجموعة من الاستجابات لمجموعة من المفردات، وعادة يفترض أن السمة المقاسة هي قدرة معينة أو خاصية من خصائص الفرد الذى يختبر بها، بحيث لا توجد علاقة منتظمة بين مستويات السمة المقاسة لدى أفراد مختلفين واحتمالات الاستجابة الصحيحة لمفردات مختلفة (أحمد محمد التقي، ٢٠١٣؛ صلاح الدين محمود علام، ٢٠٠٥).

وقد تميزت نظرية الاستجابة للمفردة بقوتها على المستوى التنظيري وقدرتها على إعطاء تقديرات أفضل للمستويات الحقيقية للأفراد على متصل السمة، كما أنها توفر تقديراً للقدرة مستقلاً عن خصائص العينة وبمستوى قياس يحقق مميزات القياس ذي الفئات المتساوية (أمينة محمد كاظم، ١٩٨٨؛ صلاح الدين محمود علام، ٢٠٠٥؛ عبدالرحمن عبدالله النفيعي،٢٠١٢).

وقد انبثق عن نظرية الاستجابة للمفردة مجموعة من النماذج التي جميعها تهدف إلى تحديد العلاقة بين أداء الفرد في الاختبار وهو ما يمكن ملاحظته ملاحظة مباشرة وبين السمات أو القدرات التي تكمن وراء هذا الأداء وتفسره، ومن أهم هذه النماذج وأكثرها شيوعاً نموذج راش Rasch Model أحادي البارامتر، ونموذج لورد Lord Model ثنائي البارامتر، ونموذج بيرنبوم Birnbaum Model ثلاثي البارامتر (صلاح الدين محمود علام، ٢٠٠٥؛ Magis & Raîche, 2012; Penfield, 2014).

ويعد النموذج اللوجستي الثلاثي البارامتر هو النموذج الكامل والحالة العامة بين النماذج البارامترية الثلاثة السابقة حيث يوصف المنحنى المميز للمفردة وفق هذا النموذج من خلال ثلاثة بارامترات تشتق رياضياً من البيانات الإمبريقية وهي بارامتر صعوبة الفقرة، وبارامتر التمييز، وبارامتر التخمين، ويتميز هذا النموذج عن النموذجين الآخرين بمراعاة عامل التخمين وهو ما يتوقع أن يحصل في كثير من اختبارات الصح والخطأ أو الاختيار من متعدد مما يمكن أن يؤثر على دقة تقدير قدرات الطلاب في هذه الأنواع من الاختبارات (آن أناستازي، سوزانا أوربينا، ٢٠١٥؛ حمدي يونس أبو جراد، ٢٠١٧؛ معين سلمان النصراوين، محمد وليد موسى البطش، ٢٠١٨).

وتعتبر الاختبارات التحصيلية وسيلة من الوسائل المهمة التي يُعوَل عليها قياس وتقويم قدرات الطلاب ومعرفة مستواهم التحصيلي والتأكد من مدى تحقق الأهداف التعليمية المختلفة، وتعدّ الاختبارات التحصيلية من نوع الاختيار من متعدد أكثر أشكال التقويم انتشاراً في التربية، ومما زاد في انتشار هذا النوع من الاختبارات وتفوقها على كافة أشكال الفقرات الموضوعية الأخرى، إذ يمكن بواسطتها قياس أهداف بسيطة وأخرى مركبة في مختلف المواضيع الدراسية، وعلى اختلاف المراحل التعليمية (باسل خميس أبو فودة، نجاتي أحمد يونس،٢٠١٢ ;Campbell, 2015; Slepkov & Godfrey, 2019).

وتُعد صياغة فقرات الاختيار من متعدد عملاً فنياً وإبداعياً، وثمة من أعتبر ذلك فناً وعلماً في آن واحد، ولذا يتوجب توزيع قواعد صياغة الفقرات على معديّ تلك الفقرات إذا كانوا من غير المتخصصين، وتكثيف البرامج التدريبية والتطبيقية المتعلقة بجودة صياغة فقرات الاختيار من متعدد وفق الإرشادات الخاصة بها من قِبل خبراء بناء الأسئلة والمتمرسين عليها، ولا يخلو كتاب في القياس والتقويم دُوّن من قِبل متخصصي هذا الفن من إرشادات تتعلق بصياغة فقرات الاختبارات التحصيلية من نوع الاختيار من متعدد (Aiken & Groth-Marnat, 2006; Breakall, Randles & Tasker, 2019).

ويشير باسل خميس أبو فودة (٢٠١٤) إلى مجموعة من القواعد والإرشادات في كتابة فقرات اختبار الاختيار من متعدد منها: التأكد من أن الجذر يطرح مشكلة محددة وواضحة، وجعل البدائل قصيرة ما أمكن، وتجنب صيغ النفي، والتأكد من أن بدائل الإجابة الخطأ تؤلف إجابات معقولة ظاهرياً، وأن تكون جذابة للمفحوصين الذين تنقصهم المعرفة، وأن لا يتضمن الاختبار فقرات تعتمد في إجابتها على فقرات أخرى، وتجنب الخداع والغموض في جذر الفقرة وبدائلها، وجعل بدائل الفقرة متساوية في طولها.

وتأتي أهمية دراسة قواعد صياغة فقرات الاختيار من متعدد بالأثر المتوقع لها في أداء الفرد وهذا ما يراه Hambleton & Swaminathan, (1985) من أن مستوى الأداء على الفقرة أو الاختبار يتوقف على خصائص الفقرة أو الاختبار وعلى خصائص الفرد، ولقد أكد (Gleason, Alley & Baker, 2010; Slepkov & Godfrey, 2019) على ضرورة فحص كل فقرة من أجل تحديد ما إذا كانت الفقرة تتضمن انتهاكاً للقواعد أم لا ؛ مما قد يؤثر ذلك سلباً على الخصائص السيكومترية للمفردة.

وتعد طرق تقدير الدرجات Methods of Scoring هي القاعدة التي يعطى في ضوئها قيماً كمية تعكس الدرجة المستحقة للطالب في كل فقرة من فقرات الاختبار، فهي أساليب وإجراءات تتعلق بتعليمات تطبيق الاختبار وتصحيحه يتم من خلالها تقدير درجة المفحوص على كل فقرة من فقرات الاختبار وفق نظام رقمي محدد يختلف من طريقة لأخرى (Lesage, Valcke & Sabbe, 2013; Sočan, 2015) ، وقد اشتملت الدراسة الحالية على ثلاث طرق لتقدير درجات فقرات الاختيار من متعدد وهي: الطريقة التقليدية، والطريقة التجريبية، وطريقة الاحتمال المقترح للإجابة الصحيحة.

كما يعد الهدف الأساسي من استخدام طرق تقدير الدرجات في الفقرات الموضوعية هو الوصول إلى أفضل تقدير لقدرات الأفراد، عن طريق الحصول على أكبر قدر من المعلومات الكمية، وتقليل خطأ القياس إلى أقل حد، وهو ذات الهدف النهائي الذي تهتم به جميع نظريات القياس ومنها نظرية استجابة المفردة والتي جاءت كثورة في مجال القياس النفسي والتربوي حيث قدمت الحلول للعديد من أوجه القصور في القياس الكلاسيكي، كما أصبحت وسيلة أساسية وشائعة في بناء وتطوير الاختبارات حيث أنها تقدم بديلاً عن نظرية القياس الكلاسيكية في تقدير معالم الأفراد والمفردات بأقل قدر من الخطأ (Ndalichako, & Rogers, 1997; Vanderoost, Janssen, Eggermont, Callens & De Laet, 2018 ).

وتعتبر معادلة الاختبار من أهم تطبيقات القياس والتقويم، ففي العديد من المواقف الاختبارية تستدعي الحاجة إلى تطبيق عدة صور من الاختبار الواحد بهدف الحفاظ على سرية الاختبار، كما أن بعض اختبارات القبول تحتاج تطبيق صور متعددة من الاختبار ومقارنة الدرجات، التي يحصل عليها المفحوصون المطبق عليهم صوراً من نفس الاختبار (Inal & Anil, 2018; Kolen & Brennan, 2014)، والمقصود بالمعادلة هو تحويل الدرجات على صور الاختبارات إلى مقياس مشترك موحد، بحيث تصبح القياسات المستمدة من درجات كل من الصورتين متكافئة بعد إجراء هذا التحويل، وتهدف المعادلة إلى إزالة فروقات الصعوبة؛ بحيث يمكن إجراء مقارنات بين المفحوصين المطبق عليهم نفس الاختبار بشكل متبادل؛ ولذلك يتطلب إجراء معادلة ذات مستوى عالٍ من الدقة.

وهناك العديد من الطرق لإجراء المعادلة، فإما أن تكون الطريقة المستخدمة في المعادلة تعتمد على النظرية الكلاسيكية في القياس (CTT) أو نظرية استجابة المفردة (IRT)، ولكل منهما طرقها الخاصة في المعادلة، ففي النظرية الكلاسيكية تكون من خلال طريقة المعادلة الخطية Linear Equating، أو طريقة المعادلة المئينية Equipercentile Equating، أو طريقة المعادلة الانحدارية Regression Equating، أو من خلال معادلة المتوسط الحسابي Mean Equating؛ أما طرق المعادلة التي تعتمد على نظرية استجابة المفردة تكون باستخدام معادلة الدرجات الحقيقية True–Score Equating، أو باستخدام معادلة درجات القدرة Ability Score Equating، أو باستخدام معادلة الدرجات المشاهدة Observed Score Equating (Angoff, 1987; Kolen & Brennan, 2014; Öztürk–Gübes & Kelecioglu, 2016) .

وقد أشار Zhonghua (2010) أنه عند إجراء المعادلة وفق تصميم المفردات المشتركة يمكن وضع تقديرات بارامترات المفردة غير المعلومة المشتقة من صور الاختبار على مقياس مشترك من خلال ثلاثة طرق: ربط التدريج المنفصل The linking Separate Calibration (LSC)، التدريج المتلازم The Concurrent Calibration ، تدريج المعلمة الثابتة The Fixed Parameter Calibration (FPC) ، وتعتمد طريقة ربط التدريج المنفصل على التحويل الخطي؛ حيث يمكن حساب معاملا التحويل (A, B) بعدة طرق من أهمها: طريقة المتوسط/ المتوسط، طريقة المتوسط/الانحراف المعياري.

يتضح مما سبق أن الدراسة الحالية تُعد محاولة للتعرف على أثر اختلاف طريقة المعادلة وطرق تقدير الدرجات وقواعد صياغة الفقرات على دقة تقدير معالم الفقرات وقدرات الأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي ثلاثي البارامتر، حيث إن هذا الميدان فى حاجة إلى مزيد من الدراسات والبحوث، والدراسة الحالية تعد بمثابة دعوة فى هذا الاتجاه .

**مشكلة الدراسة:**

شهد منتصف القرن العشرين تطورات جوهرية في منهجيات القياس النفسي وطرق تصميم الاختبارات والمقاييس وتقنيات تحليل البيانات المستمدة منها، من خلال ظهور ما يسمى بنظرية الاستجابة للمفردة (IRT) التي أُعتبرت بمثابة الثورة والمستقبل الزاهر للقياس النفسي والتربوي، (Anstasi & Urbena, 2005)، حيث قدمت إطاراً مرجعيًّا لبناء المقاييس النفسية والتربوية، وطريقة تفسير الدرجات على هذه الاختبارات مقارنة بما قدمته النظرية الكلاسيكية في القياس (Ojerinde, 2013; Van der Linden, 2009; 2010) ، وبذلك تحققت إلى حد بعيد الموضوعية المنشودة للقياس لمعالجة نواحي القصور التي ظهرت في أساليب القياس المعتمدة على نظرية القياس التقليدية.

وتعد نظرية استجابة الفقرة ثورة في عالم القياس النفسي والتربوي وذلك لكونها تراعي عدد من المتغيرات التي أغفلتها نظرية القياس الكلاسيكي مما أثمر عن قياس أكثر دقة وموضوعية في تقدير قدرات الأفراد، وقد أكد هذا التفوق عدد من الدراسات التي قارنت بين هاتين النظريتين كدراسات (Adedoyin, 2010; Ainol & Noor, 2006; Ayhan, 2015; Coggins et al., 2017; Hambleton & Jones, 1993; Reise & Revicki, 2015) .

ونظراً للانتشار الواسع لاستخدام اختبارات الاختيار من متعدد، وخاصة في المجال التعليمي؛ لما تتمتع به هذه الاختبارات من ميزات كثيرة؛ حيث أن هذه الاختبارات لديها القدرة على شمول المحتوى بشكل جيد، وكذلك سهولة التطبيق، وموضوعية التصحيح، مما جعلها تتمتع بدرجة عالية من الصدق والثبات، وبالرغم من ذلك إلا أن هذا النوع من الاختبارات تعتمد دقة نتائجه على جودة بناء فقراته الاختبارية وكذلك التقيد بقواعد صياغة فقرات اختبار الاختيار من متعدد (Lin, 2018; Slepkov & Godfrey, 2019).

ولأهمية ذلك أُجريت العديد من الأبحاث التي تتعلق بصياغة فقرات الاختيار من متعدد مثل دراسة Haladyna, Downing & Rodriguez (2002) ، والتي توصلت إلى إحدى وثلاثين توصيةً تتعلق بصياغة فقرات الاختيار من متعدد نتيجة مراجعتهم لسبعة وعشرين مرجعاً متخصصاً ومن خلال سبع وعشرين دراسة تجريبية في القياس والتقويم، وكذلك ساهمت بعض الدراسات في الكشف عن العديد من انتهاكات صياغة فقرات الاختيار من متعدد مثل دراسة(2006) Tarrant, Knierim, Hayes & Ware, والتي شملت مراجعة (٢٧٧٠) فقرة من نوع الاختيار من متعدد في ضوء تسعة عشر قاعدة، وقد أظهرت النتائج أنَّ ما يقترب من نصف الأسئلة (٤٦.٢% ) تقريباً تنتهك القواعد، وأن أكثر من (٩٠%) من الأسئلة كُتبت لقياس مستوى معرفي متدن لدى الطلبة؛ ولذا لابد من دراسة أسئلة الاختيار من متعدد وضبط جوانب القصور، وتقوية نواحي القوة لاسيما عندما يكون القرار المبني على نتائجها حاسماً وحساساً.

وبمراجعة الدراسات والبحوث السابقة أتضح أن عدداً قليلاً منها تناول البحث عن أثر انتهاك عدد من القواعد الخاصة ببناء وصياغة فقرات الاختيار من متعدد على الخصائص السيكومترية على الاختبار وفقراته (حيدر إبراهيم ظاظا، ٢٠١٢؛ صبري حسن الطراونة، ٢٠١٥)، وأن هذه الدراسات التي أُجريت لم تقدم نتائج متسقة تصب في اتجاه واحد بل كان هناك تبايناً بين الدراسات المختلفة، فقد أكدت دراسات (ابتسام عيسى خصاونة، ٢٠١٢؛ نضال الشريفين، رانيا الصبح، ٢٠١١؛ Pachai, DiBattista & Kim, 2015) أن هناك فروقاً ذات دلالة إحصائية في معاملات صعوبة الفقرات محكمة البناء ونظائرها المخالفة لقواعد الصياغة لصالح الفقرات المخالفة لقواعد الصياغة، وكانت لصالح الفقرات المحكمة البناء في معاملات التمييز، بينما توصلت دراسات كل من (إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٠؛ Crehan & Haladyna, 1991) إلى عدم وجود فروق ذات دلالة إحصائية في تقدير معلمة الصعوبة تعزى إلى نموذج الاختبار (محكم، مخالف) لقواعد صياغة الفقرات؛ ولذلك لا يزال هذا الموضوع يحمل مزيداً من الدراسة والبحث، كما أن بعض القواعد المتعلقة بصياغة فقرات الاختيار من متعدد لم تحظ بالقدر الكافي من الاهتمام البحثي من قبل الباحثين وخاصة المتعلقة (بوضع البدائل بشكل أفقي وليس عمودي، وكذلك البدائل الغير معقولة ظاهرياً)، كما أن معظم الدراسات التي تناولت دراسة قواعد صياغة فقرات الاختيار من متعدد استخدمت المنظور الكلاسيكي في تحليل نتائج الاختبارات كدراسة (ابتسام عيسى خصاونة، ٢٠١٢؛ إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٢؛ Mueller & Schrock, 1982) والتي تضمنت دراسة المؤشرات الدالة على الإجابة الصحيحة، وصياغة المتن على شكل سؤال أو جملة غير مكتملة وأثر ارتكاب تلك المخالفات على الخصائص السيكومترية للاختبار وفقراته من المنظور الكلاسيكي، فضلاً عن قلة تلك الدراسات التي تناولت أثر انتهاك قواعد الصياغة على التقديرات المختلفة لنظرية الاستجابة للمفردة كدراسة (إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٠؛ محمد صيتان الصمادي، ٢٠١٥؛ نضال الشريفين، رانيا الصبح، ٢٠١١) والتي هدفت إلى التعرف على أثر تضمين الانتهاكات عند صياغة فقرات الاختيار من متعدد على التقديرات المختلفة لنظرية استجابة المفردة.

وقد أجريت عديد من الدراسات تناولت تقييم الطرق الكلاسيكية المختلفة لتقدير درجات اختبارات الاختيار من متعدد من حيث كفاءتها في تقدير قدرات الأفراد كالدراسات التي أجراها كل من (ساري سليم سواقد،١٩٩٢؛ صلاح شريف عبدالوهاب، ٢٠٠١ ؛ عفاف راضي اللحياني، ٢٠١٢؛ يوسف عبدالقادر أبوشندي، راشد سيف المحرزي، إيهاب محمد عمارة، ٢٠١٨؛Lau, Lau, Hong & Usop, 2011)، إلا جميع هذه الدراسات كانت محكات الحكم فيها تستند إلى نظرية القياس الكلاسيكي، كما أن النتائج التي خرجت بها جاءت متباينة حول أفضلية أي من هذه الطرق.

لهذا جاءت الدراسة الحالية لتكمل البناء البحثي حول هذا الموضوع وذلك من خلال المقارنة بين ثلاثة من هذه الطرق (الطريقة التقليدية، الطريقة التجريبية، وطريقة الاحتمال المقترح للإجابة الصحيحة) وفق معيار لم تتطرق له الدراسات السابقة وهو علاقة إحصائيات الفقرات والأفراد (صعوبة الفقرات، تمييز الفقرات، وقدرات الأفراد) المستمدة من هذه الطرق الثلاث بنظيراتها (معالم الفقرات والأفراد) المقدرة باستخدام النموذج اللوجستي الثلاثي البارامتر أحد نماذج نظرية الاستجابة للمفردة المهمة والذي يتميز بمراعاة أثر عامل التخمين على تقدير قدرات الأفراد، وهو ذات العامل الذي أهتمت به الطريقتين التجريبية، الاحتمال المقترح للإجابة الصحيحة اللتان تضمنتهما الدراسة الحالية، كما أن هذا النموذج يعد أكثر نماذج نظرية الاستجابة للمفردة فاعلية في تقدير قيم البارامترات (صلاح محمود علام، ٢٠٠٧، ٢٠٧).

ويمكن لنتائج هذه الدراسة أن تسهم في مساعدة الباحث الذي يرغب في استخدام نماذج نظرية استجابة المفردة ولكنه لم يتمكن من ذلك لصعوبة تحقيق متطلبات هذه النماذج في معرفة البديل المناسب من طرق تقدير الدرجات الكلاسيكية موضع الدراسة، والذي يكون الأقرب لتدريج قدرات الأفراد وفق النموذج اللوجستي الثلاثي البارامتر، وكذلك تزوده بتصور عن مقدار التوافق بين تدريج الدرجات المستمد من كل طريقة من الطرق الكلاسيكية الثلاث وتدريج القدرات المستمد من النموذج اللوجستي الثلاثي البارامتر.

وحيث أن طرق معادلة الاختبارات وفق نظرية استجابة المفردة تتغلب على الكثير من المشكلات التي عجزت طرق معادلة الاختبارات وفق النظرية الكلاسيكية عن حلها؛ وذلك لعدم قدرتها على تحقيق أغلب شروط المعادلة، فعلى سبيل المثال نجد أن طريقتي معادلة النسب المتساوية، والمعادلة الخطية وفق النظرية التقليدية تشترط وجود عينات متساوية القدرة، وفي حال عدم تحقق الشروط اللازمة تصبح معادلة الاختبار غير فعالة (Kolen& Brennan, 2014)، بينما نجد الأمر مختلف بالنسبة لطرق معادلة الاختبارات وفق نظرية استجابة المفردة فعندما يكون النموذج المستخدم مطابقاً لبيانات الاختبارات المراد معادلتها، تكون تقديرات معالم المفردات مستقلة عن قدرة الأفراد الذين طبق عليهم الاختبار، وتقديرات القدرة لعينة الأفراد تكون مستقلة عن المفردات المستخدمة في هذا التقدير، وبالتالي يمكن وضع تقديرات البارامترات للمفردة المشتقة من صور الاختبارات، وقدرات الأفراد على مقياس مشترك (Zhonghua, 2010).

كما تلعب دالة معلومات الاختبار دوراً مهماً في النظرية الحديثة في القياس، إذ يمكن من خلالها تحديد الخطأ المعياري في التقدير؛ حيث تتمتع دالة معلومات الاختبار، والتي تمثل مجموع دوال معلومات الفقرات عند مستوى معين من القدرة بميزة، وهي كون دالة معلومات الاختبار مستقلة عن عينة المفحوصين، وبذلك تقدم النظرية الحديثة في القياس مميزات إضافية، فيما يتعلق بزيادة القدرة على تقدير أخطاء القياس (حابس سعد الزبون، ٢٠١٣؛ نضال كمال الشريفين،٢٠١٢؛ Ayala, 2008).

ويرتبط تطوير نظريات القياس النفسي ونماذجها بكيفية معالجة أخطاء القياس، حيث يكون لتعيين الخطأ في النموذج تأثير كبير على كيفية تقدير درجات الخطأ، وفى داخل النظرية الكلاسيكية للاختبار من الممكن افتراض أن الخطأ يمكن تقسيمه بشكل طبيعي ويمكن افتراض أن حجم الأخطاء ثابت خلال مقياس درجة الاختبار (Campbell, 2015)، ومن ناحية أخرى ففي ظل نظرية الاستجابة للمفردة من الممكن افتراض أن حجم الأخطاء ربما يكون مرتبطاً بدرجة الممتحن الفعلية، ويتم حساب الخطأ المعياري للقياس بشكل منفصل بالنسبة لكل من قياس الفرد وتدريج المفردة (Hambleton, 2004; Ojerinde, 2013).

وفي الوقت الذي تتعدد فيه عيوب النظرية الكلاسيكية في القياس تزداد مزايا نظرية الاستجابة للمفردة بنماذجها المختلفة، وتتمثل هذه المزايا في استقلال خصائص المفردات عن عينة الأفراد المستخدمة في التحليل، وكذلك استقلال تقدير قدرات الأفراد عن عينة المفردات المكونة للمقياس، كما أن تفسير درجات الأفراد يتم في ضوء المفردات وليس في ضوء الجماعة المرجعية كما في النظرية الكلاسيكية، وتتحقق خصائص الميزان الفتري وربما القياس النسبي دون ضرورة أن يكون توزيع مستويات القدرة في المجتمع المستهدف اعتداليًا، ويتم تقدير الخطأ المعياري لكل مختبر على حده وليس خطأ معيارياً واحداً لكل المختبرين، مع إمكانية المقارنة بين أداء الأفراد الذين اختبروا باختبارات مختلفة تقيس نفس السمة (رحاب سعيد الحكماني ،٢٠٠٨ ؛ صلاح الدين محمود علام، ٢٠٠٥؛ عزالدين عبدالله النعيمي، ٢٠١٥).

وبناءً على ما سبق عرضه من طرح نظري وبعض نتائج الدراسات والبحوث السابقة ينصب اهتمام الدراسة الراهنة على التعرف على أثر اختلاف طريقة المعادلة وطرق تقدير الدرجات وقواعد صياغة الفقرات على دقة تقدير معالم الفقرات وقدرات الأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي ثلاثي البارامتر ومن ثم تتحد مشكلة الدراسة الحالية فى الإجابة عن التساؤلات التالية:

١- هل تتحقق افتراضات نظرية الاستجابة للمفردة الاختبارية على استجابات أفراد عينة الدراسة على نموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟

٢- ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات صعوبة الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر؟

٣- ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تمييز الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر؟

٤- ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تخمين الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير الخطأ المعياري لمتوسط معاملات تخمين الفقرات تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟

٥-ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على دقة تقديرات معالم القدرة للأفراد في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير متوسط الخطأ المعياري لتقدير قدرات الأفراد تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟

٦-هل هناك فروق ذات دلالة إحصائية بين التقديرات الخاصة بدالة معلومات الاختبار تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟

٧-ما تقديرات قدرات أفراد العينة في اختبار الاختيار من متعدد المستخدم في الدراسة وذلك وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) وكذلك وفق النموذج اللوجستي الثلاثي البارامتر؟

٨-ما درجة الارتباط/الاختلاف بين قدرات الطلاب عند تقديرها باستخدام النموذج اللوجستي الثلاثي البارامتر بتقديرات درجاتهم عند استخدام كل من الطرق الكلاسيكية لتقدير الدرجات التي شملتها الدراسة (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) ؟

٩-ما درجات ارتباط قيم معاملات صعوبة/تمييز الفقرات عند استخدام كل من الطرق الكلاسيكية لتقدير درجات الاختيار من متعدد (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) مع قيم معاملات صعوبة/تمييز الفقرات عند استخدام النموذج اللوجستي الثلاثي البارامتر؟

١٠- هل تختلف دقة معادلة درجات الاختبارات باختلاف طريقتي المعادلة (المتوسط/المتوسط، المتوسط/الانحراف المعياري) باستخدام النموذج اللوجستي الثلاثي البارامتر، لأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار(٢٥، ٥٠)؛ في ضوء محكي التحيز وجذر متوسط مربع الخطأ؟

## أهداف الدراسة :

**تسعى الدراسة الحالية إلى تحقيق الأهداف التالية :**

١-التحقق من افتراضات نظرية الاستجابة للمفردة الاختبارية على استجابات أفراد عينة الدراسة على نموذجي الاختبار (المحكم، المخالف) لقواعد صياغة الفقرات.

٢-التعرف على أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات صعوبة الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر.

٣-التعرف على أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تمييز الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر.

٤- التعرف على أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تخمين الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير الخطأ المعياري لمتوسط معاملات تخمين الفقرات تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات.

٥- التعرف على أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على دقة تقديرات معالم القدرة للأفراد في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير متوسط الخطأ المعياري لتقدير قدرات الأفراد تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟

٦- التعرف على دلالة الفروق الإحصائية بين التقديرات الخاصة بدالة معلومات الاختبار والتي تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر.

٧- التعرف على تقديرات قدرات أفراد العينة في اختبار الاختيار من متعدد المستخدم في الدراسة وذلك وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) وكذلك وفق النموذج اللوجستي الثلاثي البارامتر.

٨- التعرف على درجة الارتباط/الاختلاف بين قدرات الطلاب عند تقديرها باستخدام النموذج اللوجستي الثلاثي البارامتر بتقديرات درجاتهم عند استخدام كل من الطرق الكلاسيكية لتقدير الدرجات التي شملتها الدراسة (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة).

٩- التعرف على درجات ارتباط قيم معاملات صعوبة/تمييز الفقرات عند استخدام كل من الطرق الكلاسيكية لتقدير درجات الاختيار من متعدد (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) مع قيم معاملات صعوبة/تمييز الفقرات عند استخدام النموذج اللوجستي الثلاثي البارامتر؟

١٠- المقارنة بين طرق المعادلة لتحديد الطريقة الأدق في معادلة درجات الاختبارات (المتوسط/ المتوسط، المتوسط/الانحراف المعياري) باستخدام النموذج اللوجستي الثلاثي البارامتر، لأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار(٢٥، ٥٠)؛ في ضوء محكي التحيز وجذر متوسط مربع الخطأ؟

## أهمية الدراسة :

**تأتي أهمية الدراسة الحالية فى ضوء الجوانب التالية :**

١- إن الدراسة الحالية تتكامل مع الدراسات والبحوث التي تطرقت للمقارنة بين نماذج نظرية القياس الكلاسيكي ونماذج نظرية الاستجابة للمفردة، مما يمكن أن يساهم في الوصول لبناء معرفي متسق حول هذين الإطارين النظريين.

٢-إثراء الدراسات والبحوث المتعلقة بموضوع انتهاك قواعد صياغة فقرات اختبار الاختيار من متعدد؛ وذلك بدراسة المؤشرات الإحصائية المتعلقة بأثر بعض انتهاك صياغة الفقرات على الخصائص السيكومترية لفقرات الاختبار، فضلاً عن دراسة الجودة المتعلقة ببناء الاختبار والذي تؤلف الفقرة جزءاً منه وفق القياس الكلاسيكي وفي ضوء النموذج اللوجستي ثلاثي البارامتر وتقديم تغذية راجعة لمعديّ الاختبارات والمهتمين بشؤونها على مستوى الأفراد والمؤسسات.

