مقارنة بين مداخل الكشف عن تحيز المفردات عبر الزمن لاختبار TIMSS للرياضيات في البيئة المصرية

د. محمود علي موسى

http://dx.doi.org/10.29009/ijres.2.4.11

مقارنة بين مداخل الكشف عن تحيز المفردات عبر الزمن لاختبار TIMSS للرياضيات في البيئة المصرية

د. محمود علي موسى

مدرس علم النفس التربوي، كلية التربية بالإسماعيلية، جامعة قناة السويس، مصر

mahmod567@yahoo.com

قدمت للنشر في ١/ ١/ ٢٠١٩م **قبلت للنشر في ١/ ٦/ ٢٠١٩م**

الملخص: هدفت الدراسة للمقارنة بين مداخل الكشف عن تحيز مفردات الاختبارات التحصيلية عبر الزمن للمفردات التي تقع ضمن مستوى القياس الفتري. واعتمدت الدراسة على المنهج الوصفي التحليلي والمنهج السببي المقارن في الاجابة عن تساؤلات الدراسة. واعتمدت الدراسة على بيانات الأرشيف لدورتي اختبار الاتجاهات الدولية لتعلم الرياضيات (TIMSS 2015)، (TIMSS 2007). واستخدمت الدراسة الدالة التمييزية المنتظمة في الكشف عن تحيز المفردات، وقارنت الدراسة بين نتائج طريقتي تحليل التغاير وتحليل الانحدار المتعدد بطريقة stepwise. وتوصلت الدراسة إلى وجود تحيز في المفردة Alg2 ببعد الجبر.

الكلمات الدلالية: اختبار TIMSS، الدالة التمييزية للمفردات، تحيز الاختبار.

http://dx.doi.org/10.29009/ijres.2.4.11

Cross-time Item bias approaches Comparison of TIMSS test among Egyptian environment

Mahmoud Ali Mousa Mehanna
Lecturer of Educational psychology, College of Education, Suez Canal university, Egypt.
mahmod567@yahoo.com

Received 1st January 2019 **Accepted 1st June 2019**

Abstract: The study aimed that comprise the cross-time item bias approaches of scale items' achievement tests. Analytical descriptive and comparative causal approaches had been used. TIMSS 2007, and TIMSS 2015 data archive had been used. The uniform differential item functioning methods used to test the item bias. The study comprises between ANCOVA and Multiple regression. The finding was the Alg2 item was biased in Algebra dimension.

Keywords: Item Bias, Differential Item Functioning, Test Fair.

http://dx.doi.org/10.29009/ijres.2.4.11

Summary

Introduction:

Psychological tests must be fair for all students, but some of it was biased in gender, nationality, and culture. Research can use the statistical approach to test that there was a biased item in which groups make a differences and had a variance in test items.

Differential item functioning (DIF) is the best method to test the bias in test items comprise the differences between ability levels for unequaled response for each true item.

Differential item functioning (DIF) is a statistical item characteristic which shows the range to which the item could be measuring different abilities for individuals in separate subgroups. Mean item scores for subgroups having the same overall score on the test which determine whether item is measuring in essentially the same way for all subgroups comparing. The presence of DIF requires review and judgment, and it does not necessarily indicate the presence bias.

Differential item functioning analysis provides an unexpected behavior indices of test items. The item doesn't display DIF if people from different groups have a different probability to determine a certain response; it displays DIF if and only if people from different groups with the same underlying true ability have a different probability of giving a certain response.

Problems and research motivation:

This study explained the item bias using three systematic statistical indices such as Type I error, Number of item biased in each one, effect sizes convergence. First study could be used statistical control procedures such as data distribution and outliers in variables.

These procedures may lead to biased explanations that previous studies did not address in the research of the function of discriminating vocabulary and then identify the consistency of the approaches adopted by the researcher in the study to determine the discriminatory function of TIMSS 2007 and TIMSS 2015 items.

http://dx.doi.org/10.29009/ijres.2.4.11

Objectives:

The study aimed that comprise the cross-time item bias approaches of scale items' achievement tests. Then determine of item bias in TIMSS 2007 and TIMSS 2015. ANCOVA and Multiple regression w\had been selected to determine the item bias and differentiate the item characteristics using some indices such as Type I error, Number of Biased items in each one, effect sizes convergences.

Hypotheses:

- Differential item functioning was sensitivity of data distribution of continuous variables and items and Outliers data.
- Item Bias approaches had been differentiated with respect to Type I error, Number of items biased in each one, effect sizes convergence indices.

Method and producers

Method:

- Research Design: Analytical descriptive and comparative causal approaches had been used.
- Participants: The study was based on the mathematics test of 13,164 students in the second grade of TIMSS 2007 and the number of 7095 students and students of TIMSS 2015. The average age is 14.34 years with a standard deviation of 0.89 years.
- Instruments: TIMSS 2007, and TIMSS 2015 data archive had been used.
- statistical techniques: The uniform differential item functioning methods used to test the item bias. The study comprises between ANCOVA and Multiple regression. The finding was the Alg2 item was biased in Algebra dimension.

http://dx.doi.org/10.29009/ijres.2.4.11

Procedures:

- Obtain the results of the TIMSS 2007 and TIMSS 2015 tests and prepare them for statistical analysis in the SPSS program.
- Verify the distribution of the two sets of data, linearity, and check for the existence of outliers in the variable data belonging to both groups.
- Choose the grouping variable (TIMSS 2007 and TIMSS 2015) as a independent and items which calculate its DIF as a dependent variable.
- Create a Rest score variable and consider the covariate variable in the ANCOVA method, and independent in the regression analysis method. The rest score is calculated by the difference between the degree of the total dimension of the content and the degree of the item to be calculated.
- Check for differences between the two groups using independent sample tests to verify differences between the two groups on the degree of content exclusion. To determine the type of DIF procedures (regular, irregular, mixed).
- The differentiation between the methods of calculating the discriminating function of the individual by means of the following indicators: (Type I error, Number of items biased in each one, and effect sizes convergence).

Results:

The results of the analysis found that the dimensions of the content of the two sets of analysis were statistically significant, which means that the items weren't non-normal distributions.

Reflecting on the TIMSS 2007 boxplot, it is clear that there are positive and negative outliers for the dimension of algebra, then data and inference dimensions, which affect the data distributions and may give misleading indicators.

http://dx.doi.org/10.29009/ijres.2.4.11

The results of the study showed that the item agreement Alg3, Alg4, and Alg5 in algebra were biased to the reference group TIMSS 2015. This means that the items may be repeated in presence within the test in the two sets.

Thus, the magnitude of the effect size indicated that a weak effect on the Eta square and the multiple correlation coefficient indicators.

http://dx.doi.org/10.29009/ijres.2.4.11

مقدمة

يجب أن تكون الاختبارات النفسية عادلة لجميع المتقدمين، إذ لا ينبغي لا يجب أن يكون الاختبارات متحيزة لخصائص المقدمين سواء لجنس أو عرق أو ثقافة معينة. وهذا يحتم على الباحث استخدام منهجاً احصائياً للتحليل يدرس به ما إذا كانت المفردات تعمل بشكل تمييزي بين المجموعات التي طبق عليها الاختبار واكتشاف مصادر التباين في مفردات الاختبار، ويعد أسلوب الدالة التمييزية للمفردة differential item functioning (DIF) أحد الأساليب البارزة في حساب الأداء التمييزي للمفردة DIF عندما يكون للمجموعتين موضع المقارنة نفس المستويات من القدرة مع عدم تساوي الاستجابة على كل مفردة بشكل صحيح. وعليه فلا تتمتع مجموعة واحدة بفرصة متساوية في الحصول على مفردة بالرغم من أن أعضائها لديهم مستويات قدرة مماثلة للمجموعة الأخرى.

وتُعد الدالة التمييزية أحد سُبل تطوير الاختبارات النفسية في ضوء الثقافات المتباينة، ودراسة مدى قابليتها للتطبيق عبر مجموعات متعددة. فكلما ابتعدت الدالة التمييزية عن البناء الكلي للمقياس أو الاختبار دل ذلك على عدم تحيز المفردة. ومتى صيغت المفردة بصورة أكثر صرامة في قياس سمة عامة تحتم على الباحث التأكد من خلو الاختبار من التحيز، ومن ثم يصبح تفسير النتائج المتحصل عليها في ضوء بناء المقياس أكثر منطقية (Karami, 2012).