٣-الوقوف الحقيقي على وضعين مختلفين لبناء فقرات اختبار الاختيار من متعدد (البناء المحكم، البناء المخالف) لقواعد الصياغة، وأثر ذلك على خصائص الأفراد والمفردات، مما يمكن من التحقق من صحة النتائج وإمكانية التعميم على مواقف مشابهه، ودعوة القائمين على العملية التعليمية بضرورة الاهتمام بجودة صياغة فقرات الاختيار من متعدد، وذلك باتباع إرشادات وقواعد الصياغة المستنبطة من الدراسات والبحوث المتخصصة في هذا الجانب، لأهمية القرارات التي تبنى عليها نتائج تلك الاختبارات.

٤-تعد الدراسة الحالية دراسة تقويمه لعدد من طرق تقدير الدرجات لاختبارات الاختيار من متعدد وذلك من خلال معيار لم تتطرق له الدراسات والبحوث السابقة التي حاولت تقويمها؛ وهو درجة ارتباط إحصائيات كل من الفقرات، والأفراد المستمدة من هذه الطرق بنظائرها المستمدة من النموذج اللوجستي ثلاثي البارامتر.

٥-توفير معلومات تفيد الباحثين وغيرهم في الكشف عن أكثر طرق تقدير الدرجات اتساقاً مع النموذج اللوجستي ثلاثي البارامتر، مما يمكنهم من استخدام البديل المناسب من بين هذه الطرق في حال عدم قدرتهم على استخدام النموذج.

٦-التعرف على عملية معادلة الاختبارات، والتي تعتبر من أهم تطبيقات نظرية الاستجابة للمفردة وأهميتها في وضع درجات الاختبارات على مقياس مشترك واحد، وكذلك إعطاء تصور واضح لمصممي الاختبارات لاختيار طريقة المعادلة التي تتناسب مع طبيعة الاختبار للوصول إلى أفضل دقة لمعادلة درجات صور الاختبار.

## مصطلحات الدراسة :

### ١- اختبار الاختيار من متعدد Multiple Choice Test :

يعرف اختبار الاختيار من متعدد بأنه السؤال الي يتكون من جزأين أولهما يعرف بالأساس وهو الخاص بالقضية التي يسأل عنها الطالب والثاني عدداً من الاختيارات تسمى البدائل، يختار من بينها الطالب الإجابة الصحيحة (Chang, Lin & Lin, 2007; Lau et al., 2011; Lin, 2018).

ويتحد اختبار الاختيار من متعدد في الدراسة الحالية إجرائياً: تبعاً لدرجة الطالب على نموذجي الاختبار (المحكم، المخالف) لقواعد صياغة فقرات الاختيار من متعدد لمقرر الاختبارات والمقاييس والذي تم تطبيقه على طلاب كلية التربية بجامعة أم القرى، خلال الفصل الدراسي الأول من العام الدراسي ٢٠١٧/٢٠١٨ م.

**٢- صياغة فقرات الاختيار من متعدد Crafting Items Multiple Choice :**

تتحد فقرات الاختيار من متعدد في الدراسة الحالية إجرائياً: بأنها فقرات الاختيار من متعدد لمقرر الاختبارات والمقاييس وفقاً لصورتين تقيسين نفس المحتوى:

**الصورة الأولى:** فقرات محكمة البناء تم صياغتها وفقاً لقواعد وإرشادات كتابة فقرات اختبار الاختيار من متعدد (Breakall et al., 2019; Rodriguez & Albano, 2017).

**الصورة الثانية:** فقرات فيها انتهاك لبعض قواعد صياغة فقرات الاختيار من متعدد وبالتحديد (اختلاف طول البدائل الصحيحة، البدائل غير المعقولة منطقياً، وضع البدائل بشكل أفقي وليس عمودي، وجود البديل "جميع ما ذكر" كإجابة صحيحة، وجود البديل "لا شيء مما ذكر" كإجابة صحيحة).

**٣- القياس الكلاسيكي Classic Measurement :**

يقصد به مجموعة الطرق الإحصائية الكلاسيكية التي تستخدم في حساب مفاهيم: الصعوبة ، التمييز، الصدق ، الثبات الخاصة باختبار مقرر "الاختبارات والمقاييس" موضع الدراسة.

أ - صعوبة الفقرة وفق القياس الكلاسيكي: هي النسبة المئوية من الطلاب الذين أجابوا عن الفقرة إجابة صحيحة، وأن أفضل درجة صعوبة للفقرة تلك التي تعطي أكبر تباين عندما تكون صعوبة الفقرة تقارب (٠.٥)، وأي فقرة ضمن توزيع لدرجات الصعوبة يتراوح بين (٠.٣ - ٠.٧) بمتوسط حسابي قدره (٠.٥) تعتبر ملائمة لجودة بناء الاختبار (Adedoyin, 2010; Cappelleri et al., 2014; Hambleton & Swaminathan,1985; Eleje, Onah & Abanobi, 2018) .

ب- تمييز الفقرة وفق القياس الكلاسيكي: هو مؤشر إحصائي يستخدم للكشف عن مدى قدرة الفقرة على التمييز بين الطلاب ذوي مستويات القدرة المختلفة، ويتم حسابه كلاسيكياً اعتماداً على طريقة المقارنة الطرفية، وفي ضوء المعيار الذي وضعه Ebel & Frisbie, (1991)، حيث أشارا إلى أن أية فقرة قيمة معامل تميزها سالب أو أقل من (٠.٢) تحذف، وأية فقرة ذات قدرة تمييزية أكبر من أو تساوي (٠.٢) وأقل من (٠.٤) تعتبر ذات تمييز مقبول وينصح بتحسينها، أما الفقرات ذات التمييز يساوى (٠.٤) فأكثر فتعتبر ذات تمييز جيد ويمكن الاحتفاظ بها (آن أناستازي، سوزانا أوربينا،٢٠١٥؛ صلاح الدين محمود علام ،٢٠١٥؛ ليندا كروكر، وجيمس الجينا، ٢٠١٧)..

**٤- نظرية الاستجابة للفقرة Item Response Theory(IRT)**

هي مجموعة الطرق الإحصائية التي تستخدم في حساب معالم الفقرات والأفراد "من صعوبة وتمييز وتخمين ودالة معلومات"، والمكونة من نماذج أحادية البعد هي : نموذج راش الأحادي البارامتر، نموذج لورد ثنائي البارامتر، نموذج بيرنبوم الثلاثي البارامتر (أحمد محمد التقي، ٢٠١٣؛ صلاح الدين محمود علام،٢٠٠٥؛ Ayala, 2008; Ayhan, 2015; DeMars, 2010; Natarajan, 2009; Coggins et al., 2017; Nering & Ostini, 2010; Reise & Revicki, 2015).

**٥- النموذج اللوجستي الثلاثي البارامتر Three-Parameter Logistic Model**

هو أحد نماذج نظرية الاستجابة للمفردة أحادية البعد ثنائية الاستجابة، ويراعي هذا النموذج إمكانية تفاوت فقرات الاختبار في صعوبتها، وقوتها التمييزية، واحتمالية الإجابة عليها عن طريق التخمين، حيث يفترض أن احتمال وصول الفرد للإجابة الصحيحة عن أي فقرة اختبارية $P(\theta)$، هو دالة في متغيرين هما: قدرة الطالب فيما يقيسه الاختبار $(\theta)$، وخصائص هذه الفقرة الاختبارية والتي تشمل في هذا النموذج كل من: بارامتر الصعوبة (b)، وبارامتر التمييز(a)، وبارامتر التخمين(c) (حسين عبدالنبي القيسي، ٢٠١٤؛ شاهر خالد سليمان، علي محمد الصالح،٢٠١٧؛ صلاح الدين محمود علام،٢٠٠٧ ؛ طه الخرشه، ٢٠١٦؛ فريال محمد أبو عواد، ٢٠١٨؛ DeMars, 2010; Kim & Lee, 2017; Tay, Huang, & Vermunt, 2016)، على النحو التالي.

أ- صعوبة الفقرة وفق النموذج اللوجستي ثلاثي البارامتر(b): تمثل نقطة على مقياس القدرة مقابلة لاحتمال الإجابة الصحيحة عن الفقرة بمقدار (1+C)/2 حيث (C) عبارة عن احتمالية إجابة الأفراد ذوي المستويات المنخفضة من القدرة إجابة صحيحة عن الفقرة عن طريق التخمين، ويقدر بارامتر الصعوبة بوحدة اللوجيت.

ب- تمييز الفقرة وفق النموذج اللوجستي ثلاثي البارامتر(a): يمثل ميل المنحنى المميز للفقرة عند نقطة انقلاب المنحنى، وهي النقطة التي يكون فيها احتمال إجابة الفرد عن الفقرة إجابة صحيحة يساوي (1+C)/2 .

ج- بارامتر تخمين الفقرة وفق النموذج اللوجستي ثلاثي البارامتر(c): هو مؤشر إحصائي يعكس احتمال إجابة الأفراد من ذوي المستويات المنخفضة فيما يقاس عن المفردة إجابة صحيحة، وهو بذلك الجزء المقطوع من محور الصادات، ويسمى مستوى شبه الصدفة Pseudo-Chance، أو الخط التقاربي الأدنى للمنحنى Lower Asymptote. (Georgiev, 2008; Nering & Ostini, 2010; Raykov & Marcoulides, 2016; Reise & Revicki, 2015)

**٦- دالة المعلومات للاختبار Test Information Function(TIF):**

تمثل هذه الدالة بعلاقة منحنية بين متغيرين هما مستويات القدرة التي يمثلها المحور الأفقي والمعلومات التي يوفرها الاختبار ككل، وتعبر هذه الدالة عن كمية المعلومات المقدم من المجموع الكلي لمفردات الاختبار عند مستوى معين من القدرة (Hambleton & Jones, 1994; Jinming, 2012; Joo, Lee & Stark, 2018; Reise & Revicki, 2015)

**٧- دالة المعلومات للفقرة Item Information Function(IIF):**

تمثل هذه الدالة بعلاقة منحنية بين متغيرين هما مستويات القدرة التي يمثلها المحور الأفقي والمعلومات المقدمة من خلال الفقرة التي يمثلها المحور الرأسي، وتعبر هذه الدالة عن كمية المعلومات التي تقدمها الفقرة عن مستوى القدرة التي تقيسها (David, 2013; Joo et al., 2018; Nering & Ostini, 2010; Van der Linden, 2016)

**٨- طرق تقدير الدرجات Methods of Scoring:**

تمثل القاعدة التي يُعطى في ضوئها قيماً كمية تعكس الدرجة المستحقة للطالب في كل فقرة من فقرات الاختبار، وقد اشتملت الدراسة الحالية على ثلاث طرق لتقدير درجات فقرات الاختيار من متعدد وهي: الطريقة التقليدية، والطريقة التجريبية، وطريقة الاحتمال المقترح للإجابة الصحيحة (Dimiter, 2016; Lau et al., 2011; Lesage et al., 2013; Ndalichako, & Rogers, 1997; Sočan, 2015; Vanderoost et al., 2018; Zhonghua, 2010) ، على النحو التالي:

أ- الطريقة التقليدية Conventional Scoring Method: حيث يطلب من الطالب في هذه الطريقة أن يختار أحد بدائل فقرة الاختيار من متعدد لتعبر عن إجابة هذه الفقرة، فإذا كان البديل الذي تم اختياره صحيحاً فإن درجة الطالب على هذه الفقرة تكون (١) أما إذا كان البديل الذي تم اختاره خاطئاً فإن درجته ستكون (صفراً).

ب- الطريقة التجريبية Experimental Method: تختلف إجابة الطالب ودرجته على الفقرة في هذه الطريقة، بحسب مدى ثقته بمعرفة البديل الصحيح، فإذا كان الطالب متأكداً من معرفة البديل الصحيح، فإنه يضع أمامه (١) وكان هذا البديل هو البديل الصحيح يحصل الطالب على ثلاث درجات؛ وإذا كان الطالب يشك في صحة بديلين فإنه يضع أمام أحدهما (١) وأمام الثاني (٢) وكان أحد البديلين هو البديل الصحيح يحصل الطالب على درجتين؛ وإذا كان الطالب يشك في صحة ثلاثة بدائل فإنه يضع أمام أحدهما (١) وأمام الثاني (٢) وأمام الثالث (٣) وكان أحدهم هو البديل الصحيح يحصل الطالب على درجة واحدة، ويحصل الطالب على درجة (صفر) إذا لم يقع البديل الصحيح ضمن البدائل التي اختارها، أو إذا قام الطالب باختيار جميع بدائل الفقرة.

ج- طريقة الاحتمال المقترح للإجابة الصحيحة Method of Probability Assigned to the Correct Answer: يقوم الطالب في هذه الطريقة بإعطاء نسب مئوية تعبر عن مدى تقديره لصحة كل بديل من بدائل فقرة الاختيار من متعدد، بحيث يكون مجموع هذه النسب مساويا لـ ١٠٠٪، ويتم تقدير درجة المفحوص بأخذ النسبة المئوية التي اقترحها للبديل الصحيح لتعبر عن درجته على الفقرة.

**٩- طرق المعادلة Functioning Method:**

يقصد بها إجراءات رياضية تم تطويرها لتحقيق إجراءات المعادلة وتم تحديد طريقتين من طرق المعادلة (Hambelton & Swaminthan, 1985; Inal & Anil, 2018; Kolen & Brennan, 2014; Zhonghua, 2010) وهما:

أ- طريقة المتوسط/المتوسط Mean & Mean Method : تعتمد هذه الطريقة على حساب متوسط معلمات التمييز $(a)$ والصعوبة $(b)$ لجميع المفردات المشتركة بين صورتي الاختبار (X&Y)، لتقدير الثابتين $(\alpha,\beta)$، ويتم الحصول على التقديرات للمعالم المطلوبة من خلال استبدال تقديرات الثابتين $(\alpha,\beta)$ في المعادلات الثلاثة التالية $\theta yi = \alpha\,\theta xi + \beta$ ، $byj = \alpha\,bxj + \beta$ ، $\alpha yj = \frac{\alpha\,xj}{\alpha}$ .

ب- طريقة المتوسط/الانحراف المعياري Mean & Sigma Method : تعتمد هذه الطريقة على حساب المتوسط والانحراف المعياري لمعلمات الصعوبة $(b)$ لجميع المفردات المشتركة بين صورتي الاختبار (X&Y)، لتقدير الثوابت $(\alpha,\beta)$ ، وبعد تحديد ثوابت المعادلة $(\alpha,\beta)$ يمكن تحويل تقدير معالم المفردات على الاختبار X إلى مقياس واحد هو مقياس الاختبار Y باستخدام معادلات التحويل الأساسي وهي: $a^*y = \frac{\ddot{a}x}{a}$ , $b^*y = a\,bx + \beta$ .

**١٠- دقة المعادلة: Accuracy of Equation:**

هو أسلوب إحصائي يستخدم للتأكد من مدى دقة وصحة عملية المعادلة باستخدام اختبار الفقرات المشتركة (Harris& Crouse, 1993)، فإذا تمت دقة المعادلة بنجاح فإنه من الممكن مناقشة التغير الحقيقي عبر صور متكافئة من الاختبارات ومقارنة المفحوصين الخاضعين لتطبيق هذه الاختبارات، ويوجد العديد من المحكات للاستدلال ولتقييم دقة المعادلة لصورتي الاختبار، منها:

أ- التحيز Bias : يمكن استخدام التحيز كمحك لتقييم دقة المعادلة عندما يتم تطبيق صور الاختبار على نفس مجموعة المفحوصين، ويمكن حساب التحيز لدالة المعادلة؛ بطرح نتائج دالة المعادلة الحقيقية من دالة المعادلة المقدرة وكلما قلت قيمته دل ذلك على دقة المعادلة (Dimiter, 2016; Kellere, 2007; Kolen & Brennan, 2014; Öztürk-Gübes & Kelecioglu, 2016) .

ب- الجذر التربيعي لمتوسط مربعات الأخطاء (Root Mean Square Error (RMSE:
تبرز أهمية الجذر التربيعي لمتوسط مربعات الأخطاء لأنه يعكس مقدار التحيز، وكذلك يعكس دقة نتائج المعادلة مقارنة بمعيار المعادلة المستخدم، وكلما قلت قيمة هذا الاحصائي واقترب من الصفر دل ذلك على زيادة دقة المعادلة والعكس صحيح (Albano et al., 2018; Inal & Anil, 2018; Petersen, Kolen & Hoover, 1989; Zhonghua, 2010).

## إجراءات الدراسة:

### أولاً : عينة الدراسة:

**١ - عينة تقنين الأدوات:**

تم تقنين الأداة المستخدمة فى الدراسة على عينة من طلاب كلية التربية (شعبة التربية الخاصة، شعبة التربية الفنية، شعبة التربية الرياضية، شعبة التربية الأسرية) جامعة أم القرى بمكة المكرمة، قوامها ٢٨٤ طالباً وطالبة، تتراوح أعمارهم بين (٢٠.٦ – ٢١,٢) سنة، بمتوسط عمري قدره ٢٠.٩ سنة، وانحراف معياري قدره ٠.٢٥ سنة، وقد روعي أن تتوافر فيها معظم خصائص ومواصفات العينة الأساسية للدارسة الحالية.

**٢ - عينة الدراسة الأساسية:**

تكونت عينة الدراسة الأساسية من طلاب كلية التربية (شعبة التربية الخاصة، شعبة التربية الفنية، شعبة التربية الرياضية، شعبة التربية الأسرية، شعبة رياض الأطفال) جامعة أم القرى بمكة المكرمة، بلغ عددهم ١٤٠٠ طالباً وطالبة، منهم (٦٧٨) طالباً و(٧٢٢) طالبة، حيث تراوحت أعمارهم (٢٠.٤ – ٢١,٧) سنة، بمتوسط عمري قدره ٢١.٠٥ سنة، وانحراف معياري قدره ٠.٤٦ سنة، ، تم اختيارهم بالطريقة العشوائية الطبقية خلال العام الجامعي ٢٠١٦ / ٢٠١٧ م.

### ثانياً: أداتا الدراسة:

- **الاختبار التحصيلي في مقرر الاختبارات والمقاييس: ( إعداد : الباحث )**

لتحقيق هدف الدراسة والمتمثل في أثر اختلاف طريقة المعادلة وطرق تقدير الدرجات وقواعد صياغة الفقرات على دقة تقدير معالم الفقرات وقدرات الأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي ثلاثي البارامتر، قام الباحث بتصميم أداة الدراسة، وعملية تصميم الأداة في المقام الأول تعتمد على القيام بعدة خطوات متسلسلة تؤدي في النهاية إلى تجنب كثير من الأخطاء وتتيح إمكانية إعداد أدة جيدة يُعتمد عليها في المجال المعني.

وللكشف عن أثر انتهاك بعض قواعد صياغة فقرات الاختيار من متعدد على دقة تقدير معالم الفقرات والأفراد في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر وللإجابة عن تساؤلات الدراسة وتحقيق أهدافها؛ صمم الباحث أداتا الدراسة (نموذجي اختبار) من نوع الاختيار من متعدد من أربعة بدائل في مقرر الاختبارات والمقاييس، وذلك في ضوء دراسة الأدبيات التربوية، ومن خلال الاطلاع على الأطر النظرية والدراسات والبحوث السابقة، وقد اقتضى بناء نموذج الاختبار المحكم ووضعه في صيغته المعدة للتحكيم الاسترشاد بالأسس العامة المتبعة عند بناء فقرات اختبار الاختيار من متعدد التي أوردها (أحمد عودة، ٢٠١٤؛ آن أناستازي، سوزانا أوربينا، ٢٠١٥؛ صلاح الدين محمود علام، ٢٠١٥؛ Breakall et al., 2019; Campbell, 2015; Haladyna et al., 2002; Rodriguez & Albano, 2017; Thomas et al., 2002).

## [ ١ ]: بناء الاختبار التحصيلي لمقرر الاختبارات والمقاييس (محكم البناء):

تم بناء أداة الدراسة المتمثلة في الاختبار التحصيلي لمقرر الاختبارات والمقاييس لدى طلاب وطالبات كلية التربية، جامعة أم القرى، في ضوء المنهجية العلمية في بناء الاختبارات، وفي ضوء معايير بناء اختبارات الاختيار من متعدد، وفقاً للخطوات التالية:

١- تحديد الغرض من الاختبار: تمثل في بناء اختبار تحصيلي لقياس التحصيل الدراسي لدى طلاب وطالبات كلية التربية بجامعة أم القرى في مقرر ( الاختبارات والمقاييس ).

٢- تحديد النطاق السلوكي: تمثل في الموضوعات المدرجة في توصيف مقرر الاختبارات والمقاييس التابع لقسم علم النفس في كلية التربية بجامعة أم القرى، والتي تمثل النطاق السلوكي المختار.

٣- تحليل المحتوى الدراسي: تم تحليل محتوى مقرر الاختبارات والمقاييس وفق التوصيف إلى ستة موضوعات رئيسة على النحو التالي:

أ- القياس النفسي: تعريف القياس النفسي، خصائصه، أنواعه، أدواته، مستوياته، أغراضه في العملية التربوية، الأسس العلمية للقياس النفسي والتربوي.

ب- التقويم التربوي: تعريف التقويم التربوي، خصائصه، أنواعه، أسسه ووظائفه، أدواره ومجالاته في العملية التعليمية، أساليب وأدوات التقويم التربوي.

ج-الخصائص السيكومترية: الصدق (مفهومه، طرقه، العوامل المؤثرة فيه)، الثبات (مفهومه، طرقه، العوامل المؤثرة فيه) المعايير (الدرجة المعيارية، الدرجة التائية).

د- المقاييس النفسية: تعريف المقاييس النفسية، تصنيفها، أغراضها، أسس بناء المقاييس النفسية، كيفية تطويرها، ومجالات استخداماتها في العملية التعليمية.

هـ- الاختبارات العقلية: تعريف الاختبارات العقلية، تصنيفها، أغراضها، أسس بناء الاختبارات العقلية، كيفية تطوريها، ومجالات استخداماتها في العملية التعليمية.

و- الاختبارات التحصيلية: تحليل المحتوى الدراسي، صياغة الأهداف السلوكية وفقاً لتصنيف بلوم، بناء جدول مواصفات الاختبار التحصيلي، بناء الاختبار التحصيلي (بناء الفقرات الاختبارية، إخراج وتطبيق وتصحيح الاختبار التحصيلي، تحليل نتائج الاختبار التحصيلي على مستوى الفقرات والدرجة الكلية وتفسيرها، كتابة التقارير النهائي).

٤- اشتقاق مخرجات التعلم: تم اشتقاق مخرجات التعلم بحيث أن يكون الطالب قادراً على أن يفرق بين مصطلحات القياس والتقييم والتقويم، ويوضح طبيعة العلاقة بينها، يُعرف مصطلح القياس تعريفاً علمياً دقيقاً، ومنه يستنتج عناصر عملية القياس، يذكر أغراض القياس في العملية التربوية، يتعرف على الأسس العلمية للقياس النفسي والتربوي، يميز بين أنواع القياس، يفرق بين مستويات القياس في ضوء ما تعكسه الدرجات من خصائص الدرجة الحقيقية في كل مستوى، يحدد المفهوم العلمي لمصطلح التقويم التربوي، يستنتج المكونات الأربعة الأساسية للتعريف الشامل للتقويم التربوي، يستنتج أدوار ومجالات ووظائف وأدوات التقويم التربوي من خلال منظومة التدريس، يخطط لبناء اختبار تحصيلي وفقاً للأسس العلمية لبناء الاختبارات التحصيلية، يحلل محتوى المقرر الدراسي إلى مكوناته الأساسية من المعارف والخبرات، يفرق بين مجالات الأهداف التربوية، يصيغ الأهداف السلوكية مسترشداً بتصنيف بلوم لمستويات المجال المعرفي، مراعياً مكونات الهدف السلوكي الجيد، يختار نوع الفقرات الاختبارية المناسبة لقياس مدى تحقق الهدف السلوكي، يصيغ الفقرات الاختبارية بأنواعها مراعياً معايير الصياغة الجيدة لكل نوع، يصمم جدول مواصفات الاختبار التحصيلي يدوياً وباستخدام برنامج جدول المواصفات الحاسوبي، ويوظفه في اختيار عينة فقرات الاختبار بحيث تكون ممثلة للاختبار، يخرج الورقة الاختبارية بصورة ملائمة وفقاً للمعايير المحددة لذلك.

٥- صياغة الأهداف السلوكية: تم صياغة الأهداف السلوكية التي تشمل جميع موضوعات مقرر الاختبارات والمقاييس في ضوء أدبيات وشروط صياغة الأهداف السلوكية وتصنيفها؛ حيث تم صياغة هدف سلوكي أو أكثر لكل موضوع من الموضوعات، مع تحديد مستوى كل هدف سلوكي بصورة تسمح لقياسه بفقرات الاختيار من متعدد؛ حيث تم كتابته على المستويات المعرفية لتصنيف بلوم (التذكر، الفهم، التطبيق، التحليل، التركيب، التقويم) بعد تحليل محتوى المقرر، وقد بلغ عدد الأهداف الاجمالية للمحتوى (١٤٠) هدفاً، تم عرضها على عشرة محكمين من المتخصصين في القياس والتقويم (ملحق، ٣)، من أجل التحقق من مدى وضوحها، وسلامة صياغتها اللغوية، وعدم تكرارها، ومدى ملائمتها لهدف الدراسة، ومناسبتها للمحتوى المعرفي المستهدف، وبناءً على ذلك تم إجراء التعديلات المتعلقة بالصياغة في (٧) أهداف، وتم حذف (٥) أهداف، وقد بلغ عدد الأهداف في صورتها النهائية (١٣٥) هدفاً موزعة على النحو التالي: (٣٥) هدف تذكر، (٣٠) هدف فهم، (٢٥) هدف تطبيق، (٢٠) هدف تحليل، (١٥) هدف تركيب، (١٠) هدف تقويم.

٦- بناء جدول المواصفات: قام الباحث ببناء جدول مواصفات للاختبار التحصيلي لمقرر الاختبارات والمقاييس في ضوء الوزن النسبي للزمن المستغرق في تدريس كل موضوع، وكذلك في ضوء الوزن النسبي للأهداف في كل مستوى؛ حيث تم ربط الأهداف السلوكية بالمحتوي الدراسي، وذلك من خلال تقدير أوزان تتناسب مع كل من المحتوى الدراسي، والأهداف السلوكية؛ ولأن الاختبار التحصيلي للمقرر تم تحديد حجمه بستين فقرة وهو ما يتناسب مع الوقت المخصص للاختبار، لذا تم بناء جدول المواصفات والذي أُستخدم لانتقاء فقرات الاختبار، وجدول (١) التالي يوضح جدول مواصفات الاختبار التحصيلي لمقرر الاختبارات والمقاييس.

**جدول (١)**

**جدول مواصفات الاختبار التحصيلي لمقرر الاختبارات والمقاييس**

| الموضوعات | عدد المحاضرات | مستويات الأهداف السلوكية | | | | | | عدد الفقرات | الأوزان النسبية للموضوعات |
|---|---|---|---|---|---|---|---|---|---|
| | | التذكر (٣٥) | الفهم (٣٠) | التطبيق (٢٥) | التحليل (٢٠) | التركيب (١٥) | التقويم (١٠) | | |
| القياس النفسي | ٢ | ٢ | ٢ | ١ | ١ | ١ | ١ | ٨ | ١٢.٥٪ |
| التقويم التربوي | ٢ | ٢ | ٢ | ٢ | ١ | ١ | – | ٨ | ١٢.٥٪ |
| الخصائص السيكومترية | ٣ | ٣ | ٣ | ٢ | ٢ | ١ | – | ١١ | ١٨.٨٪ |
| المقاييس النفسية | ٣ | ٣ | ٢ | ٢ | ٢ | ١ | ١ | ١١ | ١٨.٨٪ |
| الاختبارات العقلية | ٣ | ٣ | ٢ | ٢ | ١ | ٢ | ١ | ١١ | ١٨.٨٪ |
| الاختبارات التحصيلية | ٣ | ٣ | ٢ | ٢ | ٢ | ١ | ١ | ١١ | ١٨.٨٪ |
| المجموع | ١٦ | ١٦ | ١٣ | ١١ | ٩ | ٧ | ٤ | ٦٠ | ١٠٠٪ |
| الأوزان النسبية للأهداف | | ٢٥.٩٪ | ٢٢.٢٪ | ١٨.٥٪ | ١٤.٨٪ | ١١.١٪ | ٧.٤٪ | ١٠٠٪ | |

٧- كتابة الفقرات الاختبارية: تم بناء فقرات الاختبار التحصيلي بكتابة الفقرات الاختبارية والتي تتناغم مع الأهداف السلوكية للمقرر والمخرجات السابقة وتحققها في ضوء قواعد وتعليمات كتابة الفقرات الاختبارية بوجه عام وفقرات الاختيار من متعدد بوجه خاص؛ حيث تم دراسة كل هدف سلوكي على حدة، وصياغة أفضل الفقرات التي تقيس ذلك الهدف، بأنماط استجابة مختلفة من نوع الاختيار من متعدد رباعي البدائل، وذلك لكل هدف سلوكي تمت صياغته في الخطوة السابقة؛ بحيث يتناغم مع الهدف السلوكي والمستوى العقلي المعرفي له، وقد حرص الباحث في بناء الفقرات على مراعاة الخصائص والشروط الواجب توافرها في نمط الاختيار من متعدد، وذلك لقياس كل هدف من الاهداف السلوكية التي تم صياغتها، وقد بلغ عدد الفقرات (٦٠) فقرة، موزعة حسب الأهمية النسبية للموضوعات الواردة في جدول المواصفات السابق ذكره.