ويعاني استخدام الدالة التمييزية للمفردة من بعض المحددات تتمثل في حجم العينات الكبيرة الذي قد يسبب خطأ من النوع الأول بمفردة غير متحيزة عند استخدام اختبارات الدلالة الاحصائية (Duncan, 2007). كما أن الدالة التمييزية للمفردة تتأثر بنوع القدرة المقاسة وتوزيع بيانات العينة فالتناقض في توزيعات القدرة عبر المجموعات المعيارية يخلق عدم استقرار نتائج DIF (Sweeney, 1996). كما تتأثر درجات الاختبار بمصادر التباين بين المجموعات في القدرة المقاسة، ومن ثم ينتج عن القياس بعد استبعاد المفردات ذات التحيز مؤشرات صادقة وثابتة للسمة المقاسة (Duncan, 2007).

http://dx.doi.org/10.29009/ijres.2.4.11

وتعتمد فكرة الدراسة على حساب الدالة التمييزية لمفردات اختبار TIMSS في الرياضيات لطلاب الصف الثاني الاعدادي. وهذا الاختبار يعقد دورياً كل أربعة سنوات بدءاً من عام ١٩٩٥، ومبررات هذا أن الدورة الحالية تطبق على الصف الرابع وعقب انتهاء الدورة الحالية يصبح الطلاب في الدورة القادمة بالصف الثاني الاعدادي. وهنا ركزت الدراسة على نتائج الاختبار التحصيلي لدورة ٢٠٠٧ ودورة ٢٠١٥ إذ لم تطبق اختبارات دورة ٢٠١١ بسبب قيام الثورة ومن ثم تصبح عينتي الدورتين مختلفين تماماً. وهذا يعد مؤشراً لنمو تعلم الرياضيات بمرور الزمن في ضوء برنامج PISA. كما اهتمت الدراسة بتوظيف الأساليب الاحصائية توظيفاً جديداً للتحقق من الدالة التمييزية للمفردات في كلا الصورتين مع استخدام بعض المحددات الاحصائية مثل توزيع بيانات المتغيرات الداخلة للتحليل واستبعاد القيم المتطرفة والضبط الاحصائي لمتغير الأداء المتحيز (المتغير المصاحب) والتحقق من الخطأ من النوع الأول وعدد المفردات المتحيزة وتقارب حجم الأثر كمحددات للحكم على تمييز المفردات في كلا الصورتين.

الاتجاهات الدولية لدراسة العلوم والرياضيات TIMSS

يتطلب تلبية متطلبات الأمم والشعوب المستقبلية في الاقتصاد العالمي تطوير رأس المال البشري في مجالات العلوم والتكنولوجيا والهندسة والرياضيات STEM. وقد أدت العولمة لدراسة تعلم العلوم والرياضيات في جميع أنحاء العالم بدلاً من دراسة التحصيل الدراسي للطالب داخل دولة واحدة؛ بل دراسة المتغيرات التعليمية عبر نطاق أوسع من الدول متعددة اللغات والثقافات. وقدمت بعض الدراسات التقييمات واسعة عبر الدول (ILSA) international large-scale assessments ومنها على سبيل المثل برنامج التقييم للطلاب الدوليين Programme for International Student Assessment (PISA) والاتجاهات الدولية لدراسة العلوم والرياضيات Trends in International Mathematics and Science Study (TIMSS) بالمزيد من مقارنات بين أداء الطلاب الدوليين للعلوم والرياضيات. واستناداً إلى تلك النتائج حاول صناع القرار بالعديد من البلدان إصلاح التعليم باستخدام المعلومات الناتجة من تلك التقييمات كمرجعية في اتخاذ القرار (Liou & Hung, 2015).

http://dx.doi.org/10.29009/ijres.2.4.11

محتوى وأبعاد اختبار الاتجاهات الدولية لدراسة الرياضيات TIMSS

اهتم TIMSS بقياس الأداء لطلاب الصفوف الرابع والثامن. وقد صمم الاختبار ليتماشى مع مناهج العلوم والرياضيات في أنظمة التعليم للدول المشاركة في التقييم، وذلك للوصول لمعلومات قيمة حول أداء الطلاب بجميع دول العالم (Stephens et al., 2016). وتتم إدارة TIMSS كل أربع سنوات؛ فطلاب الصف الرابع في الدورة الحالية تنتقل إلى الصف الثامن الطلاب في الدورة المقبلة. ويعتبر TIMSS 2015 هو الاختبار السادس في سلسلة التقييمات منذ بدأه منذ عام ١٩٩٥ (Martin, Mullis & Foy, 2015). ويعتمد جوهر الاختبار على حل المشكلات في مادة الرياضيات. فالرياضيات علم حياتي يبرز في المجالات اليومية المختلفة كالعد، وإدارة الأموال، وبعض المجالات المهنية التي تقوم أسس رياضية مثل الأعمال مصرفية والتجارة والطب والهندسة والمجالات البرمجية وتكنولوجيا المعلومات. يتم تنظيم كل إطار من أطر التقييم الثلاثة لبرنامج TIMSS 2015 حول بعدين (Grønmo, Lindquist, Arora & Mullis, 2015):