٨- التحقق من صدق المحتوى: تم التحقق من صدق محتوى الاختبار التحصيلي لمقرر الاختبارات والمقاييس، والمكون من (٦٠) فقرة من خلال عرضه مرفقاً معه الأهداف السلوكية وجدول مواصفات الاختبار على عشرة محكمين من المتخصصين في القياس والتقويم ( ملحق، ٣) وذلك بغرض الحكم على مدى ارتباط الأهداف السلوكية بالأهداف العامة للمقرر، ومدى تمثيلها للمجال السلوكي الذي يقيسه الاختبار، إضافة إلى مدى قياس الفقرات للأهداف السلوكية المرتبطة بها، والمستوى العقلي المعرفي لها، وأخيراً الحكم على مدى جودة صياغة فقرات الاختبار وتحقيقها لمواصفات الفقرات الجيدة، مع إمكانية حذف أو دمج أو إضافة فقرات بما يرونه مناسباً، وبناء على ملاحظات المحكمين تم حذف (٤) فقرات اختيارية، وبذلك تكون الاختبار في صورته النهائية من (٥٦) فقرة أجمع المحكمين على أنها عينة كافية وممثلة للمجال السلوكي الذي يقيسه الاختبار.

٩- الدراسة الاستطلاعية (إجراءات التطبيق الأولى): تم تطبيق الاختبار بصورته الأولية على العينة الاستطلاعية المكونة من (٢٨٤) طالباً وطالبة، للتجريب الأولي للاختبار بغرض التعرف على الزمن المناسب للتطبيق، والتأكدمن وضوح التعليمات والصياغة اللغوية للفقرات، وملائمة بدائل كل فقرة، والتعرف على الصعوبات التي يمكن أن تواجه الطلاب أثناء الإجابة على فقرات الاختبار، ولم يكشف التجريب الأولي للاختبار عن أية ملاحظات ذات أهمية فيما يتعلق بوضوح الفقرات، كما لم تظهر أي صعوبات عند تطبيق الاختبار، وقد انتهى جميع الطلاب من الإجابة عن فقرات الاختبار خلال ساعة ونصف مما يدل على أنه الوقت المناسب للإجابة عن فقرات الاختبار في صورته النهائية، كما تم استخدام عينة الدراسة الاستطلاعية في التحقق من الخصائص السيكومترية للفقرات، ومن ثم تحليل مفردات الاختبار واستخراج معاملات الصعوبة والتمييز من منظور النظرية الكلاسيكية في القياس.

١٠- الدراسة الأساسية (الصورة النهائية للاختبار): بعد إعداد فقرات الاختبار في صورته النهائية تم توفير نسخ الاختبار اللازمة للتطبيق على أفراد العينة، كما تم توفير عدد كافي من نماذج ورق الإجابة الإلكترونية التي تصحح بآلات التصحيح، وأقلام الرصاص، وطبق الاختبار (بنماذجه الأربعة) على عينة الدراسة التي تكونت من (١٥٠٠) طالباً وطالبة من الطلاب عينة الدراسة الأساسية، وذلك في نهاية الفصل الدراسي الأول من العام الجامعي ٢٠١٨/٢٠١٧م، وبعد الانتهاء من التطبيق تم تصحيح أوراق الإجابة باستخدام آلة التصحيح والحصول على ملف الكتروني للبيانات، والذي تم معالجته ببرامج التحليل الإحصائي المناسبة لنوعي القياس الكلاسيكي والموضوعي للحصول على النتائج اللازمة للإجابة عن تساؤلات الدراسة.

- **الخصائص السيكومترية للاختبار التحصيلي لمقرر الاختبارات والمقاييس (محكم البناء):**

♦ **صدق الاختبار:**

لفحص مدى تمتع الاختبار التحصيلي لمقرر الاختبارات والمقاييس بدلالات صدق كافية قام الباحث بإجراء طرق الصدق التالية:

١- صدق المحتوى (صدق المحكمين): تم التحقق من صدق المحتوى للاختبار من خلال عرض فقرات الاختبار، والأهداف السلوكية التي تقيسها، وجدول مواصفات الاختبار على عشرة محكمين من المتخصصين في مجال القياس والتقويم ( ملحق، ٣)، وقد أجمع المحكمين على كفاية وتمثيل فقرات الاختبار للمجال السلوكي الذي يقيسه الاختبار، كما أجمع المحكمين على جودة (٥٠) فقرة من فقرات الاختبار؛ حيث كانت نسبة اتفاقهم (١٠٠٪)، و(٦) فقرات كان بينهم اختلاف يسير في جودتها حيث كانت نسبة اتفاقهم (٩٠٪)، كما بلغت قيمة معامل كابا Kappa لاتفاق المحكمين (٠.٩٧)، وهي قيمة دالة عند مستوى دلالة (٠.٠١)، وتدل على اتساق المحكمين في حكمهم على جودة فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس وكفايتها.

٢-الاتساق الداخلي (صدق التكوين الفرضي): تم إيجاد الاتساق الداخلي للاختبار كمؤشر من مؤشرات صدق التكوين الفرضي وذلك من خلال حساب معامل الارتباط ثنائي التسلسل الحقيقي المصحح Point Biserial Correlation Coefficient (PTBIS) بين درجات الطلاب على كل فقرة من فقرات الاختبار والدرجة الكلية للاختبار بعد حذف درجة الفقرة من الدرجة الكلية للاختبار، كما هو موضح في جدول (٢) التالي.

**جدول (٢)**

**قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح بين كل فقرة والدرجة الكلية للاختبار المحكم البناء**

| الفقرة | PTBIS | الفقرة | PTBIS | الفقرة | PTBIS | الفقرة | PTBIS |
|---|---|---|---|---|---|---|---|
| ١ | ٠.٨٤ | ١٥ | ٠.٨٩ | ٢٩ | ٠.٨٤ | ٤٣ | ٠.٩٣ |
| ٢ | ٠.٩٢ | ١٦ | ٠.٨٥ | ٣٠ | ٠.٨٨ | ٤٤ | ٠.٩٠ |
| ٣ | ٠.٩٢ | ١٧ | ٠.٨٤ | ٣١ | ٠.٩٤ | ٤٥ | ٠.٩١ |
| ٤ | ٠.٩٥ | ١٨ | ٠.٩١ | ٣٢ | ٠.٩٦ | ٤٦ | ٠.٨٦ |
| ٥ | ٠.٨٥ | ١٩ | ٠.٩٤ | ٣٣ | ٠.٩٣ | ٤٧ | ٠.٩٤ |
| ٦ | ٠.٨٣ | ٢٠ | ٠.٨٨ | ٣٤ | ٠.٨٥ | ٤٨ | ٠.٨٨ |
| ٧ | ٠.٩٢ | ٢١ | ٠.٩٥ | ٣٥ | ٠.٨٣ | ٤٩ | ٠.٨٤ |
| ٨ | ٠.٩٠ | ٢٢ | ٠.٩٠ | ٣٦ | ٠.٩٢ | ٥٠ | ٠.٩٠ |
| ٩ | ٠.٩٠ | ٢٣ | ٠.٩٢ | ٣٧ | ٠.٩٦ | ٥١ | ٠.٨٧ |
| ١٠ | ٠.٨٩ | ٢٤ | ٠.٨٦ | ٣٨ | ٠.٨٥ | ٥٢ | ٠.٩٥ |
| ١١ | ٠.٨٦ | ٢٥ | ٠.٩٠ | ٣٩ | ٠.٩٤ | ٥٣ | ٠.٩٣ |
| ١٢ | ٠.٩٥ | ٢٦ | ٠.٨٧ | ٤٠ | ٠.٩٠ | ٥٤ | ٠.٨٩ |
| ١٣ | ٠.٨٩ | ٢٧ | ٠.٩٣ | ٤١ | ٠.٩٧ | ٥٥ | ٠.٨٥ |
| ١٤ | ٠.٩٤ | ٢٨ | ٠.٩٥ | ٤٢ | ٠.٩١ | ٥٦ | ٠.٨٨ |

PTBIS: معامل الارتباط ثنائي التسلسل الحقيقي المصحح

يتضح من جدول (٢) السابق أن قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح قد تراوحت بين (٠.٨٣-٠.٩٧) بمتوسط حسابي (٠.٨٩٩) وانحراف معياري (٠.٠٣٩) وهي قيم مرتفعة دالة عند مستوى دلالة (٠.٠١)، تدل على قوة ارتباط درجة الفقرات بالدرجة الكلية للاختبار، وبالتالي الاتساق الداخلي للاختبار.

## ♦ معاملات الصعوبة والتمييز:

### ١- معاملات الصعوبة:

تم حساب معامل الصعوبة Item-Difficulty Index كلاسيكياً بحساب النسبة المئوية من الطلاب الذين أجابوا عن الفقرة إجابة صحيحة، ويشير (Cappelleri et al., 2014; Hambleton & Swaminathan, 1985; Eleje et al., 2018) إلى أن أفضل درجة صعوبة للفقرة تلك التي تعطي أكبر تباين عندما تكون صعوبة الفقرة تقارب (٠.٥)، وأي فقرة ضمن توزيع لدرجات الصعوبة يتراوح بين (٠.٣ - ٠.٧) بمتوسط حسابي قدره (٠.٥) تعتبر ملائمة لجودة بناء الاختبار، ولقد تراوحت قيم معاملات صعوبة الفقرات، وفقًا لإجابات الطلاب على نموذج الاختبار المحكم البناء بين (٠.٢١-٠.٦٦) بمتوسط حسابي (٠.٤٤٧) وانحراف معياري (٠.٠٩٧)، وقد حصلت الفقرة (١٤) على أعلى معامل صعوبة، بينما حصلت الفقرة (٣٥) على أدنى معامل صعوبة، وقد تم حذف ثلاث فقرات هي (٢٦، ٣٥، ٥٢)، وذلك نظراً لتدنى معاملات الصعوبة الخاصة بكل فقرة منها في ضوء المعايير السابقة لمعاملات الصعوبة، حيث كانت معاملات الصعوبة لهذه الفقرات (٠.٢٢، ٠.٢١، ٠.٢٣) على الترتيب.

### ٢- معاملات التمييز:

تم حساب معامل التمييز Item- Discrimination Index كلاسيكياً (اعتماداً على طريقة المقارنة الطرفية)، وفي ضوء المعيار الذي وضعه Ebel & Frisbie, (1991) في اختيار معامل التمييز المقبول لمفردات الاختبار، حيث أشار إلى أن أية فقرة قيمة معامل تميزها سالب أو أقل من (٠.٢) تحذف ولا داعي للاحتفاظ بها، وأية فقرة ذات قدرة تمييزية أكبر من أو تساوي (٠.٢) وأقل من (٠.٤) تعتبر ذات تمييز مقبول وينصح بتحسينها، أما الفقرات ذات التمييز يساوى (٠.٤) فأكثر فتعتبر ذات تمييز جيد ويمكن الاحتفاظ بها، ولقد تراوحت قيم معاملات تمييز الفقرات وفقًا لإجابات الطلاب على نموذج الاختبار المحكم البناء تراوحت ما بين (٠.١٣-٠.٧١) وبمتوسط حسابي (٠.٤٨٢) وانحراف معياري (٠.١٢٣) ، وقد حصلت الفقرة (٤٩) على أعلى معامل تمييز، بينما حصلت الفقرة (٢٠) على أدنى معامل تمييز، وقد تم حذف ثلاث فقرات هي (٢٠، ٣١، ٤٢)، وذلك نظراً لتدنى معاملات التمييز بكل فقرة منها في ضوء المعايير السابقة لمعاملات التمييز، حيث كانت معاملات التمييز لهذه الفقرات (٠.١٣، ٠.١٨، ٠.١٥) على الترتيب، وعليه أصبح الاختبار المحكم البناء في صورته النهائية مكون من (٥٠) فقرة يمكن الوثوق بها من أجل التطبيق النهائي على عينة الدراسة.

### ♦ ثبات الاختبار:

تم حساب ثبات الاختبار التحصيلي لمقرر الاختبارات والمقاييس (المحكم البناء) بتطبيقه على العينة الاستطلاعية، وذلك باستخدام طريقة التجزئة النصفية باستخدام كل من معادلة "سبيرمان – براون"، معادلة "جتمان" ، وطريقة تحليل التباين باستخدام معادلة " كيودر– ريتشاردسون"، حيث كانت قيم معاملات الثبات (٠.٨٧، ٠.٨٥ ، ٠.٨٣) على الترتيب، وجميعها دالة عند مستوى دلالة ٠.٠١، وهى قيم مرتفعة مما يعطي مؤشراً جيداً على ثبات الاختبار.

### [٢]:بناء الاختبار التحصيلي لمقرر الاختبارات والمقاييس (المخالف لقواعد صياغة الفقرات):

بعد إعداد الاختبار المحكم البناء تم إعداد نموذج الاختبار المخالف لقواعد الصياغة عن طريق إدخال المخالفات الخمسة على بدائل الاختبار المحكم، وهذه المخالفات وهي: اختلاف طول البدائل الصحيحة، البدائل غير المعقولة منطقياً (ظاهرياً)، وضع البدائل بشكل أفقي وليس عمودي، وجود البديل "جميع ما ذكر" كإجابة صحيحة، وجود البديل "لا شيء مما ذكر" كإجابة صحيحة.

ولقد تم توزيع تلك المخالفات بشكل عشوائي على فقرات نموذج الاختبار المخالف لقواعد الصياغة على النحو التالي:

- اختلاف طول البدائل الصحيحة: وتمثلها أرقام الفقرات ( ١ ، ٦ ، ١١ ، ١٦ ، ٢١ ، ٢٦ ، ٣١ ، ٣٦ ، ٤١ ، ٤٦ ).
- البدائل غير المعقولة منطقياً (ظاهرياً): وتمثلها أرقام الفقرات ( ٢ ، ٧ ، ١٢ ، ١٧ ، ٢٢ ، ٢٧ ، ٣٢ ، ٣٧ ، ٤٢ ، ٤٧ ).
- وضع البدائل بشكل عمودي وليس أفقي: وتمثلها أرقام الفقرات ( ٣ ، ٨ ، ١٣ ، ١٨ ، ٢٣ ، ٢٨ ، ٣٣ ، ٣٨ ، ٤٣ ، ٤٨ ).
- وجود البديل " جميع ما ذكر" كإجابة صحيحة: وتمثلها أرقام الفقرات ( ٤ ، ٩ ، ١٤ ، ١٩ ، ٢٤ ، ٢٩ ، ٣٤ ، ٣٩ ، ٤٤ ، ٤٩ ).
- وجود البديل "لا شيء مما ذكر" كإجابة صحيحة: وتمثلها أرقام الفقرات ( ٥ ، ١٠ ، ١٥ ، ٢٠ ، ٢٥ ، ٣٠ ، ٣٥ ، ٤٠ ، ٤٥ ، ٥٠ ).

### • الخصائص السيكومترية للاختبار التحصيلي لمقرر الاختبارات والمقاييس (المخالف لقواعد صياغة الفقرات):

### ♦ صدق الاختبار:

لفحص مدى تمتع الاختبار التحصيلي لمقرر الاختبارات والمقاييس بدلالات صدق كافية قام الباحث بإجراء طرق الصدق التالية:

١- صدق المحتوى (صدق المحكمين): تم التحقق من صدق المحتوى للاختبار من خلال عرض فقرات الاختبار، والأهداف السلوكية التي تقيسها، وجدول مواصفات الاختبار على عشرة محكمين من المتخصصين في القياس والتقويم ( ملحق، ٣)، وقد أجمع المحكمين على كفاية وتمثيل فقرات الاختبار للمجال السلوكي الذي يقيسه الاختبار، كما أجمع المحكمين على جودة فقرات الاختبار حيث كانت نسبة اتفاقهم (١٠٠٪) ، وهذه النسبة تدل على جودة فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس وكفايتها.

٢-الاتساق الداخلي (صدق التكوين الفرضي): تم إيجاد الاتساق الداخلي للاختبار كمؤشر من مؤشرات صدق التكوين الفرضي وذلك من خلال حساب معامل الارتباط ثنائي التسلسل الحقيقي المصحح (PTBIS) بين درجات الطلاب على كل فقرة من فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس والدرجة الكلية للاختبار بعد حذف درجة الفقرة من الدرجة الكلية للاختبار، ولقد تراوحت قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح بين (٠.٨١ – ٠.٩٥) بمتوسط حسابي (٠.٨٩٤) وانحراف معياري (٠.٠٣٧) وهي قيم مرتفعة دالة عند مستوى دلالة (٠.٠١)، تدل على لاتساق الداخلي للاختبار.

## ♦ معاملات الصعوبة والتمييز:

**١- معاملات الصعوبة:**

تم حساب معامل الصعوبة كلاسيكياً بحساب نسبة الطلاب الذين أجابوا بشكل صحيح عن الفقرة إلى العدد الكلي للطلاب الذين أجابوا بالفعل عن تلك الفقرة، ولقد تراوحت قيم معاملات صعوبة الفقرات لنموذج الاختبار المخالف لقواعد الصياغة ما بين (٠.٣٠– ٠.٥٩) وبمتوسط حسابي (٠.٣٨٥) وانحراف معياري (٠.٠٦٩) ، وقد حصلت الفقرة (٩) على أعلى معامل صعوبة، بينما حصلت الفقرة (٤٥) على أدنى معامل صعوبة.

**٢- معاملات التمييز:**

تم حساب معامل التمييز كلاسيكياً (اعتماداً على طريقة المقارنة الطرفية)، ولقد تراوحت قيم معاملات تمييز الفقرات لنموذج الاختبار المخالف لقواعد الصياغة ما بين (٠.٢٥–٠.٦١) بمتوسط حسابي (٠.٤٢٥) وانحراف معياري (٠.٨٥٨)، وقد حصلت الفقرة (١٨) على أعلى معامل تمييز، بينما حصلت الفقرة (٤٣) على أدنى معامل تمييز.

## ♦ ثبات الاختبار:

تم حساب ثبات الاختبار التحصيلي لمقرر الاختبارات والمقاييس (المخالف لقواعد صياغة الفقرات) بتطبيقه على العينة الاستطلاعية، وذلك باستخدام طريقة التجزئة النصفية باستخدام كل من معادلة "سبيرمان – براون"، معادلة "جتمان" ، وطريقة تحليل التباين باستخدام معامل كيودر– ريتشاردسون، حيث كانت قيم معاملات الثبات ( ٠.٨٠، ٠.٧٨ ، ٠.٧٤ ) على الترتيب، وجميعها دالة عند مستوى دلالة ٠.٠١ ، وهى قيم مرتفعة مما يعطي مؤشراً جيداً على ثبات الاختبار.

## [٣]: طرق تقدير الدرجات للاختبار التحصيلي لمقرر الاختبارات والمقاييس المحكم البناء:

بعد إعداد الاختبار التحصيلي لمقرر الاختبارات والمقاييس المحكم البناء تم إخراج ثلاث نماذج منه تختلف فقط في تعليمات الإجابة علي كل نموذج، وهذه النماذج بحسب طريقة تقدير الدرجات المستخدمة: (نموذج الطريقة التقليدية، نموذج الطريقة التجريبية، نموذج طريقة الاحتمال المقترح للإجابة الصحيحة)، وكانت تعليمات الإجابة على فقرات هذه النماذج على النحو التالي:

١- نموذج الطريقة التقليدية: وتمثلها إجابات الطلاب على النموذج المحكم البناء، حيث يطلب من الطالب في هذه الطريقة أن يختار أحد بدائل فقرة الاختيار من متعدد لتعبر عن إجابة هذه الفقرة، فإذا كان البديل الذي تم اختياره صحيحاً فإن درجة الطالب على هذه الفقرة تكون (١) أما إذا كان البديل الذي تم اختاره خاطئاً فإن درجته ستكون (صفراً).

٢- نموذج الطريقة التجريبية: في هذه الطريقة تختلف إجابة الطالب ودرجته على الفقرة، بحسب مدى ثقته بمعرفة البديل الصحيح، فإذا كان الطالب متأكداً من معرفة البديل الصحيح، فإنه يضع أمامه (١) وكان هذا البديل هو البديل الصحيح يحصل الطالب على ثلاث درجات، وإذا كان الطالب يشك في صحة بديلين فإنه يضع أمام أحدهما (١) وأمام الثاني (٢) وكان أحد البديلين هو البديل الصحيح يحصل الطالب على درجتين، وإذا كان الطالب يشك في صحة ثلاث بدائل فإنه يضع أمام أحدهما (١) وأمام الثاني (٢) وأمام الثالث (٣) وكان أحدهم هو البديل الصحيح يحصل الطالب على درجة واحدة، ويحصل الطالب على درجة (صفر) إذا لم يقع البديل الصحيح ضمن البدائل التي اختارها، أو إذا قام الطالب باختيار جميع بدائل الفقرة.

٣- نموذج طريقة الاحتمال المقترح للإجابة الصحيحة: في هذه الطريقة يقوم الطالب بإعطاء نسب مئوية تعبر عن مدى تقديره لصحة كل بديل من بدائل فقرة الاختيار من متعدد، بحيث يكون مجموع هذه النسب مساويا لـ ١٠٠ ٪، ويتم تقدير درجة المفحوص بأخذ النسبة المئوية التي اقترحها للبديل الصحيح لتعبر عن درجته على الفقرة.

وحيث أن نموذج الطريقة التقليدية يمثل نموذج الاختبار المحكم البناء الذي تم التحقق منه مسبقاً، لذا سوف يتم التحقق من الخصائص السيكومترية للاختبار التحصيلي وفقاً لطريقتي تقدير الدرجات (الطريقة التجريبية، طريقة الاحتمال المقترح للإجابة الصحيحة) على النحو التالي:

- **الخصائص السيكومترية للاختبار التحصيلي لمقرر الاختبارات والمقاييس وفقاً لطريقتي تقدير الدرجات(التجريبية ،الاحتمال المقترح للإجابة الصحيحة):**

♦ **صدق الاختبار:**

لفحص مدى تمتع الاختبار التحصيلي لمقرر الاختبارات والمقاييس وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة) بدلالات صدق كافية قام الباحث بإجراء طرق الصدق التالية:

١- صدق المحتوى (صدق المحكمين): تم التحقق من صدق المحتوى للاختبار من خلال عرض فقرات الاختبار وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة)، والأهداف السلوكية التي تقيسها، وجدول مواصفات الاختبار على عشرة محكمين من المتخصصين في القياس والتقويم ( ملحق، ٣)، وقد أجمع المحكمين على كفاية وتمثيل فقرات الاختبار للمجال السلوكي الذي يقيسه الاختبار، كما أجمع المحكمين على جودة فقرات الاختبار حيث كانت نسبة اتفاقهم (١٠٠٪) ، وهذه النسبة تدل على جودة فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس وكفايتها.

٢- الاتساق الداخلي(صدق التكوين الفرضي): تم إيجاد الاتساق الداخلي للاختبار وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة) كمؤشر من مؤشرات صدق التكوين الفرضي وذلك من خلال حساب معامل الارتباط ثنائي التسلسل الحقيقي المصحح(PTBIS) بين درجات الطلاب على كل فقرة من فقرات الاختبار والدرجة الكلية للاختبار بعد حذف درجة الفقرة من الدرجة الكلية للاختبار للعينة الاستطلاعية.

**جدول (٣)**

**قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح بين كل فقرة والدرجة الكلية للاختبار وفقاً لطريقتي تقدير الدرجات(التجريبية،الاحتمال المقترح للإجابة الصحيحة) للعينة الاستطلاعية**

| طريقة تقدير الدرجات | عدد الفقرات | أدنى قيمة لمعامل الارتباط | أعلى قيمة لمعامل الارتباط | المتوسط | الانحراف المعياري |
|---|---|---|---|---|---|
| الطريقة التجريبية | ٥٠ | ٠.٨١ | ٠.٩٦ | ٠.٨٩٦ | ٠.٠٤٠ |
| طريقة الاحتمال المقترح للإجابة الصحيحة | ٥٠ | ٠.٧٩ | ٠.٩٥ | ٠.٨٩١ | ٠.٠٤٤ |

يتضح من جدول (٣) السابق أن قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح قد تراوحت بين (٠.٨١ – ٠.٩٦) بمتوسط حسابي (٠.٨٩٦) وانحراف معياري (٠.٠٤٠) للطريقة التجريبية، وأن قيم معاملات الارتباط ثنائي التسلسل الحقيقي المصحح قد تراوحت بين (٠.٧٩– ٠.٩٥) بمتوسط حسابي (٠.٨٩١) وانحراف معياري (٠.٠٤٤) لطريقة الاحتمال المقترح للإجابة الصحيحة؛ وهي قيم مرتفعة ودالة إحصائياً عند مستوى دلالة (٠.٠١)، تدل على قوة ارتباط درجة الفقرات بالدرجة الكلية للاختبار، وبالتالي الاتساق الداخلي لفقرات الاختبار.

## ♦ معاملات الصعوبة والتمييز:

**١-معاملات الصعوبة:**

تم حساب معامل الصعوبة كلاسيكياً بحساب النسبة المئوية من الطلاب الذين أجابوا بالفعل عن الفقرة إجابة صحيحة، ولقد تراوحت قيم معاملات صعوبة الفقرات وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة) كانت مناسبة في ضوء المعايير السابقة لمعاملات الصعوبة، حيث تراوحت ما بين (٠.٣١-٠.٦٦) بمتوسط (٠.٤٥٦)، وانحراف معياري (٠.٠٩١) للطريقة التجريبية، كما أن قيم معاملات الصعوبة قد تراوحت ما بين (٠.٣٢-٠.٦١) بمتوسط (٠.٤٥٢)، وانحراف معياري (٠.٠٧٧) لطريقة الاحتمال المقترح للإجابة الصحيحة؛ وهذه القيم قريبة من القيمة المثالية للصعوبة (٠.٥) والتي تجعل تباين الفقرة يصل إلى أقصى ما يمكن.

**٢-معاملات التمييز:**

تم حساب معامل التمييز كلاسيكياً (اعتماداً على طريقة المقارنة الطرفية)، ولقد تراوحت قيم معاملات تمييز الفقرات وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة) كانت مناسبة في ضوء المعايير السابقة لمعاملات التمييز، حيث تراوحت قيم معاملات تمييز ما بين (٠.٣٤-٠.٦٦) بمتوسط حسابي ( ٠.٤٨١)، وانحراف معياري (٠.٠٩٩) للطريقة التجريبية، كما أن قيم معاملات التمييز قد تراوحت ما بين (٠.٣٣-٠.٦٣) بمتوسط حسابي (٠.٤٧٧)، وانحراف معياري (٠.٠٩٤) لطريقة الاحتمال المقترح للإجابة الصحيحة؛ وأن معظم القيم قريبة من القيمة (٠.٤)، مما يدل على أن فقرات الاختبار جيدة جداً وفقا لنفس المعيار.

## ♦ ثبات الاختبار:

تم حساب ثبات الاختبار التحصيلي لمقرر الاختبارات والمقاييس وفقاً لطريقتي تقدير الدرجات (التجريبية، الاحتمال المقترح للإجابة الصحيحة) للعينة الاستطلاعية، وذلك باستخدام طريقة التجزئة النصفية باستخدام كل من معادلة "سبيرمان – براون"، معادلة "جتمان" ، وطريقة تحليل التباين باستخدام معامل "ألفا- كرونباخ"، ولقد تراوحت قيم معاملات الثبات ما بين (٠.٧٨-٠.٨٧) للطريقة التجريبية، كما أن قيم معاملات الثبات قد تراوحت ما بين (٠.٧٩-٠.٨٨) لطريقة الاحتمال المقترح للإجابة الصحيحة؛ وجميعها دالة عند مستوى دلالة ٠.٠١ ، وهى قيم مرتفعة مما يعطي مؤشراً جيداً على ثبات الاختبار.

## ثالثاً المعالجة الإحصائية:

تم استخدام الأساليب الإحصائية الآتية فى معالجة النتائج التي تم الحصول عليها بعد تطبيق أداة الدراسة على عينة الدراسة الأساسية وهى: (المتوسطات الحسابية، الانحرافات المعيارية، الخطأ المعياري، معامل الارتباط ثنائي التسلسل الحقيقي، معاملات الصعوبة والتمييز والتخمين، اختبار " ت " ، التحليل العاملي) ، وقد تم استخدام جميع الأساليب الإحصائية من خلال حزمة البرامج الإحصائية الاجتماعية برنامج SPSS(22) ، كما تم استخدام برنامج التحليل الإحصائي XCalibre 4.1.7 في حساب (صعوبة وتمييز وتخمين فقرات الاختبار، دالة معلومات الاختبار ككل والفقرات الاختبارية، وتقدير قدرات الأفراد).

**رابعاً: منهج الدراسة: تتبع الدراسة الحالية المنهج الوصفي المقارن.**

**نتائج الدراسة وتفسيرها:**

**[ ١ ] – نتائج التساؤل الأول وتفسيرها:**

والذي ينص على أنه " هل تتحقق افتراضات نظرية الاستجابة للمفردة الاختبارية على استجابات أفراد عينة الدراسة على نموذجي الاختبار(المحكم ، المخالف) لقواعد صياغة الفقرات؟"

وللإجابة عن التساؤل السابق تم التحقق من افتراضات نظرية الاستجابة للمفردة والمتمثلة في أحادية البعد، والاستقلال الموضعي، والتحرر من السرعة، وذلك على النحو التالي:

**أ- التحقق من افتراض أحادية البعد:**

للتحقق من افتراض أحادية البعد تم إجراء التحليل العاملي على إجابات أفراد عينة الدراسة عن فقرات نموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات وذلك لمعرفة عدد العوامل التي يزيد قيم الجذر الكامن لها عن الواحد الصحيح باستخدام برنامج SPSS(22) على النحو التالي:

**أولاً: التحقق من شروط إجراء التحليل العاملي: تم التحقق من مدى ملاءمة بيانات عينة الدراسة لإجراء التحليل العاملي عليها على النحو التالي:**

١-التحقق من مدى كفاية حجم العينة لإجراء التحليل العاملي: ويتم ذلك عن طريق استخدام اختبار كايزر – ماير – أولكن Kaiser –Meyer – Olkin (KMO) والذي يجب أن لا يقل عن (٠.٥) حسب محك كيزر.

٢-التحقق من تجانس العينة واختبار فرضية عدم تماثل مصفوفة الارتباط الأصلية: ويتم ذلك عن طريق دلالة قيمة مربع كاى ($\chi^2$) لاختبار بارتليت Bartlett's Test of Sphericity، بمعني أن تكون مصفوفة معاملات الارتباط ليست على صورة مصفوفة الوحدة.

٣-أن تكون القيمة المطلقة لمحدد مصفوفة معاملات الارتباط أكبر من ٠.٠٠٠٠١ ، وهذا يدل علي وجود اعتماد خطي Linear Dependency يحجب المساهمة الخاصة لكل متغير في تحديد عدد العوامل.

٤-أن تكون قيم توافق العينة من خلال أزواج المتغيرات الثنائية والتي يتم الحصول عليها من اختبار كايزر (MSA) مقبولة أكبر من ٠.٥.

حيث أتضح تحقق شروط استخدام التحليل العاملي في بيانات هذه الدراسة؛ حيث كانت قيمة ($\chi^2$) ( ١٠٢٩٠.٨٥ للاختبار المحكم، ٩١٢٧.٩٨ للاختبار المخالف) بدرجات حرية ١٢٢٥ دالة مما يعني أن المصفوفة غير متماثلة وأن هناك علاقة بين المتغيرات وهذا يشير إلى تحقق شرط تجانس العينة ومناسبة البيانات لمتابعة إجراء التحليل العاملي، وكانت قيمة اختبار (KMO) (٠.٩٥٤ للاختبار المحكم، ٠.٩٤٥ للاختبار المخالف) دالة حيث أنها أكبر من ٠.٥، أي أن حجم عينة الدراسة كان كافياً ومناسباً بإجراء التحليل العاملي، وأن القيمة المطلقة لمحدد مصفوفة الارتباط كانت (٠.٠٠٠٢) أكبر من ٠.٠٠٠٠١ ، وهي قيمة لا تساوي الصفر ومن ثم لا تكون المصفوفة من النوع المفرد، كما اتضح ملائمة المعاينة (MSA) والموجودة في قطر مصفوفة معاملات الارتباط الصورية، حيث كانت جميع القيم الحرجة لكل فقرة أكبر من ٠.٥، بالتالي يمكن اكمال التحليل والوثوق بدرجة كبيرة في نتائجة.

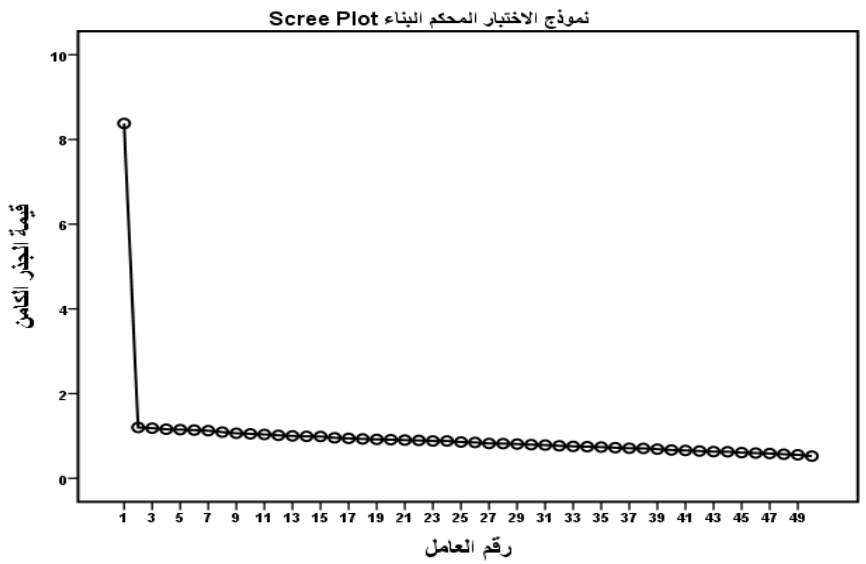
## ثانياً: إجراء التحليل العاملي لبيانات نموذجي الاختبار:

تم إجراء التحليل العاملي Factorial Analysis ، باستخدام طريقة " المكونات الأساسية " Principal Component Analysis (PCA) التي اقترحها هوتلنج Hottelling وهى من أفضل طرق التحليل العاملي من حيث الدقة ويستخلص فيها كل عامل أقصى تباين ممكن، كما تم إجراء التدوير المتعامد للمحاور (العوامل) باستخدام طريقة " الفاريماكس " Varimax Rotation، للكشف عن البنية العاملية الكامنة وللتحقق من أحادية البعد للبيانات، من خلال التأكد من وجود عامل واحد مسيطر، والذي يفترض أن الاختبار يقيسه، وقد تم استخدام محك " كايزر" Kaiser ، الذي اقترحه "جتمان" بأخذ العوامل التي جذرها الكامن Eigenvalue يساوي أو أكبر من الواحد الصحيح، من أجل تحقيق النقاء والوضوح السيكولوجي لتشبعات الفقرات على العوامل وذلك كما ذكر صفوت فرج (١٩٩١) بواسطة حزمة البرامج الإحصائية فى العلوم الاجتماعية SPSS(22) ، وذلك للبيانات المتعلقة باستجابات فقرات عينة الدراسة الأساسية المكونة من (١٥٠٠) طالباً وطالبة بكلية التربية – جامعة أم القرى بمكة المكرمة، وقد تم اعتماد عدة محكات يمكن أن يستدل من خلالها على تحقق افتراض أحادية البعد على النحو التالي:

١- أفرز التحليل العاملي لنموذج الاختبار الأول المحكم (١٣) عاملاً، قيمة الجذر الكامن لكل منها تزيد عن الواحد، وتفسر مجتمعة ما نسبته (٤٣.١٦٣ ٪) من التباين الكلي للاختبار، حيث كانت قيمة الجذر الكامن للعامل الأول (٦.٥٥٨) ، ويفسر ما نسبته (١٣.١١٥ ٪) من التباين الكلي للاختبار، وقد كانت قيمة الجذر الكامن للعامل الثاني (١.٩٤١) ، ويفسر ما نسبته (٣.٨٨٣ ٪)، وقد أفرز التحليل العاملي لنموذج الاختبار الثاني المخالف لقواعد الصياغة (١٤) عاملاً، قيمة الجذر الكامن لكل منها تزيد عن الواحد، وتفسر مجتمعة ما نسبته (٤٤.١٨٧ ٪) من التباين الكلي للاختبار، حيث كانت قيمة الجذر الكامن للعامل الأول (٥.٣٦٧ ) ، ويفسر ما نسبته (١٠.٧٣٤ ٪) من التباين الكلي للاختبار، وقد كانت قيمة الجذر الكامن للعامل الثاني (٢.٠٤٥) ، ويفسر ما نسبته (٤.٠٩٠ ٪) ؛ وتُعتَمد في التحليل العاملي أحادية البعد من خلال نسبة الجذر الكامن للعامل الأول إلى الجذر الكامن للعامل الثاني، بحيث تكون هذه النسبة لا تقل عن (٢) ، وقد كان ناتج قسمة قيمة الجذر الكامن للعامل الأول، على قيمة الجذر الكامن للعامل الثاني يساوي (٣.٣٨) للاختبار المحكم البناء، و(٢.٦٣) للاختبار المخالف لقواعد الصياغة، وهذه النسبة تزيد عن المعيار (٢) (Georgiev, 2008; Hambleton& Swaminathan, 1985; Nering & Ostini, 2010; Reise & Waller, 2003).

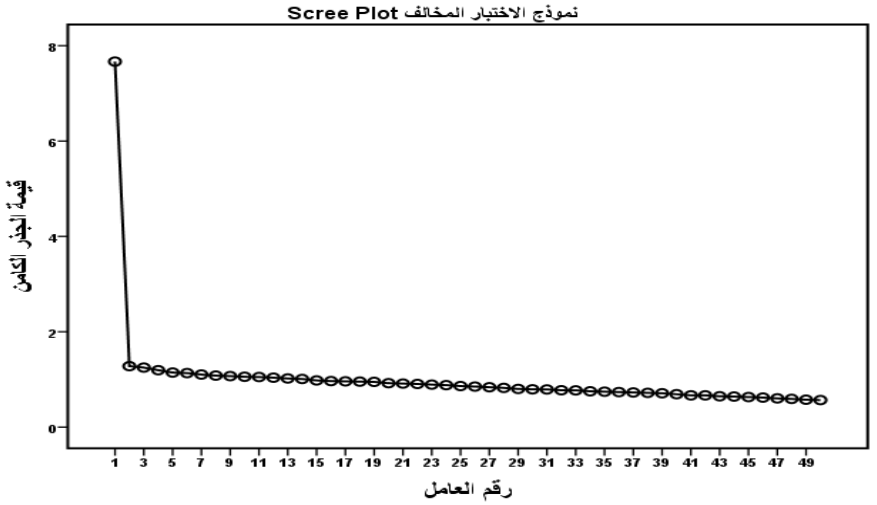
٢- عند النظر إلى نسبة الفرق بين الجذر الكامن للعامل الأول والجذر الكامن للعامل الثاني، , إلى الفرق بين الجذر الكامن للعامل الثاني والجذر الكامن للعامل الثالث في كل من الاختبار المحكم والمخالف، تبين أن النسبة كبيرة، وأن النسبة بين بقية الجذور الكامنة المتتالية الأخرى كانت متقاربة؛ بمعنى أنه يوجد شبه استقرار في نسب التباين المفسر لجميع العوامل باستثناء العامل الأول، وهذا مؤشر على تحقق افتراض أحادية البعد للإختبار (Hambleton & Swaminthan, 1985; Hambleton et al., 1991; Onder, 2007; Reise & Revicki, 2015).

وقد تم تمثيل الجذور الكامنة للعوامل جميعها بيانياً لنموذجي الاختبار (المحكم، المخالف) لقواعد صياغة الفقرات، عن طريق اختبار الفحص البياني Scree Plot ؛ حيث يلاحظ وجود عامل سائد لنموذجي الاختبار (المحكم ، المخالف) على بقية العوامل الأخرى، وهذا ما أكده كل من (Gorsuch, 1983; Field, 2009) من أن اختبار المنحدر لكاتيل Kattell's Scree test يكون أكثر دقة عندما تكون العينة كبيرة، كما هو موضح بالشكلين (١) ، (٢).




**شكل (١)**

**التمثيل البياني لقيم الجذور الكامنة للعوامل المكونة لنموذج الاختبار المحكم البناء**




**شكل (٢)**

**التمثيل البياني لقيم الجذور الكامنة للعوامل المكونة لنموذج الاختبار المخالف لقواعد صياغة الفقرات**

ويمكن الاعتماد على حساب معاملات الارتباط بين درجات كل فقرة من فقرات الاختبار والدرجة الكلية للاختبار، وذلك ضمن الافتراض بأنه إذا كانت ارتباطات معظم الفقرات بالدرجة الكلية للاختبار تزيد عن (٠.٢) فإن ذلك مؤشرٌ على أحادية البعد (Hattie,1985; Reise & Revicki, 2015)، حيث كانت جميع قيم معاملات الارتباط دالة إحصائياً عند مستوى (٠.٠١)، وتراوحت قيم معاملات الارتباط للنموذج المحكم بين (٠.٧٢- ٠.٨٩)، بينما تراوحت قيم معاملات الارتباط للنموذج المخالف بين ( ٠.٧٠- ٠.٨٧) ، ومن ثم فإن جميع الفقرات تجاوزت قيم معاملات ارتباطها (٠.٢) ؛ مما يشير إلى أن هذه الفقرات تتشارك في قياس بعد واحد تعبر عنه الدرجة الكلية، وبذلك يمكن اعتبار أن نموذجي الاختبار قد حققا افتراض أحادية البعد.

كما تم استخدام طريقة تحليل البواقي من نماذج نظرية الاستجابة للفقرة الاختبارية أحادية البعد، وذلك للتحقق من افتراض أحادية البعد لبيانات الدراسة، حيث تم تحليل البيانات باستخدام برنامج نوهارم3 NOHARM ، حيث تم الكشف عن أُحادية البعد من خلال مؤشري الملائمة الإحصائية التاليين:

**أولاً: مؤشِر تاناكا Tanaka's Index of Goodness** وهو مؤشِر يدل على حسن المطابقة ما بين النَموذج المستخدم والبيانات، ويعمل كمعامل تحديد وتلخيص لنسبة التباين المفسَّر بواسطة النَموذج، ومعادلته هي:

$$Y_{ULS} = 1 - \frac{T_r\left(\underline{R}^2\right)}{T_r\left(\underline{C}^2\right)}$$

ويعتبر مؤشِر تاناكا دليلاً على المستوى المقبول من المطابقة ما بين النَموذج والبيانات إذا بلغت قيمته 0.95 فأكثر، أما المطابقة التامة بين النَموذج والبيانات فتحصل عندما تبلغ قيمته واحد صحيح (Jasper, 2010)، وفي الدراسة الحالية كانت القيمة تساوي (٠.٩٨٥) للاختبار المحكم، وتساوي (٠.٩٧٨) للاختبار المخالف، ومن ثم فإن مؤشِر تاناكا لعينة المعايرة قد تحقق فيه المستوى الجيد من المطابقة ما بين النَموذج والبيانات لزيادة قيمته عن ٠.٩٥ .

**ثانياً: مؤشر جذر متوسط مربعات البواقي:** (Root Mean Square of Residuals, RMSR) ، ويوفر برنامج NOHARM مصفوفة البواقي بهدف إجراء عملية مطابقة البيانات للنموذج، وتعبِّر القيم في هذه المصفوفة عن الفروق بين التباينات المصاحبة المشاهدة، والتباينات المصاحبة الناتجة من إجراء مطابقة البيانات للنموذج، وعليه تكون المطابقة تامة إذا كانت الفروق بينها مساوية للصفر، بعد ذلك يقوم البرنامج بتلخيص مصفوفة البواقي عن طريق حساب جذر متوسط مربعات البواقي RMSR، وهكذا فإن القيمة المنخفضة لهذا المقدار هي مؤشِر على حسن المطابقة، ويمكن تقدير قيمة هذا المؤشِر كما ذكر (Fraser & McDonald, 1988) بمقارنة قيمة هذا المؤشر مع قيمة الخطأ المعياري للبواقي كمعيار، والتي يتم حسابها من خلال قسمة القيمة ٤.١ على الجذر التربيعي لحجم العينة.

وفي الدراسة الحالية كانت قيمة مؤشر جذر متوسط مربعات البواقي RMSR تساوي (٠.٠٠٨) للاختبار المحكم، وتساوي (٠.٠١١) للاختبار المخالف وهى قيم صغيرة جداً وقريبة من الصفر، كما أنها أقل من القيمة الحرجة (نقطة القطع لاحتمالية قبول قيمة الإحصائي) التي حسبت من المعادلة ( $\frac{4.1}{\sqrt{n}}$ حيث n: ١٥٠٠) والبالغة ( ٠.١٠٦).

**ب- التحقق من افتراض الاستقلال الموضعي:**

يقصد بالاستقلال الموضعي أنه عند مستوى قدرة معين فإنه لا يوجد ارتباط بين احتمالية إجابة الأفراد على سؤال ما إجابة صحيحة واحتمالية إجابتهم إجابة صحيحة على سؤال آخر، ولهذا فقد أطلق على هذا الافتراض الاستقلال الشرطي (استجابة الفرد على الفقرات في المقياس مستقلة احصائياً)، ويرى كل من (DeMars, 2010; Hambleton et al.,1991; Hulin, Drasgow & Parsons, 1983; Raykov & Marcoulides, 2016; Reise & Revicki, 2015) أن هذا الشرط يتحقق ضمنياً بتحقق شرط أحادية البعد؛ حيث أن هناك ارتباطاً وثيقاً بين تحقق افتراض أحادية البعد وتحقق افتراض الاستقلال الموضعي.

كما تم التحقق من افتراض الاستقلال الموضعي من خلال مؤشر $\mathbf{Z_{Q_3}}$، والذي يتم حسابه وفقاً للخطوات الآتية:

**أولاً: حساب مؤشر ين (Yen's Index)** المعروف بـ $\mathbf{Q_3}$، ويُعرف على أنه معامل ارتباط بيرسون للبواقي الناتجة من النَموذج المتعلق بنظرية الاستجابة للمفردة بين زوج من الفقرات بعد ضبط السمة المقدّرة، ويتم حساب قيمة الباقي لاستجابة المفحوص على الفقرة وفقاً للمعادلة:

$$d_{ij} = u_{ij} - T_j(\hat{\theta}_i)$$

وعند إعطاء الوزن (٠) للإجابة الخاطئة، والوزن (١) للإجابة الصحيحة في حالة البيانات ثنائية الاستجابة كما هي الحالة في الدِراسة الحالية تكون:

$$T_j(\hat{\theta}_i) = P_j(\hat{\theta}_i)$$

وبحساب جميع قيم البواقي لاستجابات جميع المفحوصين عند كل مستوى قدرة مقدَّرة على فقرتين من فقرات الاختبار مثل $\mathbf{j}'$،$\mathbf{j}$ يكون:

$$Q_{3jj'} = r(d_j, d_{j'})$$

وفي الدراسة الحالية تم حساب مؤشر ($Q_3$) لين لفحص الاستقلال الموضعي لأزواج فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس؛ حيث تم تحليل البيانات باستخدام برمجية Local (LDID) الذي وضعه كل من (Kim, Cohen & Lin, 2006)، وإيجاد معامل الارتباط بين البواقي لأزواج فقرات إحصائيًا نموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، وذلك بعد معايرة البيانات باستخدام النموذج اللوجستي ثلاثي البارامتر باستخدام برمجية XCalibre 4.1.7 ، وقد كانت جميع قيم معامل الارتباط بين البواقي لجميع أزواج فقرات الاختبار أقل من (٠.١٦٤) وهي قيمة أقل من درجة القطع التي افترضها (ين) والبالغة (٠.٢٠) مما يعني تحقق افتراض الاستقلال الموضعي لأزواج فقرات الاختبار.

**ثانياً: حساب قيم مؤشر** $Z_{Q_3}$ من خلال عمل تحويل فشر Fisher Transforming لقيم $Q_3$ وفقاً للمعادلة:

$$Z_{Q_3} = \frac{1}{2} Ln \frac{1 + Q_3}{1 - Q_3}$$

وللحكم على تحقق الاستقلال الموضعي لفقرتين يجب أن تقع قيمة $Z_{Q_3}$ المحسوبة لهاتين الفقرتين ضمن فترة ثقة بانحرافين معياريين عن المتوسط الحسابي لقيم $Z_{Q_3}$ المحسوبة.

وإذا كانت أزواج الفقرات التي تتحقق بها الاستقلالية الموضعيَّة أكبر منها لأزواج الفقرات التي تتحقق بها الاعتمادية فيعتبر هذا مؤشر على تحقق الاستقلال الموضعي للاختبار ككل.

وفي الدراسة الحالية تم حساب مؤشر $Z_{Q_3}$ لفحص الاستقلال الموضعي لأزواج فقرات الاختبار التحصيلي لمقرر الاختبارات والمقاييس؛ حيث تم تحليل البيانات باستخدام برمجية LDID لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، حيث تم الاعتماد على نتائج فترة الثقة لقيم $Q_3$ المحولة إلى قيم Z فشر المناظرة لها، فقد تم رصد عدد أزواج الفقرات التي وقعت قيمة $Z_{Q_3}$ لها ضمن فترة الثقة المحققة لشرط الاستقلال الموضعي على أنها إما معتمدة أو مستقلة؛ حيث كانت عدد أزواج الفقرات التي وقعت خارج مدى فترة الثقة (١٥٦) زوجاً ؛ أي ما نسبته(١٢.٧٤٪) من عدد الأزواج الكلي (١٢٢٥)، بينما كان عدد أزواج الفقرات التي وقعت ضمن مدى فترة الثقة (١٠٦٩) زوجاً؛ أي ما نسبته (٨٧.٢٦٪) من عدد الأزواج الكلي للاختبار المحكم، وكانت عدد أزواج الفقرات التي وقعت خارج مدى فترة الثقة (٢٢١) زوجاً ؛ أي ما نسبته (١٨.٠٤٪) من عدد الأزواج الكلي (١٢٢٥)، بينما كان عدد أزواج الفقرات التي وقعت ضمن مدى فترة الثقة (١٠٠٤) زوجاً؛ أي ما نسبته (٨١.٩٦٪) من عدد الأزواج الكلي للاختبار المخالف، وهذا يبين أن عدد أزواج الفقرات التي حققت الاستقلالية الموضعية أعلى بكثير من عدد أزواج الفقرات التي حققت التبعية الموضعية، وهذا مؤشر على تحقق افتراض الاستقلال الموضعي لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات.

**ج- التحقق من افتراض التحرر من السرعة:**

يعد وجود عامل واحد رئيس يقع خلف الاستجابة على فقرات الاختبار مؤشراً على أن عامل السرعة ليس عاملاً مؤثراً في الاستجابة على فقرات الاختبار؛ حيث يري (Hambleton, 2004; Hambleton & Swaminathan, 1985; Ueckert, 2018) أن هناك افتراض أساس عام لجميع نماذج نظرية الاستجابة للمفردة، وهو أن الاختبار الذي يسعى النموذج لمطابقة بياناته لم يتم تطبيقه تحت ظرف السرعة، بمعنى أن الأفراد الذين أخفقوا في الإجابة على فقرات الاختبار لم يكن ذلك بسبب إخفاقهم في السرعة الكافية لإنجاز الاختبار، وإنما يعود ذلك إلى محدودية قدراتهم، كما ولقد راعى الباحث أثناء تطبيقه للاختبار إعطاء الطلاب الوقت الكافي للانتهاء من الإجابة عن فقرات الاختبار.

## [ ٢ ] – نتائج التساؤل الثاني وتفسيرها:

والذي ينص على أنه " ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات صعوبة الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر؟"، وللإجابة عن هذا التساؤل تمت الاجابة عن التساؤلات الفرعية التالية:

**أولاً:** ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات صعوبة الفقرات في ضوء القياس الكلاسيكي؟"

للإجابة عن التساؤل الفرعي السابق تم حساب مؤشرات معامل الصعوبة كلاسيكياً، ولقد تراوحت قيم معاملات صعوبة الفقرات، وفقًا لإجابات الطلاب على نموذج الاختبار محكم البناء ما بين (٠.٣٠٩- ٠.٦٢٧) بمتوسط حسابي (٠.٤٠٩) وانحراف معياري (٠.٠٦٣)، وقد حصلت الفقرة (٣٦) على أعلى معامل صعوبة، بينما حصلت الفقرة (٢٣) على أدنى معامل صعوبة، بينما تراوح قيم معاملات صعوبة الفقرات وفقًا لإجابات الطلاب على نموذج الاختبار المخالف لقواعد الصياغة ما بين (٠.٣٠١-٠.٥٦١) بمتوسط حسابي ( ٠.٣٦٦) وانحراف معياري (٠.٠٥٠)، وقد حصلت الفقرة (٦) على أعلى معامل صعوبة، بينما حصلت الفقرة (٤١) على أدنى معامل صعوبة.

وللكشف عن الفروق في معامل صعوبة الفقرات تبعاً لنموذجي الاختبار (المحكم ، المخالف) تم إجراء الاختبار الإحصائي (T-test) لاختبار دلالة الفروق بين متوسطي صعوبة فقرات الاختبار المحكم والمخالف لقواعد الصياغة، كما هو موضح في الجدول (٤) التالي:

**جدول(٤)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي صعوبة فقرات الاختبار (المحكم والمخالف) كلاسيكياً**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٠.٤٠٩ | ٠.٠٦٣ | ٤.٣٠٦ | ٠.٠٠٠ دالة عند مستوى (٠.٠١=α) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٠.٣٦٦ | ٠.٠٥٠ | | |

يتضح من جدول (٤) السابق أن قيمة (ت=٤.٣٠٦) دالة إحصائية عند مستوى (٠.٠١=α) وتؤكد هذه النتيجة على أن متوسط صعوبة فقرات الاختبار المحكم كان أعلى من متوسط صعوبة فقرات الاختبار المخالف، أي أن فقرات الاختبار المحكم أسهل من فقرات الاختبار المخالف، وهذا بدوره يعطي أهمية لاتباع قواعد صياغة فقرات الاختيار من متعدد؛ حيث تبين تأثر صعوبة فقرات الاختبار المخالف لقواعد صياغة الفقرة بإدخال الانتهاكات عليها، وهذه النتيجة تتفق مع دراسة (ابتسام عيسى خصاونة، ٢٠١٢؛ Chang et al., 2007)، والتي أظهرت أن الفقرات المتضمنة للمخالفات في صياغتها، أقل صعوبة وبدرجة دالة إحصائيًا مقارنةً بالفقرات المتحررة من هذه المخالفات محكمة البناء.

**ثانياً:** ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات صعوبة الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير الخطأ المعياري لمتوسط معاملات صعوبة الفقرات تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟"

للإجابة عن هذا التساؤل الفرعي تم استخدام برنامج XCalibre 4.1.7 لتحليل بيانات كل من نموذجي الاختبار لتقدير معالم صعوبة الفقرات، والجدولان (٥) ، (٦) التاليين يوضحان تقديرات معالم صعوبة الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر وذلك لنموذجي الاختبار (المحكم ، المخالف) لقواعد الصياغة، كما تم تحويل الدرجة باللوجيت إلى وحدة المنف.