١. بعد المحتوى Content ويشير إلى تحديد الموضوعات المقرر تقييمها. ويتضمن محتوى الاختبار بالصف الثامن مستويات الاعداد Number ويمثل ٣٠٪ من محتوى الاختبار، والجبر Algebra وتمثل ٣٠٪، والهندسة Geometry وتمثل ٢٠٪، والبيانات والاحتمالات Data and chance وتمثل ٢٠٪. ويرى الباحث أنه في بعض المحتوى فإن التحيز أو التمايز في المفردات قد يحدث نتيجة طبيعة المناهج في سياقها العملي المقدم نتيجة لتطوير الاختبار عبر الزمن مما يتطلب اختلاف التجهيز المعرفي أو الأساليب المعرفية المختلفة حسب طبيعة كل بيئة من بيئات القطر الواحد خصوصاً في دراسات عبر الزمن.

٢. البعد المعرفي Cognitive ويشير إلى عمليات التفكير التي يتعين على البرنامج تقييمها. ويتضمن الاختبار بالصف الثامن المستويات المعرفية التالية: مستوى

http://dx.doi.org/10.29009/ijres.2.4.11

المعرفة Knowing ويمثل ٣٥٪ من محتوى الاختبار، والتطبيق Appling ٤٠٪، ومستوى التبرير وحل المشكلات Reasoning ٢٥٪.

ويتكون الاختبار من مجموعة متنوعة من المشكلات الرياضية، وحل المعادلات الخطية، والنسب المئوية، والتعميمات، وتبرير استنتاجاتهم. وطبيعة الأشكال الهندسية والاحتمالات (Stephens, Landeros, Perkins & Tang, 2016). ويدار اختبار TIMSS كل أربعة سنوات منذ عام ١٩٩٥ . ويتم استعراض تحديثات الأطر المرجعية للأداء في كل دورة ليعكس التطورات في المناهج، مع التأكيد على محددات المقارنة وإجراءات المعاينة والتقييم (Stephens et al., 2016).

مستويات المنهج في اختبار الاتجاهات الدولية لدراسة العلوم والرياضيات TIMSS:

يهدف TIMSS لمساعدة الدول في تقييم تعلم العلوم والرياضيات والتعلم عبر الوقت ومستويات السنة الدراسية. ويركز TIMSS على المنهج الدراسي، وقد حددت ثلاثة مستويات من المنهج في الدراسات السابقة وهي (Thomson, Wernert, O'Grady & Rodrigues, 2016):

- المنهج المقصود intended curriculum: ويشير إلى مناسبة المنهج لمستوى دولة معينة أو مستوى النظام التعليمي المراد اختبار طلابه. ويبحث عن إجابة لتساؤل "ما يتوقع أن يتعلمه الطلاب في العلوم والرياضيات بجميع دول العالم؟" و "كيف تختلف الدول في أهدافها المقصودة؟" و "ما هي خصائص نظم التعليم والمدارس والطلاب المؤثرة في تطوير هذه الأهداف؟" و "كيف ينبغي تنظيم مثيرات التعلم لتسهيل هذا التعلم؟".
- المنهج التطبيقي implemented curriculum: ويشير للصورة التي يطرح بها معلمي الرياضيات المنهج الدراسي وينتهجها، ومراعاة الممارسات التعليمية للدول وتحقيق التقارب بين العوامل المؤثرة فيها.

http://dx.doi.org/10.29009/ijres.2.4.11

- المنهج المدروس attained curriculum: ويشير إلى المنهج الذي يتعلمه الطلاب بما يتوافق مع مواقفهم وإنجازاتهم. ودراسة المفاهيم والعمليات والمواقف التي تعلمها الطلاب في الرياضيات، وكيفية تأثيرها على تحصيل وأداء الطلاب.

دلالات التقدير عبر الثقافات والزمن TIMSS 2015 reliability scoring:

يعد الثبات أحد مؤشرات الحكم على جودة TIMSS، وقد حسب ثبات مفردات TIMSS 2015 لكل دولة على حدة، وعبر الدول المختلفة، والثبات عبر الزمن Trend reliability. وقد تلخص إجراء تحديد الثبات لكل دولة على حدة اشتقاق عينة عشوائية بسيطة وحساب قيم ثبات بناء المفردات. أما الثبات عبر الزمن لكل دولة فقد حسب عبر مرات التطبيق التي أجريت في الدورة الحالية مع الدورة السابقة باستخدام برنامج IEA Coding Expert (Johansone, 2015). ويرى ساشي وهاغ (Sachse & Haag, 2017) أن الأخطاء المعيارية للقياس لدراسات التقويم الدولية PISA وبالأخص اختبار TIMSS للرياضيات كانت متحيزة عبر الثقافات والدول المختلفة، بينما في البيئة الواحدة كانت التحيزات طفيفة في ضوء الدالة التمييزية.

الدالة التمييزية للمفردات Differential item functioning (DIF):

الدالة التمييزية هي أسلوب احصائي مرادف للتحيز الإحصائي حيث يوجد بالاختبار أو المقياس مفردة واحدة أو أكثر أقل أو مبالغ في تقديرها. وتهدف DIF إلى تحديد المفردات التي تعمل بشكل مختلف عبر اللغات المختلفة. وتختلف طرق تقييم DIF باختلاف طبيعة المفردة الاختبارية فقد يستخدم أسلوب ANCOVA إذا كانت المفردة تتبع مستوى القياس الفتري على الأقل، ويستطع ANCOVA تقيم الفروق بين المجموعات بعد ضبط المتغيرات التي قد تشوه تفسير الاختلافات (Harter & Agrawal, 2011). وتحدث الدالة التمييزية للمفردة إذا اختلف احتمال الاستجابة الصحيحة بين الأشخاص الذين لديهم نفس القيمة على نفس السمة في المجموعات الفرعية، فعلى سبيل المثال إذا كانت صعوبة أحد المفردات تتأثر بمجموعة فرعية عرقية أو مرحلة عمرية أو جنسيات مختلفة (Berger & Tutz, 2016). ويرى الباحث أن هذا المفهوم يعد مفهوماً متحيزاً خصوصاً إذا كانت السمات معرفية

http://dx.doi.org/10.29009/ijres.2.4.11

فهي تتأثر بتجهيز المعلومات وإشكاليات المكونات المعرفية وليس العرق أو الثقافة وإنما تتأثر بالمرحلة العمرية أكثر من أي شيء أخر. ويوفر تحليل DIF دلالة على سلوك غير متوقع لمفردة من مفردات الاختبار عبر مجموعات مختلفة (Gesicki, 2015).

وعندما تتوفر دالة التمييز DIF في أحد المفردات عندئذ يجب التحقق من مصدر الاختلاف في مفردات نفس الأبعاد والعوامل الداخلية للمقياس. ويتم الحكم على المفردة بأنها متحيزة في ضوء DIF إذا ارتبط مصدر التباين بالمجموعات التي عملت بشكل تمييزي على مفردة. وأكد دانكن (Duncan, 2007) على ضرورة استخدام الدالة التمييزية كإجراء مكمل كشرط للتحقق من البنية العاملية للمقاييس النفسية لتقييم مدى تميز القياسات للقدرات الحقيقية بين الممتحنين بطريقة غير متحيزة. ويرى الباحث أن المفردة في المقاييس النفسية تختلف من مكان لأخر ومن ثقافة لأخري، بينما المفردة في العلوم والرياضيات تكون بنيتها أقرب للمسلمات وليس للمفاهيم الانطباعية والشخصية.

ويحدث DIF عندما يكون لدى مجموعتين من نفس مستويات القدرة فرص مختلفة في الاستجابة على مفردة. ويطلق على المجموعة التي تتفوق في السمة المحتملة من خلال الاختبار بالمجموعة المرجعية (Karami, 2012). وتنقسم الدوال التمييزية للمفردات DIF على النحو التالي:

١. الدالة التمييزية المنتظمة DIF uniform وفيها يكون أداء أحد المجموعات أفضل من المجموعة الأخرى على جميع مستويات القدرة، وهذا يعني أن جميع أفراد المجموعة تقريباً يتفوقون على جميع أفراد المجموعة الأخرى الذين يتمتعون بنفس مستويات القدرة (Karami, 2012). وتوصل وتمور (Whitmore, 1996) لانخفاض التحيز بزيادة حجم العينة في أسلوب تحليل التغاير عندما أجرى دراسته على بيانات المحاكاة.

٢. الدالة التمييزية غير المنتظمة DIF Non-uniform يتم تفضيل أعضاء مجموعة واحدة إلى أحد أبعاد مقياس وهذا يعني وجود تفاعل بين المتغير التصنيفي

http://dx.doi.org/10.29009/ijres.2.4.11

للمجموعات ومستوى القدرة (Karami, 2012). ويزيد التمييز غير المنتظم باستخدام تحليل التغاير (Whitmore, 1996).