**جدول (٥)**

**معالم صعوبة الفقرات وتقدير الخطأ المعياري لمعلمة الصعوبة وفق النموذج الثلاثي البارامتر لنموذج الاختبار المحكم البناء**

| رقم الفقرة | معلمة الصعوبة باللوجيت | معلمة الصعوبة بالمنف | الخطأ المعياري باللوجيت | الخطأ المعياري بالمنف | رقم الفقرة | معلمة الصعوبة باللوجيت | معلمة الصعوبة بالمنف | الخطأ المعياري باللوجيت | الخطأ المعياري بالمنف |
|---|---|---|---|---|---|---|---|---|---|
| ١ | ٠.٨٧٨ | ٥٤.٣٩ | ٠.٠٧٠ | ٠.٣٥ | ٢٦ | ١.١٧٣ | ٥٥.٨٧ | ٠.٠٧٧ | ٠.٣٩ |
| ٢ | ١.٠٨٦ | ٥٥.٤٣ | ٠.٠٦٨ | ٠.٣٤ | ٢٧ | ٠.٧٤٠ | ٥٣.٧٠ | ٠.٠٤٥ | ٠.٢٣ |
| ٣ | ١.٤٦٥ | ٥٧.٣٣ | ٠.١٣٠ | ٠.٦٥ | ٢٨ | ١.٣٧١ | ٥٦.٨٦ | ٠.٤٥٤ | ٢.٢٧ |
| ٤ | ١.٥٢١ | ٥٧.٦١ | ٠.١٨٠ | ٠.٩٠ | ٢٩ | ٠.٦٥٧ | ٥٣.٢٩ | ٠.٣١٧ | ١.٥٩ |
| ٥ | ٠.٥٢٩ | ٥٢.٦٥ | ٠.٠٥٣ | ٠.٢٧ | ٣٠ | ١.١٧٣ | ٥٥.٨٧ | ٠.٣٦٠ | ١.٨٠ |
| ٦ | ٠.٦٣٤ | ٥٣.١٧ | ٠.٣٠٧ | ١.٥٤ | ٣١ | ١.٣٢٦ | ٥٦.٦٣ | ٠.٤٠٩ | ٢.٠٥ |
| ٧ | ٠.٦٥٢ | ٥٣.٢٦ | ٠.٠٥٤ | ٠.٢٧ | ٣٢ | ١.٧٣٧ | ٥٨.٦٩ | ٠.٥٧٦ | ٢.٨٨ |
| ٨ | ٠.٣٤٧ | ٥١.٧٤ | ٠.٠٩٠ | ٠.٤٥ | ٣٣ | ٠.٩٩٢ | ٥٤.٩٦ | ٠.٣٤٧ | ١.٧٤ |
| ٩ | ١.٩٣٤ | ٥٩.٦٧ | ٠.١٢٣ | ٠.٦٢ | ٣٤ | ١.٨٨٩ | ٥٩.٤٥ | ٠.٢٢٩ | ١.١٥ |
| ١٠ | ١.٣٧٨ | ٥٦.٨٩ | ٠.٤٣٢ | ٢.١٦ | ٣٥ | ١.٧٢٤ | ٥٨.٦٢ | ٠.٥٦٩ | ٢.٨٥ |
| ١١ | ١.٥٦٢ | ٥٧.٨١ | ٠.٠٨٧ | ٠.٤٤ | ٣٦ | ١.٣٣٢ | ٥٦.٦٦ | ٠.٤١٣ | ٢.٠٧ |
| ١٢ | ١.٤٢٢ | ٥٧.١١ | ٠.٤٤٢ | ٢.٢١ | ٣٧ | ١.٠٨٦ | ٥٥.٤٣ | ٠.٠٦٨ | ٠.٣٤ |
| ١٣ | ٠.٩٨٧ | ٥٤.٩٤ | ٠.٠٨٨ | ٠.٤٤ | ٣٨ | ١.٦١٨ | ٥٨.٠٩ | ٠.٥١٢ | ٢.٥٦ |
| ١٤ | ١.٣٥٢ | ٥٦.٧٦ | ٠.٤٠٣ | ٢.٠٢ | ٣٩ | ٠.٦٥٧ | ٥٣.٢٩ | ٠.٩٨٠ | ٤.٩٠ |
| ١٥ | ١.٦٥٠ | ٥٨.٢٥ | ٠.٥٢٣ | ٢.٦٢ | ٤٠ | ١.٢٣٧ | ٥٦.١٩ | ٠.٣٨٨ | ١.٩٤ |
| ١٦ | ٠.٧٩٩ | ٥٣.٩٩ | ٠.٣٦٠ | ١.٨٠ | ٤١ | ١.٩٦٢ | ٥٩.٨١ | ٠.٧٧٧ | ٣.٨٩ |
| ١٧ | ١.٦٥٢ | ٥٨.٢٦ | ٠.١٨٠ | ٠.٩٠ | ٤٢ | ٠.٤٤٥ | ٥٢.٢٣ | ٠.٢٦٩ | ١.٣٥ |
| ١٨ | ٠.٨٨٨ | ٥٤.٤٤ | ٠.٠٦٥ | ٠.٣٣ | ٤٣ | ١.٦٠٨ | ٥٨.٠٤ | ٠.٥٠٣ | ٢.٥٢ |
| ١٩ | ٠.٨٤٥ | ٥٤.٢٣ | ٠.٣٦٠ | ١.٨٠ | ٤٤ | ٠.٩١٤ | ٥٤.٥٧ | ٠.٠٥٦ | ٠.٢٨ |
| ٢٠ | ٠.٦٥٨ | ٥٣.٢٩ | ٠.٠٨٢ | ٠.٤١ | ٤٥ | ١.٢٥٦ | ٥٦.٢٨ | ٠.٨٦٧ | ٤.٣٤ |
| ٢١ | ١.١٨٩ | ٥٥.٩٥ | ٠.٣٦٤ | ١.٨٢ | ٤٦ | ١.٥٤٥ | ٥٧.٧٣ | ٠.٤٨٣ | ٢.٤٢ |
| ٢٢ | ٠.٧٨٤ | ٥٣.٩٢ | ٠.٦٢٠ | ٣.١٠ | ٤٧ | ٠.٩٤٦ | ٥٤.٧٣ | ٠.٣٠١ | ١.٥١ |
| ٢٣ | ٢.٢٥١ | ٦١.٢٦ | ٠.١٥٥ | ٠.٧٨ | ٤٨ | ١.٦٠٨ | ٥٨.٠٤ | ٠.٥٠٣ | ٢.٥٢ |
| ٢٤ | ١.١٨٧ | ٥٥.٩٤ | ٠.٣٨١ | ١.٩١ | ٤٩ | ١.٢٤٦ | ٥٦.٢٣ | ٠.٠٦٩ | ٠.٣٥ |
| ٢٥ | ٠.٩١٥ | ٥٤.٥٨ | ٠.٠٥٦ | ٠.٢٨ | ٥٠ | ٢.٠٦٥ | ٦٠.٣٣ | ٠.٨٨٣ | ٤.٤٢ |

يتضح من جدول (٥) السابق تراوح قيم معاملات صعوبة الفقرات باللوجيت، وفقًا لإجابات الطلاب على نموذج الاختبار المحكم البناء ما بين (٠.٣٤٧- ٢.٢٥١) بمتوسط حسابي (١.٢١٧) وانحراف معياري (٠.٤٥١)، وتراوحت قيم الخطأ المعياري في تقدير معلمة الصعوبة للنموذج المحكم البناء ما بين (٠.٠٤٥- ٠.٩٨٠) بمتوسط حسابي (٠.٣٢٣)، وقد حصلت الفقرة (٢٣) على أعلى معامل صعوبة، بينما حصلت الفقرة (٨) على أدنى معامل صعوبة.

**جدول (٦)**
**معالم صعوبة الفقرات وتقدير الخطأ المعياري لمعلمة الصعوبة وفق النموذج الثلاثي البارامتر لنموذج الاختبار المخالف لقواعد صياغة الفقرات**

| رقم الفقرة | معلمة الصعوبة باللوجيت | معلمة الصعوبة بالمنف | الخطأ المعياري في تقدير معلمة الصعوبة باللوجيت | الخطأ المعياري في تقدير معلمة الصعوبة بالمنف | رقم الفقرة | معلمة الصعوبة باللوجيت | معلمة الصعوبة بالمنف | الخطأ المعياري في تقدير معلمة الصعوبة باللوجيت | الخطأ المعياري في تقدير معلمة الصعوبة بالمنف |
|---|---|---|---|---|---|---|---|---|---|
| ١ | ١.٦٢٤ | ٥٨.١٢ | ٠.٤٠٩ | ٢.٠٥ | ٢٦ | ١.٨١٥ | ٥٩.٠٨ | ٠.١٢٢ | ٠.٦١ |
| ٢ | ١.٦١١ | ٥٨.٠٦ | ٠.٤٢١ | ٢.١١ | ٢٧ | ١.٣٥٣ | ٥٦.٧٧ | ٠.٣٨٥ | ١.٩٣ |
| ٣ | ١.٥٧١ | ٥٧.٨٦ | ٠.٣٩٨ | ١.٩٩ | ٢٨ | ٢.٢٤٥ | ٦١.٢٣ | ٠.٤٣٣ | ٢.١٧ |
| ٤ | ١.٦٣٢ | ٥٨.١٦ | ٠.٤١٥ | ٢.٠٨ | ٢٩ | ٢.٣٧٨ | ٦١.٨٩ | ١.٠١١ | ٥.٠٦ |
| ٥ | ٣.١٣٣ | ٦٥.٦٧ | ٠.٣٧١ | ١.٨٦ | ٣٠ | ٢.٦٥٤ | ٦٣.٢٧ | ٠.٢٠٨ | ١.٠٤ |
| ٦ | ١.٤٤٩ | ٥٧.٢٥ | ٠.٣٨٥ | ١.٩٣ | ٣١ | ٢.٥٠٢ | ٦٢.٥١ | ٠.١٩٠ | ٠.٩٥ |
| ٧ | ٢.٥٤٥ | ٦٢.٧٣ | ١.٣٦٣ | ٦.٨٢ | ٣٢ | ٢.٥٠٥ | ٦٢.٥٣ | ٠.٣٨٠ | ١.٩٠ |
| ٨ | ٠.٤٤٧ | ٥٢.٢٤ | ٠.٢٨١ | ١.٤١ | ٣٣ | ٢.٣٩٠ | ٦١.٩٥ | ٠.٣٧٩ | ١.٩٠ |
| ٩ | ٢.٦٥٤ | ٦٣.٢٧ | ٠.٨١٨ | ٤.٠٩ | ٣٤ | ١.٩٠٧ | ٥٩.٥٤ | ٠.٥٦١ | ٢.٨١ |
| ١٠ | ٢.٢٣٥ | ٦١.١٨ | ٠.٤٥٧ | ٢.٢٩ | ٣٥ | ٢.٦٥٤ | ٦٣.٢٧ | ٠.٤٠٨ | ٢.٠٤ |
| ١١ | ١.٩٦٢ | ٥٩.٨١ | ٠.١٣٤ | ٠.٦٧ | ٣٦ | ٢.٥٠٧ | ٦٢.٥٤ | ١.٢٦٣ | ٦.٣٢ |
| ١٢ | ٢.٣٤٩ | ٦١.٧٥ | ٠.٩٧٠ | ٤.٨٥ | ٣٧ | ١.٢٢٢ | ٥٦.١١ | ٠.٤٠٦ | ٢.٠٣ |
| ١٣ | ١.١٣٤ | ٥٥.٦٧ | ٠.٣٢٥ | ١.٦٣ | ٣٨ | ١.٣٤٥ | ٥٦.٧٣ | ٠.٣٧٠ | ١.٨٥ |
| ١٤ | ٣.٣٨٢ | ٦٦.٩١ | ٠.٤٤١ | ٢.٢١ | ٣٩ | ١.٢٠٥ | ٥٦.٠٣ | ٠.٣٢١ | ١.٦١ |
| ١٥ | ١.٦٧٢ | ٥٨.٣٦ | ٠.٤٣٨ | ٢.١٩ | ٤٠ | ١.٧٤٤ | ٥٨.٧٢ | ٠.٤٧٧ | ٢.٣٩ |
| ١٦ | ٠.٨٩٩ | ٥٤.٥٠ | ٠.٢٦٨ | ١.٣٤ | ٤١ | ٢.٣٣٢ | ٦١.٦٦ | ٠.٩٣٨ | ٤.٦٩ |
| ١٧ | ١.٩٦٢ | ٥٩.٨١ | ٠.٥٣٤ | ٢.٦٧ | ٤٢ | ٠.٧٤٣ | ٥٣.٧٢ | ٠.٢٨٥ | ١.٤٣ |
| ١٨ | ١.٠٦١ | ٥٥.٣١ | ٠.٢٩٠ | ١.٤٥ | ٤٣ | ٢.٢٣٣ | ٦١.١٧ | ٠.٨١١ | ٤.٠٦ |
| ١٩ | ٠.٩٦٣ | ٥٤.٨٢ | ٠.٠٦٤ | ٠.٣٢ | ٤٤ | ٢.٢١١ | ٦١.٠٦ | ٠.٧٩٣ | ٣.٩٧ |
| ٢٠ | ٠.٧٦٢ | ٥٣.٨١ | ٠.٢٥٧ | ١.٢٩ | ٤٥ | ١.٩٥٢ | ٥٩.٧٦ | ٠.٥٩٢ | ٢.٩٦ |
| ٢١ | ١.٩٦٣ | ٥٩.٨٢ | ٠.٤١٨ | ٢.٠٩ | ٤٦ | ١.٩٣٣ | ٥٩.٦٧ | ٠.٥٨٠ | ٢.٩٠ |
| ٢٢ | ١.٢٤٢ | ٥٦.٢١ | ٠.٥٦٨ | ٢.٨٤ | ٤٧ | ١.٦٥٦ | ٥٨.٢٨ | ٠.٤٤٢ | ٢.٢١ |
| ٢٣ | ٢.٣٩٠ | ٦١.٩٥ | ٠.١٨٠ | ٠.٩٠ | ٤٨ | ١.٩٠٧ | ٥٩.٥٤ | ٠.٥٦١ | ٢.٨١ |
| ٢٤ | ١.٠٨٩ | ٥٥.٤٥ | ٠.٣٤٤ | ١.٧٢ | ٤٩ | ٣.١٧٤ | ٦٥.٨٧ | ٠.٢٨١ | ١.٤١ |
| ٢٥ | ٢.٤٩٠ | ٦٢.٤٥ | ١.٢٢٣ | ٦.١٢ | ٥٠ | ٢.٥٠٧ | ٦٢.٥٤ | ١.٢٦٣ | ٦.٣٢ |

يتضح من جدول (٦) السابق تراوح قيم معاملات صعوبة الفقرات باللوجيت، وفقًا لإجابات الطلاب على نموذج الاختبار المخالف لقواعد صياغة الفقرات ما بين (٠.٤٤٧- ٣.٣٨٢) بمتوسط حسابي (١.٩٠٦) وانحراف معياري (٠.٦٧١)، وتراوحت قيم الخطأ المعياري في تقدير معلمة الصعوبة للنموذج المخالف ما بين (٠.٠٦٤- ١.٣٦٣) بمتوسط حسابي (٠.٥٠٥)، وقد حصلت الفقرة (١٤) على أعلى معامل صعوبة، بينما حصلت الفقرة (٨) على أدنى معامل صعوبة.

وللكشف عن الفروق في دقة تقدير معامل صعوبة الفقرات تبعاً لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، تم إجراء الاختبار الإحصائي ( T–test) لاختبار دلالة الفروق بين متوسطي صعوبة فقرات الاختبار المحكم والمخالف وكذلك متوسطي الأخطاء المعيارية للفقرات (بوحدة المنف)، كما هو موضح في الجدولين (٧) ، (٨) التاليين:

**جدول(٧)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي معلمة صعوبة الفقرات لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٥٦.٠٨٩ | ٢.٢٥٦ | ٧.٨٠٨– | ٠.٠٠٠ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٥٩.٥٣٢ | ٣.٣٥٥ | | |

يتضح من جدول (٢١) السابق أن قيمة (ت=–٧.٨٠٨) دالة إحصائية عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على أن متوسط معلمة الصعوبة لفقرات الاختبار المخالف كان أعلى من متوسط معلمة الصعوبة لفقرات الاختبار المحكم، وهذه النتيجة تتفق مع نتائج تحليل معامل الصعوبة عن طريق القياس الكلاسيكي والتي أكدت على أن فقرات الاختبار المحكم أسهل من فقرات الاختبار المخالف، وإن إدخال المخالفات على الفقرات جعل الفقرات أكثر صعوبة، وقد أكدت هذه النتيجة ما توصلت إليه دراسات (محمد صيتان الصمادي، ٢٠١٥؛ Huang et al., 2007; Pachai et al., 2015) ، من أن الفقرات التي تضمنت البديل "لا شيء مما ذكر" كبديل صحيح كانت الأكثر صعوبة وذات فروق إحصائية.

**جدول(٨)**

**اختبار ( ت ) لدراسة دلالة فروق متوسطي الأخطاء المعيارية في دقة تقدير معلمة الصعوبة لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ١.٦١٥ | ١.٢٠٤ | ٣.٦٢٥– | ٠.٠٠١ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٢.٥٢٦ | ١.٥٥٢ | | |

يتضح من جدول (٢٢) السابق أن قيمة (ت=-٣.٦٢٥) وهي دالة إحصائية عند مستوى دلالة (٠.٠١) مما يؤكد وجود فروق ذات دلالة إحصائية بين متوسطي الأخطاء المعيارية في دقة تقديرات معالم صعوبة نموذجي الاختبار المحكم والمخالف، وكانت أقل قيمة لصالح الاختبار المحكم البناء؛ أي أن فقرات نموذج الاختبار المحكم البناء أكثر دقة في تقدير صعوبة الفقرات، وجاءت هذه النتيجة متوافقة مع نتائج دراسة كل من (ابتسام عيسى خصاونة، ٢٠١٢؛ حابس سعد الزبون، راجي عوض الصرايرة، ٢٠١٧؛ فريال محمد أبو عواد، ٢٠١٨؛ نضال الشريفين، رانيا الصبح، ٢٠١١)، والتي أظهرت نتائج التحليل الإحصائي لتلك الدراسات أن الاختبار المحكم قدم تقديرات أكثر دقة لمعالم صعوبة الفقرات، وأن متوسط الأخطاء المعيارية لمعلمة الصعوبة لفقرات الاختبار المخالف كان أعلى من متوسط الأخطاء المعيارية لمعلمة الصعوبة لفقرات الاختبار المحكم.

**ثالثاً:** هل توجد فروق في معاملي الارتباط بين تقديرات معاملات الصعوبة المقدرة باستخدام القياس الكلاسيكي وتلك المقدرة باستخدام النموذج اللوجستي الثلاثي البارامتر في حالة نموذج الاختبار ( المحكم ، المخالف ) لقواعد صياغة الفقرات؟ "

للإجابة عن هذا التساؤل الفرعي تم تحويل معاملات الصعوبة المحسوبة في ضوء القياس الكلاسيكي إلى قيم (Z) المعيارية لتصبح قيماً مناسبةً للقياس الفتري وصالحة لاستخدام معامل ارتباط بيرسون لحساب العلاقات الارتباطية وبعدها تم إجراء الخطوات التالية:

١- حساب معامل ارتباط ($R_1$) بين معاملات صعوبة الفقرات المقدرة في ضوء القياس الكلاسيكي لنموذجي الاختبار (المحكم ، المخالف) بعد تحويليها إلى قيم معيارية وكانت قيمة معامل الارتباط (٠.٢١٨)، وهو معامل ارتباط ضعيف ولم تكن قيمة الارتباط دالة مما يؤكد اختلاف معامل الصعوبة المحسوب كلاسيكياً لنموذجي الاختبار (المحكم ، والمخالف) لقواعد صياغة الفقرات.

٢- حساب معامل ارتباط ($R_2$) بين معاملات صعوبة الفقرات المقدرة في ضوء النموذج اللوجستي الثلاثي البارامتر لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، وكانت قيمة معامل الارتباط (٠.٤٣٧)، وهو معامل ارتباط دال إحصائياً عند مستوى (٠.٠١)، وهذا يؤكد ارتباط معاملات الصعوبة المقدرة في ضوء النموذج اللوجستي الثلاثي البارامتر لنموذجي الاختبار (المحكم ، المخالف) إلى حد أعلى من المتوسط، وأقوى بكثير من ارتباط معاملات صعوبة الفقرات المقدرة في ضوء القياس الكلاسيكي.

٣- تم استخدام معادلة Steiger, (1980) للمقارنة بين معامل الارتباط ($R_1$) وقيمته (٠.٢١٨) ومعامل الارتباط ($R_2$) وقيمته (٠.٤٣٧) وكانت قيمة (Z) تساوي (٦.٧٦٠) وهي قيمة دالة إحصائياً عند مستوى (٠.٠١) وهذا يدل على وجود فرق دال إحصائياً بين معاملي الارتباط، كما يدل على أن ارتباط معاملات الصعوبة المقدرة وفق القياس الكلاسيكي لنموذجي الاختبار (المحكم ، المخالف) تختلف جوهرياً عن ارتباط معاملات صعوبة الفقرات المقدرة في ضوء النموذج اللوجستي الثلاثي المعلم لنموذجي الاختبار (المحكم ، المخالف)، وتتفق هذه النتيجة مع ما توصل إليه دراسات (Bechger, Maris, Verstralen & Beguin, 2003; Eleje, et al., 2018; Ojerinde, 2013; Stage, 2003) من اختلاف تقدير صعوبة الفقرات لكل من النظرية الكلاسيكية في القياس ونظرية الاستجابة للفقرة، وأن تحليل الفقرة في ضوء نظرية الاستجابة للفقرة كان أفضل من النظرية الكلاسيكية في القياس.

## [ ٣ ] – نتائج التساؤل الثالث وتفسيرها:

والذي ينص على أنه " ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تمييز الفقرات في ضوء القياس الكلاسيكي والنموذج اللوجستي الثلاثي البارامتر؟"، وللإجابة عن هذا التساؤل تمت الاجابة عن التساؤلات الفرعية التالية:

**أولاً:** ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تمييز الفقرات في ضوء القياس الكلاسيكي ؟"

للإجابة عن التساؤل السابق الفرعي تم حساب معامل التمييز كلاسيكياً (اعتماداً على طريقة المقارنة الطرفية)، ولقد تراوحت معاملات تمييز الفقرات وفقًا لإجابات الطلاب على نموذج الاختبار محكم البناء ما بين (٠.٣٢٢- ٠.٦٣٠) وبمتوسط حسابي (٠.٤٤٧) وانحراف معياري (٠.٠٨٠)، قد حصلت الفقرة (٣) على أعلى معامل تمييز، بينما حصلت الفقرة (٢٣) على أدنى معامل تمييز، بينما تراوحت قيم معاملات تمييز الفقرات وفقًا لإجابات الطلاب على نموذج الاختبار المخالف ما بين (٠.٢٦٥- ٠.٥٦٦) بمتوسط حسابي (٠.٣٤٩) وانحراف معياري (٠.٠٨١)، وقد حصلت الفقرة (٢٠) على أعلى معامل تمييز، بينما حصلت الفقرة (٢٩) على أدنى معامل تمييز.

وللكشف عن الفروق في معامل تمييز الفقرات تبعاً لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، تم إجراء الاختبار الإحصائي (T-test) لاختبار دلالة الفروق بين متوسطي تمييز فقرات الاختبار المحكم والمخالف لقواعد صياغة الفقرات ، كما هو موضح في الجدول (٩) التالي:

**جدول(٩)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي تمييز فقرات**

الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٠.٤٤٧ | ٠.٠٨٠ | ٦.٨٦٨ | ٠.٠٠٠ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٠.٣٤٩ | ٠.٠٨١ | | |

يتضح من جدول (٩) السابق أن قيمة (ت=٦.٨٦٨) وهي دالة إحصائياً عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على وجود فروق بين متوسطي معامل التمييز لنموذجي الاختبار (المحكم ، المخالف) لصالح الاختبار محكم البناء، وتتشابه هذه النتيجة مع ما توصلت إليه دراسة كل من (ابتسام عيسى خصاونة، ٢٠١٢؛ إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٢؛ Huang et al., 2007; Pachai et al., 2015) والتي أظهرت نتائجها أن استخدام البديل "لا شيء مما ذكر" يؤدي إلى انخفاض في معاملات تمييز الفقرات وذلك عندما يمثل هذا البديل الإجابة الصحيحة.

**ثانياً:** ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تمييز الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير الخطأ المعياري لمتوسط معاملات تمييز الفقرات تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات ؟ "

للإجابة عن هذا التساؤل تم استخدام برنامج XCalibre 4.1.7 لتحليل بيانات كل من نموذجي الاختبار لتقدير معالم تمييز الفقرات، ولقد تراوحت قيم معاملات تمييز الفقرات باللوجيت وفقًا لإجابات الطلاب على نموذج الاختبار المحكم ما بين (٠.٥٨٤- ١.٩٦٠) وبمتوسط حسابي (١.٤٤٨) وانحراف معياري (٠.٣١٦)، وتراوحت قيم الخطأ المعياري في تقدير معلمة التمييز للنموذج المحكم ما بين (٠.٠٩١- ٠.٦٨٥) بمتوسط حسابي(٠.٣٢٤)، وقد حصلت الفقرة (٥) على أعلى معامل تمييز، بينما حصلت الفقرة (١٠) على أدنى معامل تمييز.

كما تراوحت قيم معاملات تمييز الفقرات باللوجيت وفقًا لإجابات الطلاب على نموذج الاختبار المخالف لقواعد صياغة الفقرات ما بين (٠.٧٥١- ١.٤٦٣) وبمتوسط حسابي ( ١.١١٨) وانحراف معياري (٠.١٥٧)، وتراوحت قيم الخطأ المعياري في تقدير معلمة التمييز للنموذج المخالف ما بين (٠.١٠٢-١.٠٦٦) بمتوسط حسابي (٠.٤٩٦)، وقد حصلت الفقرة (٩) على أعلى معامل تمييز، بينما حصلت الفقرة (٢٧) على أدنى معامل تمييز.

وللكشف عن الفروق في دقة تقدير معامل تمييز الفقرات تبعاً لنموذجي الاختبار (المحكم ، المخالف) تم إجراء الاختبار الإحصائي (T-test) لاختبار دلالة الفروق بين متوسطي تمييز فقرات الاختبار المحكم والمخالف وكذلك متوسطي الأخطاء المعيارية لفقرات الاختبار المحكم والمخالف، كما هو موضح في الجدولين (١٠) ، (١١) التالي:

**جدول(١٠)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي معلمة تمييز الفقرات لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ١.٤٤٨ | ٠.٣١٦ | ٨.٣٠٠ | ٠.٠٠٠ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ١.١١٨ | ٠.١٥٧ | | |

يتضح من جدول (١٠) السابق أن قيمة (ت=٨.٣٠٠) دالة إحصائية عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على أن متوسط معلمة التمييز لفقرات الاختبار المحكم كان أعلى من متوسط معلمة التمييز لفقرات الاختبار المخالف لقواعد الصياغة، أي أن وجود المخالفات في قواعد الصياغة يؤثر على التقديرات الخاصة بمعلمة التمييز، وتتفق هذه النتيجة مع ما توصلت إليه دراسات (محمد صيتان الصمادي، ٢٠١٥؛ Huang et al., 2007; Pachai et al., 2015) والتي أظهرت نتائجها أن استخدام البديل "لا شيء مما ذكر" يؤدي إلى انخفاض في معاملات تمييز الفقرات وذلك عندما يمثل هذا البديل الإجابة الصحيحة.

**جدول(١١)**

**اختبار ( ت ) لدراسة دلالة فروق متوسطي الأخطاء المعيارية في دقة تقدير معلمة التمييز لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٠.٣٢٤ | ٠.١٧٥ | -٣.٩٩٢ | ٠.٠٠٠ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٠.٤٩٦ | ٠.٢٦٨ | | |

يتضح من جدول (٢٨) السابق أن قيمة (ت=-٣.٩٩٢) وهي دالة إحصائية عند مستوى (α=٠.٠١) مما يؤكد على وجود فروق ذات دلالة إحصائية بين متوسطي الأخطاء المعيارية في دقة تقديرات معالم تمييز نموذجي الاختبار المحكم والمخالف، وذلك لأقل قيمة لصالح الاختبار المحكم البناء؛ أي أن فقرات نموذج الاختبار المحكم البناء أكثر دقة في تقدير تمييز الفقرات، وجاءت هذه النتيجة مختلفة مع نتائج دراسة (الشريفين، الصبح، ٢٠١١) والتي أظهرت نتائج التحليل الإحصائي لها عدم وجود فروق ذات دلالة إحصائياً بين متوسطات الأخطاء المعيارية في تقديرات معالم التمييز للفقرات تبعاً لنموذج الاختبار (المحكم، المخالف) لقواعد صياغة الفقرات.

**ثالثاً:** هل توجد فروق في معاملي الارتباط بين تقديرات معاملات التمييز المقدرة باستخدام القياس الكلاسيكي وتلك المقدرة باستخدام النموذج اللوجستي الثلاثي البارامتر في حالة نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟ "

للإجابة عن هذا التساؤل تم تحويل معاملات التمييز المحسوبة في ضوء القياس الكلاسيكي إلى قيم (Z) المعيارية لتصبح قيماً مناسبةً للقياس الفتري وصالحة لاستخدام معامل ارتباط بيرسون لحساب العلاقات الارتباطية وبعدها تم إجراء الخطوات التالية:

١- حساب معامل ارتباط ($R_1$) بين معاملات تمييز الفقرات المقدرة في ضوء القياس الكلاسيكي لنموذجي الاختبار (المحكم ، المخالف) بعد تحويليها إلى قيم معيارية وكانت قيمة معامل الارتباط (٠.٢٣٠)، وهو معامل ارتباط ضعيف ولم تكن قيمة الارتباط دالة مما يؤكد اختلاف معامل التمييز المحسوب كلاسيكياً لنموذجي الاختبار (المحكم ، والمخالف) لقواعد صياغة الفقرات.

٢- حساب معامل ارتباط ($R_2$) بين معاملات تمييز الفقرات المقدرة في ضوء النموذج اللوجستي ثلاثي البارامتر لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، وكانت قيمة معامل الارتباط (٠.٤٦٠)، وهو معامل ارتباط دال إحصائياً عند مستوى (٠.٠١)، وهذا يدل على ارتباط معاملات التمييز المقدرة في ضوء النموذج اللوجستي الثلاثي البارامتر لنموذجي الاختبار (المحكم ، المخالف)، وأقوى بكثير من ارتباط معاملات تمييز الفقرات المقدرة في ضوء القياس الكلاسيكي.

٣- استخدمت معادلة (Steiger, 1980) للمقارنة بين معامل الارتباط ($R_1$) وقيمته (٠.٢٣٠) ومعامل الارتباط ($R_2$) وقيمته (٠.٤٦٠) وكانت قيمة (Z) تساوي (٧.٤٠٦)، وهي قيمة دالة إحصائياً عند مستوى(α=٠.٠١) وهذا يدل على وجود فرق دال إحصائياً بين معاملي الارتباط، كما يدل على أن ارتباط معاملات التمييز المقدرة وفق القياس الكلاسيكي لنموذجي الاختبار (المحكم ، المخالف) تختلف جوهرياً عن ارتباط معاملات التمييز المقدرة في ضوء النموذج اللوجستي الثلاثي البارامتر لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات.

## [ ٤ ] – نتائج التساؤل الرابع وتفسيرها:

والذي ينص على أنه " ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على متوسط معاملات تخمين الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير الخطأ المعياري لمتوسط معاملات تخمين الفقرات تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟"

للإجابة عن التساؤل السابق تم استخدام برنامج XCalibre 4.1.7 لتحليل بيانات كل من نموذجي الاختبار لتقدير معالم تخمين الفقرات، ولقد تراوحت قيم معاملات تخمين الفقرات باللوجيت وفقًا لإجابات الطلاب على نموذج الاختبار المحكم ما بين (٠.١٢٩- ٠.٢٦٦) بمتوسط حسابي (٠.١٧٨) وانحراف معياري (٠.٠٢٣)، وتراوحت قيم الخطأ المعياري في تقدير معلمة التخمين للنموذج المحكم ما بين (٠.٠٣٣- ٠.٢٩١) بمتوسط حسابي (٠.٠٨٦)، وقد حصلت الفقرة (١٠) على أعلى معامل تخمين، بينما حصلت الفقرة (٣٥) على أدنى معامل تخمين.