٣. الدالة التمييزية المركبة combination DIF وتشير إلى أن استجابات المجموعات على بعض أبعاد المقياس غير متكافئة، بينما تكون متكافئة على الأبعاد الأخرى. ويفضل استخدام أسلوب تحليل التغاير في هذه الحالة بزيادة حجم العينة وطول الاختبار إذ يقل أخطاء القرار من النوع الأول (Whitmore, 1996).

مداخل حساب الدالة التمييزية لمفردات الاختبارات النفسية:

أ. مدخل تحليل التغاير:

يعد أسلوب تحليل التغاير ANCOVA أحد الأساليب غير الشائعة في تحديد الدالة التمييزية للمفردات، إلا أن Sireci et al. (2003) استخدمه لأول مرة في دراسة مسحية للمفاضلة بين أداء مفردات مقياس عبر المجوعات المختلفة. وقد طبق أسلوب ANCOVA جنباً إلى جنب مع الانحدار اللوجستي logistic regression لدراسة الفروق بين استجابات بعض العاملين على استطلاع رأي عبر الثقافات. ويستخدم ANCOVA لدراسة الفروق بين المجموعات على نتائج القياس عقب ضبط تأثير بعض المتغيرات. وتوصلت النتائج إلى وجود تطابق النتائج بين ANCOVA والانحدار اللوجستي في التمييز بين المجموعات وقد بلغ حجم التأثير ٠,٩٦. وأكد Sireci et al. (2006) أن مدخل ANCOVA يعد الأنسب في تقييم الدالة التمييزية للمفردات في استجابات المقاييس المصممة في ضوء طريقة ليكرت.

وتؤدي الارتباطات بين المتغيرات المستقلة إلى تعقيد حسابات ANCOVA، فإذا تأثر المتغير المصاحب بالمعالجة في التجارب النفسية فإن التغاير يصبح أسلوباً غير مناسباً لتحقيق الضبط الاحصائي. فعلى سبيل المثال رغبة الباحث في تحديد ما إذا كان التعلم النشط يحسن من الاحتفاظ بالبيانات لدى الطلاب أكثر من المحاضرة التقليدية مع ضبط متغير الدافعية. فإذا كان التعلم النشط يزيد من دافعية المتعلم؛ فإن ANCOVA غير قادراً على ضبط الدافعية؛ إذ لا توجد طريقة لتجزيء

http://dx.doi.org/10.29009/ijres.2.4.11

التباين في الاحتفاظ بالمعلومات بحيث يتشارك دافعية التعلم والاستراتيجية التعليمية في تباين كل منهما. وفي هذه الحالة سيتم ضبط متغير الدافعية عبر مستويين تأثير المعالجة وعلى تأثير المعالجة نفسها على المتغير التابع (Karpen, 2017).

وتتأثر الدالة التمييزية للمفردة بإجراءات الضبط الإحصائي خصوصاً عندما ينتهك شرط عشوائية اختيار العينة، إذ من الصعب تحديد ما إذا كان اختلاف عما قبل المعالجة ناتجًا عن خطأ عشوائي أو عن فرق حقيقي في المجموعة. وإذا كان هناك فرق حقيقي في المجموعة فإن تحليل التغاير سوف يتحكم في كل من تأثير عضوية الفرد في المجموعة group membership وتأثير المتغير المصاحب، وبالتالي تحيز تقدير تأثير عضوية المجموعة.

ويستخدم الفرق بين الدرجة الكلية للبعد ودرجة استجابة الطلاب على المفردة كمتغير مصاحب لعزل أثر المفردة في التأثير على المتغير التابع. ويتسم ANCOVA بالقوة الاحصائية لأنه يقلل تباينات الخطأ، ويزيل الجزء القابل للتنبؤ عند اختلاف تباينات الخطأ، كما أنه يعادل احصائياً مجموعات المقارنة، ويقلل الخطأ التجريبي إذا أمكن التنبؤ بجزء من تباين الخطأ المرتبط بالمتغير التابع وذلك بالمعرفة السابقة بالمتغير المصاحب (Abah, 2018). ويستخدم مؤشر مربع ايتا لتقييم DIF بحيث يكون التمييز صغيراً أقل من ٠,٠٣٥، أما المتوسطة تتراوح قيمة المؤشر بين ٠,٠٣٥ إلى ٠,٠٧٠ بينما كبيرة تزيد قيمة المؤشر فيها عن ٠,٠٧ (Harter & Agrawal, 2011).