كما تراوحت قيم معاملات تخمين الفقرات باللوجيت وفقًا لإجابات الطلاب على نموذج الاختبار المخالف لقواعد الصياغة ما بين (٠.١٧٢- ٠.٣٢٣) بمتوسط حسابي (٠.٢١٠) وانحراف معياري (٠.٠٣٧)، وتراوحت قيم الخطأ المعياري في تقدير معلمة التخمين للنموذج المخالف ما بين (٠.٠٤٤-٠.٣٤٦) بمتوسط حسابي (٠.١٩٤)، وقد حصلت الفقرة (١٧) على أعلى معامل تخمين، بينما حصلت الفقرة (٤١) على أدنى معامل تخمين.

وللكشف عن الفروق في دقة تقدير معامل تخمين الفقرات تبعاً لنموذجي الاختبار (المحكم ، المخالف) تم إجراء الاختبار الإحصائي (T-test) لاختبار دلالة الفروق بين متوسطي تخمين فقرات الاختبار المحكم والمخالف لقواعد الصياغة وكذلك متوسطي الأخطاء المعيارية لفقرات الاختبار المحكم والمخالف، كما هو موضح في الجدول (١٢) ، (١٣) التالي:

**جدول(١٢)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي معلمة تخمين الفقرات لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٠.١٧٨ | ٠,٠٢٣ | ٦.٦٦٧- | ٠.٠٠٠ دالة عند مستوى (٠.٠١=α) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٠.٢١٠ | ٠.٠٣٧ | | |

يتضح من جدول (١٢) السابق أن قيمة (ت=-٦.٦٦٧) دالة إحصائية عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على أن متوسط معلمة التخمين لفقرات الاختبار المخالف كان أعلى من متوسط معلمة التخمين لفقرات الاختبار المحكم، أي أن وجود المخالفات في قواعد الصياغة يؤثر على التقديرات الخاصة بمعلمة التخمين، وجاءت هذه النتيجة متفقة مع دراسات (إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٠؛ طه الخرشه، ٢٠١٦؛ نضال الشريفين، رانيا الصبح،٢٠١١)، والتي جاءت نتائجها الخاصة بتقديرات معلمة التخمين لتؤكد أنها كانت أعلى في نموذج الاختبار المخالف لقواعد الصياغة.

جدول(١٣)

اختبار ( ت ) لدراسة دلالة فروق متوسطي الأخطاء المعيارية في دقة تقدير معلمة التخمين لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ٥٠ | ٠.٠٨٦ | ٠.٠٦٧ | -٦.٢٣٩ | ٠.٠٠٠ دالة عند مستوى (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ٥٠ | ٠.١٩٤ | ٠.١٠٢ | | |

يتضح من جدول (١٣) السابق أن قيمة (ت=-٦.٢٣٩) دالة إحصائية عند مستوى (α=٠.٠١) مما يؤكد على وجود فروق ذات دلالة إحصائية بين متوسطي الخطأ المعياري في دقة تقديرات معلمة التخمين تبعاً لنموذجي الاختبار المحكم والمخالف، وكانت أقل قيمة لصالح الاختبار المحكم البناء؛ أي أن فقرات نموذج الاختبار المحكم البناء أكثر دقة في تقدير معلمة التخمين، وجاءت هذه النتيجة متفقة مع دراسة (إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٠؛ طه الخرشه، ٢٠١٦؛ فريال محمد أبو عواد، ٢٠١٨؛ نضال الشريفين، رانيا الصبح، ٢٠١١)، والتي جاءت نتائجها الخاصة بالخطأ المعياري المرتبط بدقة تقديرات معلمة التخمين لتؤكد أنها كانت أكثر دقة لصالح الاختبار المحكم البناء.

## [ ٥ ] – نتائج التساؤل الخامس وتفسيرها:

والذي ينص على أنه " ما أثر انتهاك بعض قواعد صياغة فقرات اختبار الاختيار من متعدد على دقة تقديرات معالم القدرة للأفراد في ضوء النموذج اللوجستي الثلاثي البارامتر؟ وهل هناك فروق ذات دلالة إحصائية في تقدير متوسط الخطأ المعياري لتقدير قدرات الأفراد تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات؟ "

للإجابة عن التساؤل السابق تم تقدير قيم معالم القدرة للأفراد لنموذجي الاختبار باستخدام برنامج XCalibre 4.1.7، والذي يعمل على تقدير القدرة باستخدام طرق منها طريقة الأرجحية العظمى(Maximum Likelihood Estimation (MLE ، وللكشف عن الفروق في دقة تقدير معالم قدرة الأفراد والخطأ المعياري في تقديرها في ضوء النموذج اللوجستي الثلاثي البارامتر تبعاً لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، تم إجراء الاختبار الإحصائي (T–test) لاختبار دلالة الفروق بين متوسطي قدرة الأفراد وكذلك بين متوسطي الأخطاء المعيارية لنموذجي الاختبار المحكم والمخالف (بوحدة المنف)، كما هو موضح في الجدول (١٤) ، (١٥) التاليين:

**جدول(١٤)**

**اختبار ( ت ) لدراسة دلالة الفروق بين متوسطي قدرة الأفراد لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ١٥٠٠ | ٥١.٣٤٥ | ٤.٧٩٥ | ٥.٩١٤ | ٠.٠٠٠ دالة عند مستوي (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ١٥٠٠ | ٥٠.٥٠٧ | ٤.٨٥٦ | | |

يتضح من جدول (١٤) السابق أن قيمة (ت=٥.٩١٤) دالة إحصائية عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على وجود فروق ذات دلالة إحصائية بين متوسطي قدرة الأفراد تبعاً لنموذجي الاختبار المحكم والمخالف لقواعد صياغة الفقرات، حيث يلاحظ أنه فيما يخص متوسط القدرة أن متوسط الاختبار المحكم كان أعلى من متوسط الاختبار المخالف مما يؤكد أن انتهاكات قواعد صياغة فقرات الاختيار من متعدد أثرت على قدرة الأفراد، وقد جاءت هذه النتيجة متعارضة مع مسلمة نظرية الاستجابة للمفردة والمتعلقة باللاتغير في معالم القدرة باختلاف معالم الفقرات ولكن قد يبدو الأمر ليس على إطلاقه وخاصة مع استخدام النموذج الثلاثي البارامتر، وتتفق هذه النتيجة مع ما جاءت به نتائج دراسة (الرشيدي ،٢٠١٠؛ الشريفين، بني عطا،٢٠١٣) والتي أكدت عدم تحقق اللاتغير في قدرات الأفراد عند تقدمهم لفقرات مختلفة الصعوبة حيث كانت الفروق بين متوسطات معالم القدرة دالة ومن ثم لم يتحقق افتراض اللاتغير في تقدير معالم الفقرات باختلاف معالم القدرة،

**جدول(١٥)**

**اختبار ( ت ) لدراسة دلالة فروق متوسطي الأخطاء المعيارية لتقدير قدرة الأفراد لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ١٥٠٠ | ٣.٢٢٨ | ٣.٢٥٨ | -٦.٠٢٩ | ٠.٠٠٠ دالة عند مستوي (α=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ١٥٠٠ | ٣.٨٦٧ | ٢.٩٦٠ | | |

يتضح من جدول (١٥) السابق أن قيمة (ت=-٦.٠٢٩) دالة إحصائية عند مستوى (α=٠.٠١) وتؤكد هذه النتيجة على وجود فروق ذات دلالة إحصائية بين متوسطي الأخطاء المعيارية في تقدير القدرة تبعاً لنموذجي الاختبار المحكم والمخالف لقواعد الصياغة، حيث يتضح أنه فيما يخص متوسط الأخطاء المعيارية فكان متوسط الخطأ المعياري للاختبار المحكم أقل من متوسط الخطأ المعياري للاختبار المخالف، أي أن فقرات نموذج الاختبار المحكم كانت أكثر دقة في تقدير قدرة الأفراد من الاختبار المخالف، وتعتبر هذه النتيجة من النتائج المنطقية؛ حيث أن الابتعاد عما صممت فقرة الاختبار لقياسه يؤدي ذلك إلى تشتت التفكير، وبالتالي تبتعد الفقرة عن قياس القدرة الحقيقة للأفراد، وهذا يؤدي إلى زيادة الأخطاء المعيارية في تقدير معلمة القدرة، وقد جاءت هذه النتيجة لتؤكد ما توصلت إليه دراسات (إبراهيم محمد يعقوب، باسل خميس أبو فودة، ٢٠١٠؛ فريال محمد أبو عواد، ٢٠١٨؛ نضال الشريفين، رانيا الصبح، ٢٠١١)، من أن نموذج الاختبار المحكم البناء كان الأكثر دقة في تقدير قدرات الأفراد؛ حيث كانت متوسطات الأخطاء المعيارية لمعالم قدرات الأفراد الأقل لصالح نموذج الاختبار المحكم البناء، وهذا يشير إلى دقة القياس.

**[ ٦ ] – نتائج التساؤل السادس وتفسيرها:**

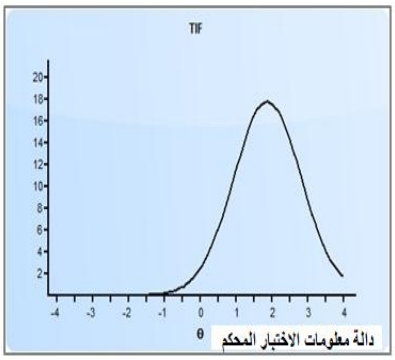
والذي ينص على أنه " هل هناك فروق ذات دلالة إحصائية بين التقديرات الخاصة بدالة معلومات الاختبار تعزى إلى نموذج الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات في ضوء النموذج اللوجستي الثلاثي البارامتر؟ "

للإجابة عن التساؤل السابق تم استخدام برنامج XCalibre 4.1.7 للحصول على دالة المعلومات لكل فقرة من فقرات الاختبار بنموذجيه (المحكم ، المخالف) لقواعد صياغة الفقرات، وبين جدول (١٦) التالي القيم العظمى لدالة المعلومات لكل فقرة من فقرات نموذجي الاختبار.

**جدول(١٦) : القيم العظمى لدالة المعلومات لكل فقرة من فقرات نموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات**

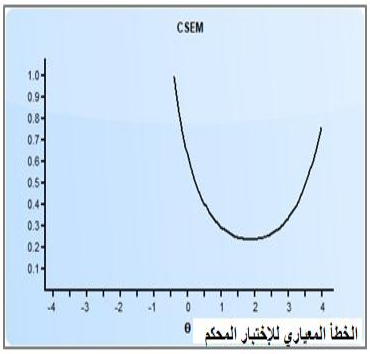
| فقرات الاختبار المحكم | | | فقرات الاختبار المخالف | | |
|---|---|---|---|---|---|
| رقم الفقرة | دالة المعلومات | رقم الفقرة | دالة المعلومات | رقم الفقرة | دالة المعلومات |
| ١ | ٠.٧٨٦ | ٢٦ | ٠.٥٣١ | ١ | ٠.٥٧١ | ٢٦ | ٠.٤٧٣ |
| ٢ | ٠.٣٠١ | ٢٧ | ٠.٤٨٥ | ٢ | ٠.١٢٤ | ٢٧ | ٠.٣٩٣ |
| ٣ | ٠.٦٥٨ | ٢٨ | ١.٤١٧ | ٣ | ٠.٣٤١ | ٢٨ | ٠.٥٩٢ |
| ٤ | ١.٥٦٤ | ٢٩ | ٠.٥٢٦ | ٤ | ٠.٢٨٥ | ٢٩ | ٠.٤٦١ |
| ٥ | ٠.٥٠٨ | ٣٠ | ١.٣٤٩ | ٥ | ٠.١٢٦ | ٣٠ | ٠.٦٣٢ |
| ٦ | ٠.٧٢٦ | ٣١ | ٠.٦٢٦ | ٦ | ٠.٤٢٧ | ٣١ | ٠.٥٨٧ |
| ٧ | ٠.٧٧٥ | ٣٢ | ١.٠٢٦ | ٧ | ٠.٥٠٠ | ٣٢ | ٠.٥٥٤ |
| ٨ | ٠.٥٠٥ | ٣٣ | ٠.٤٦٥ | ٨ | ٠.٣٦٩ | ٣٣ | ٠.٢٩٦ |
| ٩ | ٠.٤٦٢ | ٣٤ | ٠.٤٤٥ | ٩ | ٠.٣٧٨ | ٣٤ | ٠.٢٦١ |
| ١٠ | ٠.٦٠٩ | ٣٥ | ٠.٥١٨ | ١٠ | ٠.٢٠٠ | ٣٥ | ٠.٣٤٥ |
| ١١ | ٠.٧٤٠ | ٣٦ | ١.٢٩٦ | ١١ | ٠.٤٤٥ | ٣٦ | ١.٢٩٤ |
| ١٢ | ٠.٩٧٢ | ٣٧ | ١.٠٠٧ | ١٢ | ٠.٢٢٢ | ٣٧ | ٠.٧٥٠ |
| ١٣ | ٠.٧٥٨ | ٣٨ | ٠.٨٣١ | ١٣ | ٠.٧٠٩ | ٣٨ | ٠.٤٤٤ |
| ١٤ | ٠.٥٨٢ | ٣٩ | ٠.٤٩٢ | ١٤ | ٠.٣٦٧ | ٣٩ | ٠.٤١٦ |
| ١٥ | ٠.٩٤١ | ٤٠ | ٠.٦٨١ | ١٥ | ٠.٨٢٢ | ٤٠ | ٠.٦٦٤ |
| ١٦ | ٠.٧٧٦ | ٤١ | ٠.٩٥٧ | ١٦ | ٠.٣٩٢ | ٤١ | ٠.٦٥٦ |
| ١٧ | ٠.٥٢٦ | ٤٢ | ٠.٤٤٤ | ١٧ | ٠.٤٨٠ | ٤٢ | ٠.٢٢٨ |
| ١٨ | ٠.٧٦٥ | ٤٣ | ٠.٨٨٥ | ١٨ | ٠.٦١٠ | ٤٣ | ٠.٣٩١ |
| ١٩ | ٠.٩٤٢ | ٤٤ | ٠.٧٠٨ | ١٩ | ٠.٦٩٥ | ٤٤ | ٠.٣٧٨ |
| ٢٠ | ٠.٥٥٠ | ٤٥ | ١.٢٤٠ | ٢٠ | ٠.٣٩٨ | ٤٥ | ٠.٣٠٤ |
| ٢١ | ١.١٢٣ | ٤٦ | ٠.٨٩٧ | ٢١ | ٠.٤٢١ | ٤٦ | ٠.٤٠٠ |
| ٢٢ | ٠.٩٧٠ | ٤٧ | ١.١٠٢ | ٢٢ | ٠.٣٠٩ | ٤٧ | ٠.٣١٣ |
| ٢٣ | ٠.٩٠٣ | ٤٨ | ١.٠٢٥ | ٢٣ | ٠.٣٠٣ | ٤٨ | ٠.٧٠٦ |
| ٢٤ | ٠.٦٨٩ | ٤٩ | ١.١٠٢ | ٢٤ | ٠.٢٩٧ | ٤٩ | ٠.٨٠٣ |
| ٢٥ | ٠.٩٣٧ | ٥٠ | ٠.٩٠٦ | ٢٥ | ٠.٨٩٢ | ٥٠ | ٠.٤٩٩ |

يتضح من جدول (١٦) السابق أن القيم العظمى لجميع دوال معلومات فقرات الاختبار محكم البناء كانت أعلى من قيم دوال المعلومات التي تقدمها فقرات الاختبار المخالف لقواعد الصياغة، كما تم رسم منحنيات دالة معلومات الاختبار المحكم والمخالف والخطأ المعياري في تقدير فقرات نموذجي الاختبار المحكم والمخالف، والتي توضح كمية المعلومات التي يقدمها الاختبار والخطأ المعياري عند مستويات القدرة المختلفة كما في الشكلين (٣ ، ٤) للاختبار المحكم، والشكلين (٥ ، ٦) للاختبار المخالف:




**شكل ( ٣ )**

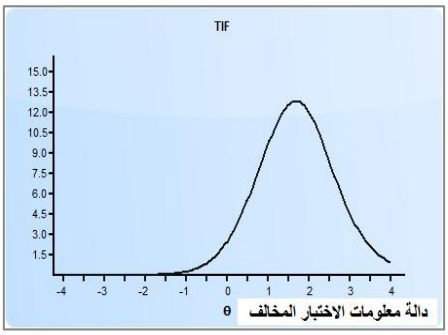
**منحنى دالة معلومات الاختبار المحكم البناء**




**شكل ( ٤ )**

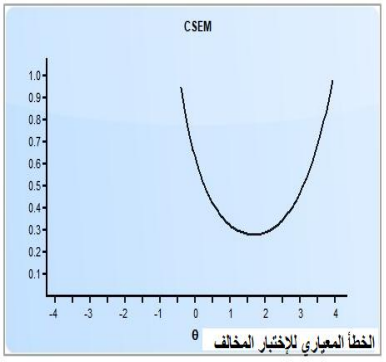
**الخطأ المعياري للتقدير للاختبار المحكم البناء**

يتضح من الشكل (٣) السابق والذي يمثل منحنى دالة معلومات الاختبار المحكم (TIF)، والذي يوضح كمية معلومات الاختبار المحكم التي يقدمها الاختبار عند مستويات القدرة المختلفة، وكان أقصى قدر من المعلومات التي يمكن تقديمها عن طريق الاختبار المحكم عند الدرجة (١٨) مقابل لمستوى قدرة ($\theta$) = (١.٨٥٠) ، كما يتضح من الشكل (٤) والذي يعرض الرسم البياني للخطأ المعياري، ويقوم بتقدير كمية الخطأ في مستوى القدرة ($\theta$) لكل مستوى من مستوياتها وهو معكوس (TIF) وكان أقل خطأ معياري للاختبار المحكم يساوي (٠.٢٣٧) عند مستوى قدرة (١.٨٥٠).




**شكل ( ٥ )**

**منحنى دالة معلومات الاختبار المخالف لقواعد صياغة الفقرات**




**شكل ( ٦ )**

**الخطأ المعياري للتقدير للاختبار المخالف لقواعد صياغة الفقرات**

يتضح من الشكل (٥) السابق والذي يمثل منحنى دالة معلومات الاختبار المخالف (TIF)، والذي يوضح كمية معلومات الاختبار المخالف التي يقدمها الاختبار عند مستويات القدرة المختلفة، وكان أقصى قدر من المعلومات التي يمكن تقديمها عن طريق الاختبار المخالف عند الدرجة (١٣) مقابل لمستوى قدرة ($\theta$)= (١.٧٠٠) ، كما يتضح من الشكل (٦) والذي يعرض الرسم البياني للخطأ المعياري، ويقوم بتقدير كمية الخطأ في مستوى القدرة ($\theta$) لكل مستوى من مستوياتها وهو معكوس (TIF) وكان أقل خطأ معياري للاختبار المخالف يساوي (٠.٢٧٩) عند مستوى قدرة (١.٧٠٠).

وللكشف عن الفروق في التقدير الخاص بدالة معلومات الاختبار في ضوء النموذج اللوجستي الثلاثي المعلم تبعاً لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، تم إجراء الاختبار الإحصائي(T–test) لاختبار دلالة الفروق بين متوسطي دالة معلومات الاختبار لنموذجي الاختبار المحكم والمخالف، كما هو موضح في الجدول (١٧) التالي:

**جدول(١٧)**

**اختبار (ت) لدراسة دلالة الفروق بين متوسطي دالة معلومات الاختبار لنموذجي الاختبار (المحكم ، المخالف) وفق النموذج الثلاثي البارامتر**

| م | العينة | العدد | المتوسط | الانحراف المعياري | ت | الدلالة |
|---|---|---|---|---|---|---|
| ١ | فقرات الاختبار المحكم | ١٦١ | ٦.٣٧٢ | ٧.١٣٠ | ٧.٢٧٠ | ٠.٠٠٠ دالة عند مستوي ($\alpha$=٠.٠١) |
| ٢ | فقرات الاختبار المخالف | ١٦١ | ٥.٠٢٤ | ٥.٢٤٥ | | |

يتضح من جدول (١٧) السابق أن (ت=٧.٢٧٠)، دالة إحصائية عند مستوى ($\alpha$=٠.٠١) وتؤكد هذه النتيجة على وجود فروق ذات دلالة إحصائية بين متوسطي دالة معلومات الاختبار تبعاً لنموذجي الاختبار المحكم والمخالف لقواعد الصياغة، كما يتضح أن المتوسط الحسابي لدالة معلومات الاختبار المحكم كان أعلى من متوسط دالة معلومات الاختبار المخالف، مما يؤكد أن انتهاكات قواعد صياغة فقرات الاختيار من متعدد أثرت على دالة معلومات الاختبار، وأن الاختبار المحكم البناء يقدم معلومات أكبر من الاختبار المخالف، وقد جاءت هذه النتيجة لتؤكد ما توصلت إليه دراسة (الشريفين، الصبح، ٢٠١١) من أن نموذج الاختبار المحكم يقدم معلومات أكبر من الاختبار المخالف.

ولقد بين (David, 2013; Jinming, 2012; Joo et al., 2018; Lord, 1980; Reise & Revicki, 2015) ، وجود عدة وسائل للكشف عن دقة تقدير المعالم وجودة الاختبارات منها محك الكفاءة النسبية للاختبار Relative Efficiency (RE) الذي يعتمد على دالة معلومات الاختبار Test Information Function (TIF) التي تلعب دوراً رئيساً في نظرية الاستجابة للفقرة؛ إذ يمكن من خلالها تحديد الخطأ المعياري في التقدير، فعندما يتم استخراج تقدير معلمة القدرة فإن تباين الخطأ في تقدير القدرة يساوي معكوس دالة المعلومات، وبالاعتماد على دالة المعلومات، فإنه يمكن تعريف الكفاءة النسبية على أنها نسبة دالة معلومات الاختبار المحكم (A) إلى دالة معلومات الاختبار المخالف (B) عند مستوى قدرة ($\theta$) كما في المعادلة التالية.

$$\mathbf{RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}}$$

وبالتالي فإن الاختبار المحكم (A) يكون أكثر كفاءة وفاعلية من الاختبار المخالف (B) عند مستوى القدرة ($\theta$) إذا كان $\mathbf{RE(\theta) > 1}$، أما إذا كان $\mathbf{RE(\theta) < 1}$ فإن الاختبار المحكم (A) يكون أقل كفاءة من الاختبار المخالف (B) وفي حالة أن $\mathbf{RE(\theta) = 1}$ يكون الاختباران A, B لهما نفس الكفاءة عند مستوى القدرة ($\theta$) ، وتشير الكفاءة هنا إلى الدقة في تقدير معلمة القدرة، وأما بالنسبة للكفاءة النسبية لنموذجي الاختبار (المحكم ، المخالف) لقواعد صياغة الفقرات، فقد تم حسابها لنموذج الاختبار المحكم إلى الاختبار المخالف، وذلك عن طريق قسمة قيم دالة معلومات الاختبار المحكم البناء إلى قيم دالة معلومات الاختبار المخالف، عند مستويات مختارة من القدرة؛ حيث أتضح أن الاختبار المحكم أكثر كفاءة وفاعلية من الاختبار المخالف لقواعد الصياغة عند مستويات القدرة ($\theta$) المختارة؛ حيث كانت قيمة الكفاءة النسبية عند جميع مستويات القدرة المنتقاة ($\theta$) أكبر من الواحد.

## [ ٧ ] – نتائج التساؤل السابع وتفسيرها:

والذي ينص على أنه " ما تقديرات قدرات أفراد العينة في اختبار الاختيار من متعدد المستخدم في الدراسة وذلك وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) وكذلك وفق النموذج اللوجستي الثلاثي البارامتر؟ ".

للإجابة عن التساؤل السابق تم تصحيح الاختبار وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية ، الاحتمال المقترح للإجابة الصحيحة)، وباستخدام برنامج SPSS(22) تم حساب الدرجة الكلية للطلاب وعدد من الإحصائيات الوصفية ذات العلاقة، وكذلك تم استخدام برنامج XCalibre 4.1.7 لتقدير قدرات الطلاب بوحدتي اللوجيت والمنف باستخدام طريقة الأرجحية العظمى(MLE) Maximum Likelihood Estimation، في ضوء النموذج اللوغاريتمي الثلاثي البارامتر.

وحيث أن معامل الاختلاف يستخدم للمقارنة بين التشتت النسبي لعدة قياسات في حالة اختلاف وحدات القياس (الشافعي، ٢٠١٤، ٧٧)، فقد تم إيجاد هذا المؤشر الإحصائي وفق القانون التالي:

معامل الاختلاف = ( الانحراف المعياري ÷ المتوسط الحسابي ) ×١٠٠.

ويعرض جدول (١٨) التالي عدد من الإحصائيات الوصفية لدرجات الطلاب المقدرة وفق طرق تقدير الدرجات الكلاسيكية الثلاث ( الطريقة التقليدية ، الطريقة التجريبية ، طريقة الاحتمال المقترح للإجابة الصحيحة)، وفي ضوء النموذج اللوغاريتمي الثلاثي البارامتر:

**جدول (١٨)**

**عدد من الإحصائيات الوصفية لدرجات الطلاب المقدرة وفق كل من طرق تقدير الدرجات الكلاسيكية والنموذج اللوجستي الثلاثي البارامتر**

| م | الطريقة | | المدى الخام (امتداد الدرجات) | أقل درجة | أعلى درجة | المدى الفعلي | المتوسط الحسابي | الانحراف المعياري | معامل الاختلاف |
|---|---|---|---|---|---|---|---|---|---|
| ١ | التقليدية | | ٠ _ ٥٠ | ١٢ | ٤٧ | ٣٥ | ٢١.٠٣ | ٧.٩٦ | ٪٣٧,٨٥ |
| ٢ | التجريبية | | ٠ _ ١٥٠ | ٢٤ | ١٤٦ | ١٢٢ | ٦٠,٦٥ | ٢٤.١ | ٪٣٩,٧٤ |
| ٣ | الاحتمال المقترح للإجابة الصحيحة | | ٠ _ ٥٠٠ | ٦٢ | ٤٨٠ | ٤١٨ | ١٩٨,٩ | ٨٣,٤٣ | ٪٤١,٩٥ |
| ٤ | النموذج اللوجستي الثلاثي البارامتر | لوجيت | −∞ إلى +∞ | – ٢,٧٧ | ٢,٦٩ | ٥.٤٦ | ٠.٢٦٩ | ٠.٩٥٩ | ٪٩.٣٤ |
| | | منف | −∞ إلى +∞ | ٣٦.١٥ | ٦٣.٤٥ | ٢٧,٣١ | ٥١.٣٤٥ | ٤.٧٩٥ | |

يتضح من جدول (١٨) السابق أنه عند استخدام الطريقة التقليدية امتدت درجات الطلاب من (١٢) درجات إلى (٤٧) درجة، بمتوسط حسابي قدرة (٢١.٠٣) درجة، وهو ما يعادل (٣٧,٨٥٪ من الدرجة الكلية)، وعند استخدام الطريقة التجريبية امتدت درجات الطلاب بين (٢٤) و (١٤٦) درجة بمتوسط حسابي قدره (٦٠.٦٥) درجة، وهو ما يعادل (٣٩,٧٤٪ من الدرجة الكلية)، وعند استخدام طريقة الاحتمال المقترح للإجابة الصحيحة امتدت درجات الطلاب بين (٦٢) و(٤٨٠) درجة، بمتوسط حسابي قدرة (١٩٨.٩) درجة، وهو يعادل (٤١,٩٥٪ من الدرجة الكلية)، وبمقارنة نسبة المتوسط الحسابي إلى الدرجة الكلية في الطرق الكلاسيكية الثلاث (٤٢.٠٦٪، ٤٠.٤٣٪، ٣٩.٧٨٪) يلاحظ أنها بوجه عام متقاربة إلا أن المتوسط الحسابي الأعلى كان عند استخدام الطريقة التقليدية، ثم الطريقة التجريبية، ثم طريقة الاحتمال المقترح للإجابة الصحيحة، أما عند استخدام النموذج اللوغاريتمي الثلاثي البارامتر فقد امتدت قدرات الطلاب بين (– ٢,٧٧) و(٢,٦٩) بمتوسط حسابي قدرة (٠.٢٦٩ لوجيت)، وامتدت الدرجات من (٣٦.١٥) و (٦٣.٤٥) بمتوسط حسابي قدره (٥١.٣٤٥ منف).

وبمقارنة التشتت النسبي للطرق الأربع نجد أن طريقة الاحتمال المقترح للإجابة الصحيحة كانت الأكثر تشتتاً حيث بلغ معامل الاختلاف لها (٤١,٩٥٪) تليها الطريقة التجريبية بمعامل اختلاف يساوي (٣٩,٧٤٪) ، ثم الطريقة التقليدية بمعامل اختلاف يساوي (٣٧,٨٥٪ ) ، فيما جاء التشتت النسبي الأقل عند تقدير الدرجات وفق النموذج اللوغاريتمي الثلاثي البارامتر بمعامل اختلاف يساوي ( ٩.٣٤٪).