ب. مدخل تحليل الانحدار:

تزداد كفاءة هذه الطريقة في التمييز بين مجموعات الدراسة بزيادة حجم العينة، ويعتبر طول اختبار مساوياً ٢٠ مفردة كحد أدنى كافياً ومرضياً للكشف عن الدالة التمييزية (Whitmore, 1996). وتعد طريقة الانحدار أدق من تحليل التباين والتغاير في حالة الدالة التمييزية المنتظمة وغير المنتظمة (Whitmore, 1996). وللحكم على تحيز المفردة في ضوء DIF يجب أن تكون قيمة مربع معامل الانحدار المتعدد هي $R2 \leq 0.130$ (Zumbo, 1999). بينما حدد جيسكي (Gesicki, 2015) نقاط القطع لحجم التأثير لمقياس R2 بأنها تتراوح بين ٠,١٣ حتى ٠,٢٦ حتى يتم الحكم على حجم التأثير بأنه مقبول.

http://dx.doi.org/10.29009/ijres.2.4.11

اختبارات الدلالة الاحصائية للدالة التمييزية Tests Significance for DIF:

يعتمد اختبار الدلالة الإحصائية للدالة التمييزية DIF على نمذجة الأداء التفاضلي DIF كبناء هرمي طبيعي لإدخال المتغيرات في النموذج. منها: (١) الإدخال الأول لمتغير الحالة conditioning variable وهو الدرجة الكلية للبعد. و (٢) متغير المجموعة group variable، و (٣) إدخال متغير التفاعل في الحساب لمعادلة الانحدار (Zumbo, 1999).

وأكد زويك وثاير مازيو (Zwick, Thayer & Mazzeo, 1997) أن اختلاف التوزيعات البيانات لمجموعات يؤدي إلى ميل المفردات للتمييز بين المجموعات بشكل كبير خصوصاً في المفردات ذات الدرجات المتصلة Polytomous. كما أسفرت نتائج الدراسة عن أن المجموعات ذات التوزيعات المتماثلة للمجموعات غير قابلة للتميز بين أداء الأفراد على المفردات. وأجرى فيتش (Finch, 2016) دراسة باستخدام بيانات المحاكاة للمقارنة بين طرق حساب الدالة التمييزية للمفردات مع مجموعات متعددة (أكثر من مجموعتين) وقد توصلت الدراسة إلى أن الانحدار المتعدد وطريقة mantel- Haenszel تفوقت في خفض الخطأ من النوع الأول وزيادة القوة الاحصائية.

مشكلة الدراسة والدراسات سابقة:

استخدم سوامنسان وروجرز (Swaminathan & Rogers, 1990) الانحدار اللوجستي لوصف الأداء التفاضلي DIF بين مجموعتين. وللتمييز بين التماثل وعدم تماثل والمجموعتين تم اختبار فرضية عدم وجود تفاضل بين المفردات في التمييز بين المجموعتين باستخدام دراسات المحاكاة وقد تبين أن الانحدار اللوجستي أكثر قوة من اختبار Mantel- Haenszel للكشف عن الأداء التفاضلي للمفردات.

اعتمدت دراسة انجلهارد وهانشي وراوتليدج (Engelhard Hansche & Rutledge, 1990) على استخدام معيار المطابقة الخارجية وذلك عن طريق استخدام ٤٢ محكم لتقدير الدرجات لكل من المعلمين البيض والسود بناء على بعض المعايير التجريبية والتحكيمية التي تتعلق بشهادات التخرج للمعلم والطفولة المبكرة والادارة والاشراف والطفولة المتوسطة وقد تم الاتفاق بين ٤٠ محكم على النتائج التي تقديرها لكل من البيض والسود من المعلمين.

http://dx.doi.org/10.29009/ijres.2.4.11

استخدمت دراسة وانغ ولان (Wang & Lane, 1996) الدالة التمييزية لمعرفة مدى فاعلية المفردات في تقييم أداء الرياضيات بشكل مختلف فيما يتعلق بنوع الجنس لعينة بلغت ١٧٨٢ من تلاميذ الصف السادس والسابع وتساوت عدد الجنسين بالعينة، وجرت محاولة لتحديد العوامل (المحتوى، العمليات المعرفية، الاختلافات في توزيعات القدرة) التي قد تكون ذات صلة بـ DIF. وصممت الدراسة أداة للتقويم مهارات التفكير الرياضي تتكون من ٣٣ مفردة متصلة الدرجات موزعة على أربعة أبعاد. وتوصلت النتائج إلى اتساق نتائج لعدد ٣١ مفردة من المفردات بينما ميزت مفردتان بين جنس الطلاب.