## [ ٨ ] – نتائج التساؤل الثامن وتفسيرها:

والذي ينص على أنه " ما درجة الارتباط/الاختلاف بين قدرات الطلاب عند تقديرها باستخدام النموذج اللوجستي الثلاثي البارامتر بتقديرات درجاتهم عند استخدام كل من الطرق الكلاسيكية لتقدير الدرجات التي شملتها الدراسة (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) ؟ " ، وللإجابة عن هذا التساؤل تمت الاجابة عن التساؤلين الفرعيين التاليين:

**أولاً:** " ما درجة ارتباط قدرات الطلاب عند تقديرها باستخدام النموذج اللوجستي الثلاثي البارامتر بتقديرات درجاتهم عند استخدام كل من الطرق الكلاسيكية لتقدير الدرجات التي شملتها الدراسة (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) ؟ "

للإجابة عن التساؤل الفرعي السابق تم حساب معامل ارتباط بيرسون بين قدرات الطلاب مقدرة باستخدام النموذج اللوغاريتمي الثلاثي البارامتر وتقديرات درجاتهم عند استخدام كل من الطرق الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، ولقد تراوحت قيم معاملات الارتباط ما بين (٠.٧١٦-٨٧٩)؛ حيث أتضح وجود علاقة ارتباطية قوية بين قدرات الطلاب المقدرة بالنموذج اللوغاريتمي ثلاثي البارامتر وكل من طرق تقدير درجاتهم باستخدام الطرق الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، وتأخذ هذه العلاقة أكبر قيمة لها عند الطريقة التقليدية حيث بلغت قيمة معامل الإرتباط (٠,٨٧٩) وهى تشير إلى علاقة ارتباطية ايجابية قوية جداً ودالة احصائياً عند مستوى دلالة (٠,٠١)، بينما كانت أقل قيمة لمعامل الارتباط مع طريقة الاحتمال المقترح حيث بلغت قيمة معامل الإرتباط (٠,٧١٦) هي تشير إلى وجود علاقة ارتباطية ايجابية قوية ودالة احصائياً عند مستوى دلالة (٠,٠١)، في حين جاءت درجات الطريقة التجريبية في مستوى وسط بين الطريقة التقليدية، وطريقة الاحتمال المقترح للإجابة الصحيحة حيث بلغت قيمة معامل الإرتباط (٠,٧٣٧) وهي تشير إلى وجود علاقة ارتباطية قوية ودالة إحصائياً عند مستوى دلالة (٠,٠١).

**ثانياً:** " ما درجة الاختلاف بين قدرات الطلاب المقدرة وفق النموذج اللوجستي الثلاثي البارامتر ووفق كل طريقة من طرق تقدير الدرجات الكلاسيكية (التقليدية ، التجريبية ، الاحتمال المقترح للإجابة الصحيحة)؟ "

تعد قيمة الارتباط قوية بين نتائج اختبارين مؤشراً على أن هذين الاختبارين يرتبان الأفراد وفق قدراتهم بشكل متشابه، إلا أن هذا لا يعني بالضرورة أن هذين الاختبارين يعطيان تقييمات متقاربة لقدرات الأفراد، لهذا تم بحث مؤشراً آخراً للمقارنة بين طرق تقدير الدرجات الداخلة في هذه الدراسة من خلال الكشف عن مدى تقارب أو تباعد هذه الطرق في تقدير قدرات الأفراد، ولإجراء هذه المقارنة تم اتباع الخطوات الآتي:

١- تحويل الدرجات المستمدة من طرق تقدير الدرجات الأربع إلى وحدة قياس موحدة، وهي الدرجات التائية، وذلك لكي نتمكن من المقارنة الكمية بين درجات الطالب على كل طريقة من طرق تقدير الدرجات الأربع.

٢- حساب متوسط الفروق المطلقة بين الدرجات التائية لكل طريقتين من طرق تقدير الدرجات، وذلك وفق المعادلة التالية(Ndalichako & Rogers,1997,586):

$$MAD_{xy} = \frac{\sum_{j=1}^{N} |T_{xj} - T_{yj}|}{N}$$

ولقد جاءت نتائج المقارنات الثنائية بين متوسط الفروق المطلقة للدرجات التائية المستمدة من طرق تقدير الدرجات الأربع (التقليدية ، التجريبية ، الاحتمال المقترح للإجابة الصحيحة ، والنموذج اللوجستي ثلاثي البارامتر) كما في جدول (١٩):

**جدول (١٩)**

**متوسط الفروق المطلقة بين درجات كل طريقتين من طرق تقدير الدرجات الداخلة في الدراسة**

| مجال المقارنة | | عدد الطلاب | متوسط الفروق المطلقة | الانحراف المعياري |
|---|---|---|---|---|
| النموذج اللوغاريتمي الثلاثي | التقليدية | ١٥٠٠ | ٣,٦ | ٢,٣ |
| | التجريبية | ١٥٠٠ | ٥,٥ | ٣,٨ |
| | الاحتمال المقترح للإجابة الصحيحة | ١٥٠٠ | ٥.٧ | ٤.١ |
| التقليدية | التجريبية | ١٥٠٠ | ٣.٥ | ٣,٢ |
| | الاحتمال المقترح للإجابة الصحيحة | ١٥٠٠ | ٣.٦ | ٣,٤ |
| التجريبية | الاحتمال المقترح للإجابة الصحيحة | ١٥٠٠ | ٣.٧ | ٣.٨ |

يتضح من جدول (١٩) السابق أن طريقة تقدير الدرجات الأكثر قرباً من درجات النموذج اللوجستي ثلاثي البارامتر هي الطريقة التقليدية حيث بلغ متوسط الفروق المطلقة بين الطريقتين (٣,٦) درجة، ويؤكد على هذه النتيجة أن الفروق المطلقة بينهما كانت الأقل تشتتاً حيث بلغت قيمة الانحراف المعياري لهذه الفروق (٢,٣) درجة وهي القيمة الأقل بين بقية المقارنات الثنائية المتبقية، وفي المقابل كانت طريقة تقدير الدرجات الأكثر اختلافاً عن النموذج اللوجستي ثلاثي البارامتر هي طريقة الاحتمال المقترح للإجابة الصحيحة حيث بلغ متوسط الفروق المطلقة بين الطريقتين (٥.٧) درجة، وبفارق بسيط عن الطريقة التجريبية بلغ (٠.٠٢) درجة.

وفيما يخص المقارنات الثنائية بين الطرق الكلاسيكية فقد أظهرت النتائج أن الطريقتين التقليدية والتجريبية هما الأقرب في تقدير الدرجات حيث بلغ متوسط الفروق المطلقة بينهما (٣,٥) درجة، في حين كانت الطريقتين التجريبية والاحتمال المقترح للإجابة الصحيحة هما الأكثر اختلافاً في تقدير الدرجات حيث بلغ متوسط الفروق المطلقة بينهما (٣,٧) درجة.

وفيما يخص المقارنة بين الطرق الأربع ككل فقد أظهرت النتائج أن الطريقتين الأكثر قرباً من بعضهما في تقدير قدرات الطلاب هما الطريقة التقليدية والطريقة التجريبية حيث بلغ متوسط الفروق المطلقة بينهما (٣,٥) درجة، في حين كانت الطريقتين الأكثر بعداً عن بعضهما البعض هما طريقة الاحتمال المقترح للإجابة الصحيحة والنموذج اللوجستي ثلاثي البارامتر حيث بلغ متوسط الفروق المطلقة بينهما (٥,٧) درجة.

## [ ٩ ] – نتائج التساؤل التاسع وتفسيرها:

والذي ينص على أنه " ما درجات ارتباط قيم معاملات صعوبة/تمييز الفقرات عند استخدام كل من الطرق الكلاسيكية لتقدير درجات الاختيار من متعدد (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) مع قيم معاملات صعوبة/تمييز الفقرات عند استخدام النموذج اللوجستي الثلاثي البارامتر؟ " ، وللإجابة عن هذا التساؤل تمت الاجابة عن التساؤلين الفرعيين التاليين:

**أولاً:** ما درجات ارتباط قيم معاملات صعوبة الفقرات عند استخدام كل من الطرق الكلاسيكية لتقدير درجات الاختيار من متعدد (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) مع قيم معاملات صعوبة الفقرات عند استخدام النموذج اللوجستي الثلاثي البارامتر؟ "

للإجابة عن التساؤل الفرعي السابق تم حساب معاملات صعوبة الفقرات وفق طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) وبارامتر صعوبة الفقرات في النموذج اللوجستي الثلاثي البارامتر وجدول (٢٠) التالي يلخص هذه الإحصائيات، حيث تم تحديد أقل وأعلى قيمة لمعاملات الصعوبة والمتوسطات الحسابية والانحرافات المعيارية وفقا لكل طريقة من طرق تقدير الدرجات الكلاسيكية والنموذج اللوجستي الثلاثي البارامتر:

**جدول (٢٠)**

**قيم صعوبة الفقرات المقدرة وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) والنموذج اللوجستي الثلاثي البارامتر**

| طريقة تقدير الدرجات | عدد الفقرات | أقل قيمة لمعامل الصعوبة | | أعلى قيمة لمعامل الصعوبة | | المدى بوحدات (Z) | المتوسط | الانحراف المعياري |
|---|---|---|---|---|---|---|---|---|
| | | بوحدات الصعوبة | بدرجات (Z) | بوحدات الصعوبة | بدرجات (Z) | | | |
| التقليدية | ٥٠ | ٠.٣٠٩ | – ١.٥٨ | ٠.٦٢٧ | ٣.٦٤ | ٥.٢٢ | ٠.٤٠٩ | ٠.٠٦٣ |
| التجريبية | ٥٠ | ٠,٢٦٠ | –٢.١٨ | ٠.٥٨١ | ٢.٢٢ | ٤.٤٠ | ٠.٤١٩ | ٠,٠٧٣ |
| الاحتمال المقترح للإجابة الصحيحة | ٥٠ | ٠.٢٣١ | – ٢.٣١ | ٠.٥٦٢ | ١.٩٩ | ٤,٣٠ | ٠.٤٠٩ | ٠.٠٧٧ |
| النموذج اللوجستي الثلاثي البارامتر | ٥٠ | ٠.٣٤٧ | – ١.٩٣ | ٢.٢٥١ | ٢.٢٩ | ٤.٢٢ | ١.٢١٧ | ٠.٤٥١ |

ولقد تم حساب مصفوفة معاملات الإرتباط البينية بين قيم صعوبة الفقرات المقدرة وفق النموذج اللوجستي الثلاثي البارامتر وقيم صعوبة الفقرات المقدرة وفق كل طريقة من طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، ولقد تراوحت قيم معاملات الارتباط بين (٠.٧٩٢ – ٠.٩٦١)؛ حيث أتضح وجود علاقة ارتباطية ذات دلالة إحصائية عند مستوى الدلالة (٠,٠١) وبدرجة ارتباط قوية بين قيم معاملات الصعوبة المقدرة وفق النموذج اللوجستي الثلاثي البارامتر ووفق طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، ويلاحظ أن وجود الإشارة السالبة لا يعني العلاقة العكسية بل يعود إلى المنطق المعكوس لقيم معاملات الصعوبة وفق طرق القياس الكلاسيكية، كما تظهر البيانات أن معاملات الصعوبة الكلاسيكية المستمدة من الطريقة التقليدية كانت هي الأكثر ارتباطاً ببارامتر صعوبة النموذج اللوجستي الثلاثي البارامتر وذلك بمعامل ارتباط بلغت قيمته (٠,٨٣٦) والتي تشير إلى درجة ارتباط قوية، في حين كانت معاملات الصعوبة المستمدة من طريقة الاحتمال المقترح للإجابة الصحيحة هي الأقل ارتباطاً بالنموذج اللوجستي الثلاثي البارامتر حيث بلغت قيمة معامل الارتباط بينهما (٠,٧٩٢) وتشير هذه القيمة إلى درجة ارتباط قوية.

وفيما يخص المقارنات الثنائية بين معاملات الصعوبة المقدرة وفق الطرق الكلاسيكية فقد أظهرت النتائج أن قيم معاملات الارتباط بين طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) كانت قوية وتتفوق على قيم ارتباط أي منها بالنموذج اللوجستي الثلاثي البارامتر، وقد كانتا الطريقتين التجريبية والاحتمال المقترح للإجابة الصحيحة هما الأعلى ارتباطاً في تقدير قيم معاملات الصعوبة حيث بلغ معامل

الارتباط بينهما (٠.٩٦١) وهي تشير إلى درجة ارتباط قوية، في حين كانتا الطريقتين الأقل ارتباطاً هما التقليدية والاحتمال المقترح للإجابة الصحيحة بقيمة ارتباط (٠.٩٣٤) ومع ذلك فإن هذه القيمة تشير إلى درجة ارتباط قوية.

وفيما يخص المقارنات الثنائية بين قيم معاملات الصعوبة المقدرة وفق طرق تقدير الدرجات الأربع ككل فقد أظهرت النتائج أن الطريقتين الأكثر ارتباطاً في تقدير قيم صعوبة الفقرات كانتا الطريقة التجريبية وطريقة الاحتمال المقترح للإجابة الصحيحة بمعامل ارتباط بلغت قيمته (٠.٩٦١) والتي تشير إلى درجة ارتباط قوية، في حين كانت الطريقتين الأقل ارتباطاً في تقدير قيم صعوبة الفقرات هما طريقة الاحتمال المقترح للإجابة الصحيحة والنموذج اللوجستي الثلاثي البارامتر بمعامل ارتباط بلغت قيمته (٠،٧٩٢) ومع ذلك فإن هذه القيمة تدل على ارتباط قوي.

**ثانياً:** ما درجات ارتباط قيم معاملات تمييز الفقرات عند استخدام كل من الطرق الكلاسيكية لتقدير درجات الاختيار من متعدد (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) مع قيم معاملات تمييز الفقرات عند استخدام النموذج اللوجستي الثلاثي البارامتر؟ "

للإجابة عن التساؤل الفرعي السابق تم حساب معاملات تمييز الفقرات وفق طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) وبارامتر تمييز الفقرات في النموذج اللوجستي الثلاثي البارامتر، وجدول (٢١) يخلص هذه الإحصائيات، حيث تم تحديد أقل وأعلى قيمة لمعاملات التمييز والمتوسطات الحسابية والانحرافات المعيارية وفقا لكل طريقة من طرق تقدير الدرجات الكلاسيكية والنموذج اللوجستي الثلاثي البارامتر:

**جدول (٢١)**

**قيم تمييز الفقرات المقدرة وفق طرق تقدير الدرجات الكلاسيكية (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة) والنموذج اللوجستي الثلاثي البارامتر**

| طريقة تقدير الدرجات | عدد الفقرات | أقل قيمة لمعامل التمييز | | أعلى قيمة لمعامل التمييز | | المدى بوحدات (Z) | المتوسط | الانحراف المعياري |
|---|---|---|---|---|---|---|---|---|
| | | بوحدات التمييز | بدرجات (Z) | بوحدات التمييز | بدرجات (Z) | | | |
| التقليدية | ٥٠ | ٠.٣٢٢ | ١.٩٤- | ٠.٦٣٠ | ١.٩١ | ٣.٨٥ | ٠.٤٧٧ | ٠.٠٨٠ |
| التجريبية | ٥٠ | ٠.٣٣١ | ١.٨٨ - | ٠,٦١٩ | ٢,٣٥ | ٤.٢٣ | ٠.٤٥٩ | ٠.٠٦٨ |
| الاحتمال المقترح للإجابة الصحيحة | ٥٠ | ٠.٣٠٨ | ٢.٠٢ - | ٠.٥٧١ | ٢.٠٣ | ٤.٠٥ | ٠.٤٣٩ | ٠.٠٦٥ |
| النموذج اللوجستي الثلاثي البارامتر | ٥٠ | ٠.٥٨٤ | ٢.٧٣- | ١.٩٦٠ | ١.٦٢ | ٤.٣٥ | ١.٤٤٨ | ٠.٣١٦ |

ولقد تم حساب مصفوفة معاملات الإرتباط البينية بين قيم تمييز الفقرات المقدرة وفق النموذج اللوجستي الثلاثي البارامتر، وقيم تمييز الفقرات المقدرة وفق كل طريقة من طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، ولقد تراوحت قيم معاملات الارتباط ما بين (٠.٦٣٦- ٠.٨٩٧)؛ حيث أتضح وجود علاقة ارتباطية ذات دلالة إحصائية عند مستوى دلالة (٠,٠١) بين قيم معاملات التمييز المقدرة وفق النموذج اللوجستي الثلاثي البارامتر ووفق طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)؛ حيث كانت الطريقة التقليدية هي الأعلى ارتباطاً بالنموذج اللوجستي الثلاثي البارامتر بمعامل ارتباط بلغت قيمته (٠,٧٢٣) وهي تشير إلى علاقة ارتباطية قوية، ثم جاءت الطريقة التجريبية بمعامل ارتباط بلغت قيمته (٠,٦٨١) ثم طريقة الاحتمال المقترح للإجابة الصحيحة بمعامل ارتباط بلغت قيمته (٠,٦٣٦).

وفيما يخص المقارنات الثنائية بين معاملات التمييز المقدرة وفق الطرق الكلاسيكية فقد أظهرت النتائج أن قيم معاملات الارتباط بين طرق تقدير الدرجات الكلاسيكية الثلاث (التقليدية، التجريبية، الاحتمال المقترح للإجابة الصحيحة)، كانت قوية وتتفوق على قيم ارتباط أي منها بالنموذج اللوجستي الثلاثي البارامتر، وقد كانتا الطريقتين التجريبية والاحتمال المقترح للإجابة الصحيحة هما الأعلى ارتباطاً في تقدير قيم معاملات التمييز؛ حيث بلغ معامل الارتباط بينهما (٠,٨٩٧) وهي تشير إلى درجة ارتباط قوية، في حين كانتا الطريقتين الأقل ارتباطاً هما التقليدية والتجريبية بقيمة ارتباط (٠,٨٤٦) ومع ذلك فإن هذه القيمة تشير إلى درجة ارتباط قوية.

وفيما يخص المقارنات الثنائية بين قيم تمييز الفقرات المقدرة وفق طرق تقدير الدرجات الأربع ككل فقد أظهرت النتائج أن الطريقتين الأكثر ارتباطاً في تقدير قيم تمييز الفقرات كانتا الطريقة التجريبية وطريقة الاحتمال المقترح للإجابة بمعامل ارتباط بلغت قيمته (٠,٨٩٧) والتي تشير إلى درجة ارتباط قوية، في حين كانت الطريقتين الأقل ارتباطاً في تقدير قيم تمييز الفقرات هما طريقة الاحتمال المقترح للإجابة الصحيحة والنموذج اللوجستي الثلاثي البارامتر بمعامل ارتباط بلغت قيمته (٠,٦٣٦).

## [ ١٠ ] – نتائج التساؤل العاشر وتفسيرها:

والذي ينص على أنه " هل تختلف دقة معادلة درجات الاختبارات باختلاف طريقتي المعادلة (المتوسط/المتوسط، المتوسط/الانحراف المعياري) باستخدام النموذج اللوجستي الثلاثي البارامتر، لأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار(٢٥، ٥٠)؛ في ضوء محكي التحيز وجذر متوسط مربع الخطأ؟ "

وللإجابة عن التساؤل السابق تمت المقارنة بين الأحجام المختلفة للعينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولين للاختبار(٢٥، ٥٠) ؛ وذلك بأخذ قيم جذر متوسط مربع الخطأ (RMSE) والمتوسط الحسابي للتحيز (BIAS) عند اختلاف حجم العينة ونسبة الفقرات المشتركة إلى عدد فقرات الاختبار الكلي، وطول الاختبار، واختلاف طريقتي المعادلة (المتوسط/المتوسط M–M، المتوسط/الانحراف المعياري M–SD)؛ بحيث يكون الأثر لمتغير حجم العينة، وطول الاختبار، عند تحليل البيانات وفق النموذج اللوجستي الثلاثي البارامتر.

وجدول (٢٢) التالي يوضح قيم (BIAS, RMSE) عند اختلاف حجم العينة ونسبة الفقرات المشتركة وطول الاختبار وطريقة المعادلة وفق النموذج اللوجستي الثلاثي البارامتر.

**جدول (٢٢)**

**قيم (BIAS, RMSE) عند اختلاف حجم العينة ونسبة الفقرات المشتركة وطول الاختبار وطريقة المعادلة وفق النموذج اللوجستي الثلاثي البارامتر**

| حجم العينة | طول الاختبار | | عدد الفقرات (٢٥) | | عدد الفقرات (٥٠) | |
|---|---|---|---|---|---|---|
| | نسبة الفقرات | الطريقة | BIAS (التحيز) | RMSE (جذر متوسط مربع الخطأ) | BIAS (التحيز) | RMSE (جذر متوسط مربع الخطأ) |
| ٥٠٠ | ٪١٠ | M–M | ٠.٠٤١ | ٠.٠٥٨ | ٠.٠٣٧ | ٠.٠٥١ |
| | | M–SD | ٠.٠٥٢ | ٠.٠٦٩ | ٠.٠٣٩ | ٠.٠٥١ |
| | ٪٢٠ | M–M | ٠.٠١٣ | ٠.٠٣٤ | ٠.٠١٣ | ٠.٠٢٧ |
| | | M–SD | ٠.٠٢٤ | ٠.٠٤٩ | ٠.٠١٥ | ٠.٠٢٧ |
| | ٪٣٠ | M–M | ٠.٠١٢ | ٠.٠٣١ | ٠.٠٠٧ | ٠.٠٢١ |
| | | M–SD | ٠.٠٣٧ | ٠.٠٥٦ | ٠.٠١٠ | ٠.٠٢١ |
| ١٠٠٠ | ٪١٠ | M–M | ٠.٠٣٦ | ٠.٠٥٢ | ٠.٠٠٤ | ٠.٠١٩ |
| | | M–SD | ٠.٠٥٨ | ٠.٠٦٠ | ٠.٠١١ | ٠.٠٢١ |
| | ٪٢٠ | M–M | ٠.٠١٠ | ٠.٠٢٩ | ٠.٠٢٥ | ٠.٠٣٥ |
| | | M–SD | ٠.٠٢١ | ٠.٠٣٦ | ٠.٠٢٩ | ٠.٠٣٨ |
| | ٪٣٠ | M–M | ٠.٠٠٤ | ٠.٠٢٩ | ٠.٠٠٩ | ٠.٠٢٣ |
| | | M–SD | ٠.٠٠٢ | ٠.٠٣٦ | ٠.٠١٣ | ٠.٠٢٣ |
| ١٥٠٠ | ٪١٠ | M–M | ٠.٠١١ | ٠.٠٢٦ | ٠.٠٢١ | ٠.٠٣٠ |
| | | M–SD | ٠.٠١٥ | ٠.٠٣٧ | ٠.٠١٢ | ٠.٠٣٢ |
| | ٪٢٠ | M–M | ٠.٠٠٤ | ٠.٠٢١ | ٠.٠٠٥ | ٠.٠١٩ |
| | | M–SD | ٠.٠١١ | ٠.٠٣٢ | ٠.٠٠٧ | ٠.٠١٩ |
| | ٪٣٠ | M–M | ٠.٠٢٢ | ٠.٠٢٥ | ٠.٠١٠ | ٠.٠٢٣ |
| | | M–SD | ٠.٠٢٥ | ٠.٠٣٧ | ٠.٠١٨ | ٠.٠٢٧ |

يتضح من جدول (٢٢) السابق أن قيم التحيز تختلف باختلاف طريقة المعادلة المستخدمة، وأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار (٢٥، ٥٠) وفق النموذج اللوجستي الثلاثي البارامتر؛ وذلك باختلاف نسبة الفقرات المشتركة إلى نسبة فقرات الاختبار الكلي (١٠٪ ، ٢٠٪ ، ٣٠٪) فكلما قلت قيمة التحيز دل ذلك على دقة أكثر في معادلة درجات الاختبارات.

وللإجابة على التساؤل السابق فقد تم حساب المتوسطات الحسابية للتحيز (BIAS) ولجذر متوسط مربع الخطأ (RMSE)، ويظهر جدول (٢٣) التالي المتوسطات الحسابية لكل من (BIAS, RMSE) وفقاً لطريقتي المعادلة (المتوسط/المتوسط ، المتوسط/الانحراف المعياري) باختلاف حجم العينة وطول الاختبار وفق النموذج اللوجستي الثلاثي البارامتر.

**جدول (٢٣)**

**قيم (BIAS, RMSE) باستخدام طريقتي المعادلة واختلاف حجم العينة وطول الاختبار وفق النموذج اللوجستي الثلاثي البارامتر**

| طريقة المعادلة | حجم العينة | عدد الفقرات (٢٥) | | عدد الفقرات (٥٠) | |
|---|---|---|---|---|---|
| | | BIAS (التحيز) | RMSE (جذر متوسط مربع الخطأ) | BIAS (التحيز) | RMSE (جذر متوسط مربع الخطأ) |
| المتوسط/المتوسط | ٥٠٠ | ٠.٠٢٢ | ٠.٠٤١ | ٠.٠١٩ | ٠.٠٣٣ |
| | ١٠٠٠ | ٠.٠١٦ | ٠.٠٣٦ | ٠.٠١٣ | ٠.٠٢٦ |
| | ١٥٠٠ | ٠.٠١٢ | ٠.٠٢٤ | ٠.٠١٢ | ٠.٠٢٤ |
| المتوسط/الانحراف المعياري | ٥٠٠ | ٠.٠٣٧ | ٠.٠٥٨ | ٠.٠٢١ | ٠.٠٣٣ |
| | ١٠٠٠ | ٠.٠٢٧ | ٠.٠٤٤ | ٠.٠١٧ | ٠.٠٢٧ |
| | ١٥٠٠ | ٠.٠١٧ | ٠.٠٣٥ | ٠.٠١٢ | ٠.٠٢٦ |

## ● النتائج المتعلقة بنماذج الاختبار الذي يتألف من (٢٥) فقرة:

### أولاً: طريقة (المتوسط/المتوسط)

يتضح من جدول (٢٣) السابق ووفقاً لطريقة (المتوسط/المتوسط) أن قيم التحيز (BIAS) كانت على التوالي (٠.٠٢٢، ٠.٠١٦، ٠.٠١٢) ؛ حيث كانت أعلى قيمة للتحيز(٠.٠٢٢)عندما كان حجم العينة (٥٠٠) ، بينما كانت أقل قيمة للتحيز (٠.٠١٢)عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة التحيز للمعادلة (BIAS).

كما يتضح من الجدول السابق أن قيم جذر متوسط مربع الخطأ (RMSE) كانت على التوالي (٠.٠٤١، ٠.٠٣٦، ٠.٠٢٤) ؛ حيث كانت أعلى قيمة لجذر متوسط مربع الخطأ (٠.٠٤١)عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة (٠.٠٢٤) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة جذر متوسط مربع الخطأ للمعادلة (RMSE).

ومن ثم يتضح أنه عند استخدام طريقة (المتوسط/المتوسط)، وفق النموذج اللوجستي الثلاثي البارامتر، وطول الاختبار(٢٥) فقرة، أنه كلما زاد حجم العينة تقل قيمة التحيز(BIAS) وقيمة جذر متوسط مربع الخطأ (RMSE).

## ثانياً: طريقة (المتوسط/الانحراف المعياري).

يتضح من جدول (٢٣) السابق ووفقا لطريقة (المتوسط/الانحراف المعياري) أن قيم التحيز (BIAS) كانت على التوالي (٠.٠٣٧، ٠.٠٢٧، ٠.٠١٧) ؛ حيث كانت أعلى قيمة للتحيز (٠.٠٣٧) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة للتحيز(٠.٠١٧) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة التحيز للمعادلة (BIAS).

كما يتضح من الجدول السابق أن قيم جذر متوسط مربع الخطأ (RMSE) كانت على التوالي (٠.٠٥٨، ٠.٠٤٤، ٠.٠٣٥) ؛ حيث كانت أعلى قيمة لجذر متوسط مربع الخطأ (٠.٠٥٨) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة (٠.٠٣٥) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة جذر متوسط مربع الخطأ للمعادلة (RMSE).

ومن ثم يتضح أنه عند استخدام طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر، وطول الاختبار(٢٥) فقرة، أنه كلما زاد حجم العينة تقل قيمة التحيز(BIAS) وقيمة جذر متوسط مربع الخطأ (RMSE).

### ثالثاً: المقارنة بين طريقتي المعادلة (المتوسط/المتوسط ، المتوسط/الانحراف المعياري)

يتضح أنه عند المقارنة بين الطريقتين، طريقة (المتوسط/المتوسط) وطريقة (المتوسط/ الانحراف المعياري)، للاختبار المكون من (٢٥) فقرة، لأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠) يتضح أن هناك اختلاف بين الطريقتين في دقة معادلة درجات الاختبار.

فيما يتعلق بقيم التحيز (BIAS) عند استخدام طريقة (المتوسط/المتوسط) كانت على التوالي (٠.٠٢٢، ٠.٠١٦، ٠.٠١٢) ، وعند استخدام طريقة (المتوسط/الانحراف المعياري) كانت على التوالي (٠.٠٣٧، ٠.٠٢٧، ٠.٠١٧) ، ويتضح أن أصغر قيمة للتحيز كانت عند استخدام طريقة (المتوسط/المتوسط) عند جميع أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وأن العينات الكبيرة تعطي دقة أكبر في معادلة درجات الاختبار من العينات الصغيرة، وعليه تشير نتائج التحيز(BIAS) أن استخدام طريقة (المتوسط/المتوسط) أكثر دقة في معادلة درجات الاختبارات مقارنة مع طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر ، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطول الاختبار(٢٥) فقرة.

وفيما يتعلق بقيم جذر متوسط مربع الخطأ (RMSE) عند استخدام طريقة (المتوسط/المتوسط) كانت على التوالي (٠.٠٤١، ٠.٠٣٦، ٠.٠٢٤)، وعند استخدام طريقة (المتوسط/الانحراف المعياري) كانت على التوالي (٠.٠٥٨، ٠.٠٤٤، ٠.٠٣٥)، ويتضح أن أصغر قيمة لجذر متوسط مربع الخطأ كانت عن استخدام طريقة (المتوسط/المتوسط) عند جميع أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وأن العينات الكبيرة تعطي دقة أكبر في معادلة درجات الاختبار من العينات الصغيرة، وعليه تشير نتائج جذر متوسط مربع الخطأ (RMSE) أن طريقة (المتوسط/المتوسط) أكثر دقة في معادلة درجات الاختبارات مقارنة مع طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطول الاختبار(٢٥) فقرة.

## ●النتائج المتعلقة بنماذج الاختبار الذي يتألف من (٥٠) فقرة:

### أولاً: طريقة (المتوسط/ المتوسط)

يتضح من جدول (٢٣) السابق ووفقاً لطريقة (المتوسط/المتوسط) أن قيم التحيز (BIAS) كانت على التوالي (٠.٠١٩، ٠.٠١٣، ٠.٠١٢)؛ حيث كانت أعلى قيمة للتحيز(٠.٠١٩) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة للتحيز (٠.٠١٢) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة التحيز للمعادلة (BIAS).

كما يتضح من الجدول السابق أن قيم جذر متوسط مربع الخطأ (RMSE) كانت على التوالي (٠.٠٣٣، ٠.٠٢٦، ٠.٠٢٤)؛ حيث كانت أعلى قيمة لجذر متوسط مربع الخطأ (٠.٠٣٣) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة (٠.٠٢٤) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة جذر متوسط مربع الخطأ للمعادلة (RMSE).

ومن ثم يتضح أنه عند استخدام طريقة (المتوسط/المتوسط)، وفق النموذج اللوجستي الثلاثي البارامتر، وطول الاختبار(٥٠) فقرة، أنه كلما زاد حجم العينة تقل قيمة التحيز (BIAS) وقيمة جذر متوسط مربع الخطأ (RMSE).

**ثانياً: طريقة (المتوسط/ الانحراف المعياري)**

يتضح من جدول (٢٣) السابق ووفقا لطريقة (المتوسط/الانحراف المعياري) أن قيم التحيز (BIAS) كانت على التوالي (٠.٠٢١، ٠.٠١٧، ٠.٠١٢)؛ حيث كانت أعلى قيمة للتحيز (٠.٠٢١) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة للتحيز(٠.٠١٢) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة التحيز للمعادلة (BIAS).

كما يتضح من الجدول السابق أن قيم جذر متوسط مربع الخطأ (RMSE) كانت على التوالي (٠.٠٣٣، ٠.٠٢٧، ٠.٠٢٦)؛ حيث كانت أعلى قيمة لجذر متوسط مربع الخطأ (٠.٠٣٣) عندما كان حجم العينة (٥٠٠)، بينما كانت أقل قيمة (٠.٠٢٦) عندما كان حجم العينة (١٥٠٠)، وتشير هذه القيم إلى أن ارتفاع حجم العينة يقلل من قيمة جذر متوسط مربع الخطأ للمعادلة (RMSE).

ومن ثم يتضح أنه عند استخدام طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر، وطول الاختبار(٥٠) فقرة، أنه كلما زاد حجم العينة تقل قيمة التحيز(BIAS) وقيمة جذر متوسط مربع الخطأ (RMSE).

**ثالثاً: المقارنة بين طريقتي المعادلة (المتوسط/المتوسط ، المتوسط/الانحراف المعياري)**

يتضح أنه عند المقارنة بين الطريقتين، طريقة (المتوسط/المتوسط) وطريقة (المتوسط/الانحراف المعياري)، للاختبار المكون من (٥٠) فقرة، لأحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠) يتضح أن هناك اختلاف بين الطريقتين في دقة معادلة درجات الاختبار.

فيما يتعلق بقيم التحيز (BIAS) عند استخدام طريقة (المتوسط/المتوسط) كانت على التوالي (٠.٠١٩، ٠.٠١٣، ٠.٠١٢)، وعند استخدام طريقة (المتوسط/الانحراف المعياري) كانت على التوالي (٠.٠٢١، ٠.٠١٧، ٠.٠١٢)، ويتضح أن أصغر قيمة للتحيز كانت عند استخدام طريقة (المتوسط/المتوسط) عند جميع أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وأن العينات الكبيرة تعطي دقة أكبر في معادلة درجات الاختبار من العينات الصغيرة، وعليه تشير نتائج التحيز (BIAS) أن استخدام طريقة (المتوسط/المتوسط) أكثر دقة في معادلة درجات الاختبارات مقارنة مع طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطول الاختبار(٥٠) فقرة.

وفيما يتعلق بقيم جذر متوسط مربع الخطأ (RMSE) عند استخدام طريقة (المتوسط/المتوسط) على التوالي (٠.٠٣٣، ٠.٠٢٦، ٠.٠٢٤)، وعند استخدام طريقة (المتوسط/الانحراف المعياري) كانت على التوالي (٠.٠٣٣، ٠.٠٢٧، ٠.٠٢٦)، ويتضح أن أصغر قيمة لجذر متوسط مربع الخطأ كانت عن استخدام طريقة (المتوسط/المتوسط) عند جميع أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وأن العينات الكبيرة تعطي دقة أكبر في معادلة درجات الاختبار من العينات الصغيرة، وعليه تشير نتائج جذر متوسط مربع الخطأ (RMSE) أن استخدام طريقة (المتوسط/المتوسط) أكثر دقة في معادلة درجات الاختبارات مقارنة مع طريقة (المتوسط/الانحراف المعياري)، وفق النموذج اللوجستي الثلاثي البارامتر، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطول الاختبار(٥٠) فقرة.

يتضح مما سبق أن طريقة (المتوسط/المتوسط) تعتبر أكثر دقة في معادلة درجات الاختبارات مقارنة مع طريقة (المتوسط/الانحراف المعياري) وفق النموذج اللوجستي الثلاثي البارامتر، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار (٢٥ ، ٥٠) ونسبة الفقرات المشتركة (١٠٪ ،٢٠٪ ، ٣٠٪)، كما أنه تقل قيم التحيز (BIAS) وجذر متوسط مربع الخطأ (RMSE)، مع ازدياد حجم العينة وطول الاختبار، عند استخدام طريقة (المتوسط/المتوسط) وفق النموذج اللوجستي الثلاثي البارامتر، عند أحجام العينات (٥٠٠، ١٠٠٠، ١٥٠٠)، وطولي الاختبار(٢٥ ، ٥٠ )، فضلاً عن أن هناك علاقة طردية بين زيادة حجم العينة وفق طولي الاختبار في دقة المعادلة، فكلما زاد حجم العينة زادت دقة المعادلة.

## مراجع الدراسة


ابتسام عيسى خصاونة (٢٠١٢). *أثر اختلاف الأوزان النسبية لقواعد انتهاك صياغة فقرات الاختيار من متعدد في الاختبارات على خصائصها السيكومترية* (رسالة دكتوراه غير منشورة). كلية التربية، جامعة اليرموك.

إبراهيم محمد يعقوب، باسل خميس أبو فودة (٢٠١٠). أثر مخالفة قواعد صياغة فقرات الاختيار من متعدد على التقديرات المختلفة لنظرية استجابة الفقرة. *مجلة كلية التربية، كلية التربية، جامعة الإسكندرية*، ٢٠ (٢)، ٥٢-٨٩.

إبراهيم محمد يعقوب، باسل خميس أبو فودة (٢٠١٢). اثر مخالفة قواعد صياغة فقرات الاختيار من متعدد على الخصائص السيكومترية للاختبار وفقراته. *مجلة جامعة دمشق للعلوم التربوية والنفسية، جامعة دمشق*، ٢٨ (١)، ٤١٩-٤٤٣.

أحمد عودة (٢٠١٤). *القياس والتقويم في العملية التدريسية*. أربد: دار الأمل للنشر والتوزيع.

أحمد محمد التقي (٢٠١٣). *النظرية الحديثة في القياس* (ط٢). عمان: دار المسيرة للنشر والتوزيع والطباعة.

أمينة محمد كاظم (١٩٨٨). دراسة نظرية نقدية حول القياس الموضوعي للسلوك (نموذج راش). الكويت: مؤسسة الكويت للتقدم العلمي.

آن أناستازي، سوزانا أوربينا (٢٠١٥). *القياس النفسي* (ترجمة: صلاح الدين محمود علام). عمان: دار الفكر للنشر والتوزيع.

باسل خميس أبو فودة (٢٠١٤). أثر إعادة ترتيب بدائل الاستجابة في صعوبة فقرة الاختيار من متعدد. مجلة *دراسات عربية في التربية وعلم النفس*، (٥٣)، ٢٦٥-٢٨٧.

باسل خميس أبو فودة ؛ نجاتي أحمد يونس (٢٠١٢). الاختبارات التحصيلية المدرسية (أسس بناء وتحليل الأسئلة). عمان: دار المسيرة للطابعة والنشر.

حابس سعد الزبون (٢٠١٣). أثر حجم العينة على تقدير دالة المعلومات للاختبار والخطأ المعياري في تقديرها باستخدام النظرية الحديثة في القياس. *مجلة جامعة النجاح للأبحاث، جامعة النجاح الوطنية*، ٢٧ (٦)، ١٣١٣-١٣٣٤.

حمدي يونس أبو جراد (٢٠١٧). فاعلية النموذج اللوجستي ثلاثي المعلمة في معايرة مفردات اختبار تحصيلي محكي المرجع في مقرر الرياضيات للصف السابع. *إربد للبحوث والدراسات، جامعة إربد الأهلية، ١٩* (١)، ٢٥٣-٢٨٨.

حيدر إبراهيم ظاظا (٢٠١٢). الكشف عن مدى انتهاك قواعد صياغة فقرة الاختيار من متعدد في أسئلة شهادة الدراسة الثانوية العامة في الأردن. *المجلة الأردنية في العلوم التربوية، جامعة اليرموك،٨*(١)، ٨١-٩١.

رحاب سعيد الحكماني (٢٠٠٨). مقارنة بين النظرية الكلاسيكية للاختبار ونظرية الاستجابة للمفردة في تقدير قدرات الأفراد ومدى استقرار مؤشرات المفردات الاختبارية. *المجلة التربوية، جامعة الكويت*، ٢٣ (٨٩)، ٢٥٣- ٢٥٩.

ساري سليم سواقد (١٩٩٢). *اختبار صحة الافتراضات النظرية لطرق التصحيح لأثر التخمين، ومقارنة أثر استخدام هذه الطرق على الخصائص السيكومترية للفقرة* (رسالة دكتوراه غير منشورة). كلية الدراسات العليا، الجامعة الأردنية.

شاهر خالد سليمان، علي محمد الصالح (٢٠١٧). أثر موقع البديل الصحيح في اختبار اختيار من متعدد على تقديرات معالم الفقرات والقدرة وفق النموذج اللوجستي ثلاثي المعلمة. *دراسات عربية في التربية وعلم النفس، رابطة التربويين العرب*، ٩٠، ٩٨-١٢٠.

صبري حسن الطراونة (٢٠١٥). مدى انتهاك قواعد كتابة فقرات الاختيار من متعدد في اختبارات الكفاءة في اللغة العربية واللغة الإنجليزية بجامعة مؤتة. *مجلة التربية، كلية التربية، جامعة الأزهر*، ١٦٣ (٢)، ٥٧١-٥٩٤.

صلاح الدين محمود علام (٢٠٠٥). *نماذج الاستجابة للمفردة الاختبارية أحادية البعد ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي.* القاهرة: دار الفكر العربي.

صلاح الدين محمود علام (٢٠٠٧). *الاختبارات التشخيصية مرجعية المحك في المجالات التربوية والنفسية والتدريبية* (ط٢). القاهرة: دار الفكر العربي.

صلاح الدين محمود علام (٢٠١٥). *القياس والتقويم التربوي والنفسي: أساسية وتطبيقاته وتوجهاته المعاصرة* (ط٦). القاهرة: دار الفكر العربي.

صلاح شريف عبد الوهاب ( ٢٠٠١). أثر بعض الطرق الوزنية لتقدير الدرجات على صدق الاختبارات مرجعية المحك ذات الاختيار من متعدد، *مجلة كلية التربية ببنها، جامعة بنها*، ١٢ (٤٩)، ٢٠٢-٢٥٥.

طه الخرشه (٢٠١٦). أثر طرق معالجة أثر التخمين على تقدير إحصائيات الأفراد والفقرات في اختبارات الاختيار من متعدد وفق النظرية الحديثة في القياس. *مجلة جامعة النجاح للأبحاث، جامعة النجاح الوطنية*، ٣٠ (١٢)، ٢٣٤٨-٢٣٦٦.

عبد الرحمن عبد الله النفيعي (٢٠١٢). الخصائص السيكومترية لاختبار المصفوفات المتتابعة المتقدم في ضوء نظرية الاستجابة للمفردة الاخبارية. *مجلة التربية، جامعة الأزهر*،١٤٧ (٢)، ١٧٥-٢١٤.

عزالدين عبدالله النعيمي (٢٠١٥). معالم الفقرات والأفراد وخاصية اللا تغير في الاختبارات الوطنية لضبط جودة التعليم في الأردن مقارنة بين النظرية الكلاسيكية والنظرية الحديثة في القياس. *مجلة اتحاد الجامعات العربية للتربية وعلم النفس، كلية التربية، جامعة دمشق*، ١٣ (١)، ١٣٦-١٥٥.

عفاف راضي اللحياني (٢٠١٢). أثر بعض طرق تقدير الدرجات للمفردات على ثبات وصدق درجات اختبار تحصيلي في الرياضيات ذي الاختيار من متعدد لدى طالبات الصف الأول الثانوي بمكة المكرمة. *دراسات عربية في التربية وعلم النفس، رابطة التربويين العرب*، ٢٢ (٢)، ٤٨٧-٥١٦

فريال محمد أبو عواد (٢٠١٨). استقصاء تقديرات معالم الفقرات والقدرة ودالة المعلومات لاختبار القدرات المعرفية باستخدام النموذج اللوجستي ثلاثي المعلمة. *دراسات نفسية وتربوية، جامعة قاصدي مرباح*،١١ (١)، ١-١٧.

ليندا كروكر، وجيمس الجينا (٢٠١٧). *مدخل الى نظرية القياس التقليدية والمعاصرة* (ترجمة: هند عبدالمجيد الحموري، زينات يوسف دعنا). عمان: دار الفكر للنشر والتوزيع.

محمد صيتان الصمادي (٢٠١٥). *أثر مخالفة قواعد صياغة فقرات الاختيار من متعدد على تقديرات معالمها ودالة معلومات الاختبار باستخدام النموذج ثلاثي المعلمة* ( رسالة دكتوراه غير منشورة). كلية التربية، جامعة اليرموك.

معين سلمان النصراوين، محمد وليد موسى البطش (٢٠١٨). مقارنة أربعة نماذج لمعالجة التخمين في الأسئلة الموضوعية/الاختيار من متعدد في إطار النموذج اللوجستي ثلاثي المعلمة وأثرها على دقة تقدير معلمة القدرة. *دراسات العلوم التربوية، الجامعة الأردنية، ٤٥* (٤)، ٣٣٢-٣٥٣.

نضال الشريفين، رانيا الصبح (٢٠١١). أثر بنية فقرات الاختيار من متعدد ومستوى القدرة لدى الأفراد على دقة التقديرات لمعالم الفقرات والأفراد وفق نظرية الاستجابة للفقرة. *مجلة جامعة أم القرى للعلوم التربوية والنفسية، ٣* (٢)، ٤٥ – ١١٠.

نضال كمال الشريفين (٢٠١٢). أثر طريقة تقدير معالم الفقرة وقدرات الأفراد على قيم معالم الفقرة، والخصائص السيكومترية للاختبار، في ضوء تغير حجم العينة. *المجلة التربوية، جامعة الكويت، ٢٦* (١٠٤)، ١٧٧-٢٣٨.

يوسف عبدالقادر أبوشندي، راشد سيف المحرزي، إيهاب محمد عمارة (٢٠١٨). دقة تقدير العلامات الحقيقية عند درجات مختلفة للارتباط الموضعي بين فقرات الاختبار فى توزيعات مختلفة للقدرة. *مجلة العلوم التربوية والنفسية، جامعة البحرين، ١٩* (٣)، ٤٦٥-٤٩١.

Adedoyin, O. (2010). Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Educational Science, 2*(2), 107-113.

Aiken, L. R. (1987).Testing With Multiple-Choice Items. *Journal of Research and Development in Education. 20* (4), 44-57.

Aiken, L. R. (2003). *Psychological Testing and Assessment* (11th ed). Boston: Pearson Education Group.

*Aiken, L. R. & Groth-Marnat, G. (2006). Psychological Testing and Assessment (12th Ed).* Boston, MA: Pearson Education Group.

Ainol, M. A. & Noor, L. A. (2006). Classical and Rasch Analyses of Dichotomously Scored Reading Comprehension Test Items. Malaysian Journal of ELT Research, 2(1), 1-20.

Albano, A. D., Christ, T. J. & Cai, L. (2018). Evaluating Equating in Progress Monitoring Measures Using Multilevel Modeling. *Measurement: Interdisciplinary Research and Perspectives, 16*(3), 168-180.

Angoff, W. H. (1987). Technical and Practical Issues in Equating. *Applied Psychological Measurement*, *11*(3), 291-300.

Anstasi, A. & Urbina, S. (2005). Psychological Testing (7th ed). New Jersey: Prentic-Hall.

Ayala, R. J. (2008). *The Theory and Practice of Item Response Theory :Methodology in the Social Sciences*. New York, NY: The Guilford Press.

Ayhan, S. (2015). Comparability of Scores from Cat and Paper and Pencil Implementations of Student Selection Examination to Higher Education, (Unpublished Master Dissertation). Bilkent University, Ankara.

Bechger, T., Maris, G., Verstralen, H. & Beguin, A. (2003). Using Classical Test Theory in Combination with Item Response Theory. *Applied Psychological Measurement, 27*(5), 319-334.

Bond, T. G. & Fox, C. M. (2015). *Applying The Rasch Model: Fundamental Measurement in the Human Sciences* (3th Ed). New York, NY: Routledge.

Breakall, J., Randles, C. & Tasker, R. (2019). Development and Use of a Multiple-Choice Item Writing Flaws Evaluation Instrument in the Context of General Chemistry. *Chemistry Education Research and Practice, 20*(2), 369-382.

Campbell, M. L. (2015). Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed to Discourage Guessing. *Journal of Chemical Education, 92*(7), 1194-1200.

Cappelleri, J. C., Jason, L.J. & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures, Clin Ther, 36(5), 648-662.

Chang, S. H., Lin, P. C. & Lin, Z. C. (2007). Measures of Partial Knowledge and Unexpected Responses in Multiple-Choice Tests. *Educational Technology & Society, 10*(4), 95-109.

Coggins, J. V., Kim, J. K. & Briggs, L. C. (2017). Comparison of IRT and CTT Using Secondary School Reading Comprehension Assessments. *Research in the Schools, 24*(1), 80-93.

Crehan, K. & Haladyna, T. M. (1991). The Validity of Two Item-Writing Rules. *The Journal OF Experimental Education*, 59(2), 183-192.

David, M. (2013). A Note on the Item Information Function of the Four-Parameter Logistic Model. *Applied Psychological Measurement, 37*(4), 304-315.

DeMars, C. (2010). *Item Response Theory: Understanding Statistics Measurement*. New York, NY: Oxford University Press.

Dimiter, M. D. (2016). An Approach to Scoring and Equating Tests with Binary Items: Piloting With Large-Scale Assessments. *Educational and Psychological Measurement, 76*(6), 954-975.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. (5th ED), New Jersey: Prentice-Hall, Englewood Cliffs.

Eleje, L. I., Onah, F. E. & Abanobi, C. C. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results. *European Journal of Educational and Social Sciences, 3* (1), 71 – 89.

Field, A. (2009). Discovering Statistics Using SPSS: Introducing Statistical Method (3rd ed.). Thousand Oaks, CA: Sage Publications.

Fraser, C. & McDonald, R. P. (1988). NOHARM: Least Squares Item Factor Analysis. Multivariate Behavior Research, 23, 267-269.

Georgiev, N. (2008). Item Analysis of C, D and E Series from Raven's Standard Progressive Matrices with Item Response Theory Two-Parameter Logistic Model. Europe's Journal of Psychology, 4(3), 1-17.

Gleason, J., Alley, A. & Baker, S. (2010). Effects of Item Writing Rules on The Reliability of Instruments to Measure The Mathematical Knowledge of Teachers. Journal of Mathematical Sciences & Mathematics Education, 5(2), 21-27.

Gorsuch, R. L. (1983). Factor Analysis ($2^{nd}$ ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Gregory, R. J. (2014). Psychological Testing: History, Principles and Applications($7^{th}$ ED). Boston: Person Education Group.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education, 15 (3), 309-334.

Hambleton, R. K. (2004). Theory, Methods, and Practices in Testing for The 21st Century. *Psicothema, 16*(4), 696-701.

Hambleton, R. K. & Jonse, R. W (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement, 12(3), 38-47.

Hambleton, R. K. & Jones, R. W. (1994). Item Parameter Estimation Errors and Their Influence on Test Information Function. Applied Measurement in Education, 7(3), 171-186.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston, Kluwer-Nijhoff Publishers.

Hambleton, R. K., Swaminthan, H. & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications.

Harris, D. J. & Crouse, J. D. (1993). A Study of Criteria Used In Equating. *Applied Measurement In Equating, 6*(3), 195-240.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. Applied Psychological Measurement, 9 (2), 139-164.

Huang, Y., Trevisan, M. & Storfer, A. (2007). The Impact of the "all-of-the-above" Option and Student Ability on Multiple Choice Tests. *International Journal for the Scholarship of Teaching and Learning, 1*(2), 1-13.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood IL: Dow Jones Irwin.

Hwang, D. (2002). Classical Test Theory and Item Response Theory: Analytical and Empirical Comparisons, *Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, 1-22. ED. 466 779.*

Inal, H. & Anil, D. (2018). Investigation of Group Invariance in Test Equating under Different Simulation Conditions. *Eurasian Journal of Educational Research*, 78, 67-86.

Jasper, F. (2010). Applied Dimensionality and Test Structure Assessment with The START-M Mathematics Test. The International Journal of Educational and Psychological Assessment, 6(1), 104-125.

Jinming, Z. (2012). The Impact of Variability of Item Parameter Estimators on Test Information Function. *Journal of Educational and Behavioral Statistics, 37*(6), 737-757.

Joo, S.-H., Lee, P. & Stark, S. (2018). Development of Information Functions and Indices for the GGUM-RANK Multidimensional Forced Choice IRT Model. *Journal of Educational Measurement, 55*(3), 357-372.

Kellere, R. R. (2007). *A Comparison of Item Response Theory True Score Equating and Item Response Theory-Based Local Equating* (Unpublished Doctoral Dissertation), University Of Massachusetts Amherst.

Kim, K. Y. & Lee, W. (2017). The Impact of Three Factors on the Recovery of Item Parameters for the Three-Parameter Logistic Model. *Applied Measurement in Education, 30*(3), 228-242.

Kim, S., Cohen, A. S. & Lin, Y. (2006). LDIP: A Computer Program for Local Dependence Indices for Polytomous Items. *Applied Psychological Measurement, 30*(6),509-510.

Kolen, M. J. & Brennan, R. L. (2014). *Test Equating, Scaling, And Linking: Methods And Practices* (3nd Ed). New York: Springer.

Kolen, M. J. &Whitney, D. R. (1982). Comparison Of Four Procedures For Equating The Tests Of General Educational Development. *Journal Of Educational Measurement, 9*(4). 279-293.

Lau. P. N., Lau, S. H., Hong, K. S. & Usop, H. (2011). Guessing, partial Knowledge, and Misconceptions in Multiple-Choice Tests. *Educational Technology& Society, 14*(4), 99-110.

Lesage. E, Valcke. M, & Sabbe (2013). Scoring Methods for Multiple Choice Assessment in Higher Education is it Still a Matter of Number Right Scoring or Negative Marking? *Studies in Educational Evaluation*, *39(3*), 188-193.

Lin, C. K. (2018). Effects of Removing Responses with Likely Random Guessing Under Rasch Measurement on a Multiple-Choice Language Proficiency Test. *Language Assessment Quarterly, 15*(4), 406-422.

Magis, D. & Raîche, G. (2012). On the Relationships Between Jeffreys Modal and Weighted Likelihood Estimation of Ability Under Logistic IRT Models. *Psychometrika. 77*(1), 163-169.

Mueller, D. & Schrock, T. (1982). Effects of Violating Three Multiple-Choice Item Construction Principles. *The Journal of Educational Research, 75* (5), 314-318.

Natarajan, V.(2009). *Basic Principles of IRT And Application to Practical Testing & Assessment*. MeritTrac Services (P) Ltd.

Ndalichako, J. & Rogers, W. T. (1997). Comparison of Finite Score Theory, Classical Test Theory and Item Response Theory in Scoring Multiple-choice Items. *Educational and Psychological Measurement*, 57(4), 580-589.

Nering, M. L. & Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models Nering*. New York, NY: Routledge /Taylor & Francis Group.

Ojerinde D. (2013). Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of The Comparability of Item Analysis Results. *A Guest Lecture Presented at The Institute of Education, University of Ibadan on 23rd May.*

Onder, I. (2007). An Investigation of Goodness of Model Data Fit Model Veri Uyumunun Araştirilmasi. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 32, 210-220.

Öztürk-Gübes, N. & Kelecioglu, H. (2016). The Impact of Test Dimensionality, Common-Item Set Format, and Scale Linking Methods on Mixed-Format Test Equating. *Educational Sciences: Theory and Practice, 16*(3), 715-734.

Pachai, M. V., DiBattista, D. & Kim, J. A. (2015). A Systematic Assessment of 'None of the Above' on Multiple Choice Tests in a First Year Psychology Classroom. *The Canadian Journal for the Scholarship of Teaching and Learning*, 6(3), 1-14.

Penfield, R. D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. *Educational Measurement: Issues and Practice, 33*(1), 36-48.

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, Norming and Equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., Pp.221-262). New York NY: Macmillan Publishing company.

Raykov, T., Marcoulides, G. A. (2016). On the Relationship between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement, 76*(2), 325-338.

Reise, S. P. & Revicki, D. A. (2015). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York, NY: Routledge.

Reise, S. P. & Waller, N. G. (2003). How Many IRT Parameters Does It Take To Model Psychopathology Items?. *Psychological Methods, 8*(2), 164-184.

Rodriguez, M. C. & Albano, A. D. (2017). *The College Instructor's Guide to Writing Test Items: Measuring student learning.* New York, NY: Routledge.

Slepkov, A. D. & Godfrey, A. T. (2019). Partial Credit in Answer-Until-Correct Multiple-Choice Tests Deployed in a Classroom Setting. *Applied Measurement in Education, 32*(2), 138-150.

Sočan, G. (2015). Empirical Option Weights for Multiple-Choice Items: Interactions with Item Properties and Testing Design. *Advances in Methodology & Statistics / Metodoloski zvezki, 12*(1/2), 25-43.

Stage, C. (2003). Classical Test Theory or Item Response Theory . The Swedish Experience, Umea University, (7), 1-30.

Steiger, J. H. (1980). Tests for Comparing Elements of A Correlation Matrix. *Psychological Bulletin, 87*(2), 245-251.

Tarrant, M., Knierim, A., Hayes, S. & Ware, J. (2006). The Frequency of Item Writing Flaws in Multiple-Choice Questions Used in High Stakes Nursing Assessments. *Nurse Education in Practice, 26*(8), 662-671.

Tay, L., Huang, Q. & Vermunt, J. K. (2016). Item Response Theory with Covariates (IRT-C): Assessing Item Recovery and Differential Item Functioning for the Three-Parameter Logistic Model. *Educational and Psychological Measurement, 76*(1), 22-42.

Thomas, M., Steven, M. & Michael, C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement In Education*, 15(3), 309-334.

Ueckert, S. (2018). Modeling Composite Assessment Data Using Item Response Theory. *Pharmacometrics & Systems Pharmacology, 7*(4), 205-218.

Van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational and Measurement*, 46(3), 247-272.

Van der Linden, W. J. (2010). Item Response Theory. *International Encyclopedia of Education*, 4, 81-88.

Van der Linden, W. J. (2016). *Handbook of Item Response Theory*. New York, NY: CRC Press /Taylor & Francis Group.

Vanderoost, J., Janssen, R., Eggermont, J., Callens, R. & De Laet, T. (2018). Elimination Testing with Adapted Scoring Reduces Guessing and Anxiety in Multiple-Choice Assessments, but does not Increase Grade Average in Comparison with Negative Marking. *PLoS ONE, 13*(10), 1-27.

Zhonghua, Z. (2010). *Comparison of Different Equating Methods and An Application to Link Testlet-Based Tests*. (Unpublished Doctoral Dissertation), The Chinese University of Hong Kong.

Zimmerman, D. W. & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*(5), 357-371.