واستخدم هاملتون (Hamilton, 1999) الدالة التمييزية للمقارنة بين الجنسين على نتائج اختبار العلوم الذي عقدته الجمعية الوطنية للتعليم العام عام ١٩٨٨ والذي يتم اختباره بصفة دورية كل عام. وتوصلت الدراسة إلى وجود مفردة واحدة ذكورية لها معامل تأثير عالي أسهمت في الاختلاف بين الجنسين على الدرجة الكلية. وكانت النتائج مماثلة لتلك التي حصل عليها عام ١٩٨٨.

بينما تحقق كول وكاوتشي وميلر وبركمان (Cole, Kawachi, Maller & Berkman, 2000) من تحيز مفردات مقياس الاكتئاب خلال متغيرات المرحلة العمرية والجنس والعرق لجميع مفردات المقياس لمؤسسة نيو هافن لعلاج الأوبئة للمسنين وكبار السن، وقد اعتمدت الدراسة على حساب مؤشر متوسط الفروق بين المجموعتين وحدود الثقة عند مستوى ثقة ٩٥٪. وقد كانت مفردات مثل "الناس غير وديين، والناس لا يحبونني" متحيزة لصالح السود وكانت فترات الثقة بها كبيرة. وقد كانت المفردة المتعلقة بالبكاء متحيزة لصالح النساء.

وقارنت دراسة ستون وكوك وليتوسيوس وكلاين (Stone, Cook, Laitusis & Cline, 2010) بين الطلاب المكفوفين والطلاب ضعاف البصر من طلاب الصف الثامن في اختبار تقويم فنون اللغة الانجليزية القائم على تقييم الحالة وذلك باستخدام طريقة Mantel- Haenszel والتي أسفرت عن وجود تمييز في أداء المجموعتين على بعد البلاغة الشعرية لصالح ضعاف الابصار الذين اعتمدوا على التخيل أكثر من الرؤية المباشرة.

http://dx.doi.org/10.29009/ijres.2.4.11

وقارن ليو (Liu, 2011) الأداء التمييزي بين موضوعات ثقافية وموضوعات العلوم الطبيعية في القطع القرائية لاختبار الفهم القرائي وهو احد الاختبارات الفرعية لاختبار TOEFL iBT وقد توصلت النتائج إلى وجود معاملات تمييزية ضئيلة أو منعدمة على معظم المفردات في الاختبار.

استخدم فينجولد (Feingold, 2013) أسلوب تحليل الانحدار الخطي لحساب حجم الأثر لنتائج المقارنة بين المجموعات في الدراسات المستعرضة والدراسات الطولية وذلك مستخدماً معامل الانحدار والانحراف المعياري للمتغير التابع ومدة الدراسة.

وحلل كارنوي وغافنسون وايفانوفا (Carnoy, Khavenson & Ivanova, 2015) نتائج TIMSS لبرنامج نتائج تحصيل الطلاب الدوليين PISA لدولة روسيا وبعض دول الجوار مثل لاتيفيا وإستونيا، حيث قدم الاختباران معلومات متناقضة حول الأداء النسبي للطلاب. وقد اسفرت النتائج عن تحقيق الطلاب الروس أداءً جيدًا نسبيًا في اختبار TIMSS للرياضيات ولكن بشكلٍ ضعيف نسبيًا في اختبار PISA.

وقارن الينا-اوليفري ولاولس وروبين وبردجمان (Elena Oliveri, Lawless, Robin & Bridgeman, 2018) بين نتائج اختبارات القبول في الرياضيات بين الطلاب من خلفيات غير أمريكية والطلاب الامريكيون وقارن الباحث بين نتائج الطلاب على المفردات باستخدام باختبار Mantel-Haenszel ونظرية الاستجابة للمفردة وقد اتفقت النتائج على أساس اتساق النتائج لمجموعتين في مفردات الاستدلال الرياضي التي تحتوي على الرسوم والجداول لصالح الامريكيين. بينما في اختبار المنطق الرياضي وجد أداء تمييزي خصوصاً في الاستعانة في الاسئلة ببعض المراجع الجغرافية وبعض المسميات المستخدمة في الاختبار بالنسبة لغير الأمريكيين.

قام انعابي ودويدين (Innabi & Dodeen, 2018) بدراسة على دراسة الاتجاهات الدولية لتعلم العلوم والرياضيات لعام ٢٠١٥ لدراسة الدالة التمييزية بين الذكور والاناث وأثبتت الدراسة ان الاناث يتفوقن على الذكور في الصف الثامن وقد استخدمت الدراسة أسلوب Mantel- Haenszel.

http://dx.doi.org/10.29009/ijres.2.4.11

وأظهرت النتائج أن الأولاد أكثر عرضة من البنات للإجابة الصحيحة على المشكلات الرياضية الأكثر صعوبة وغير المألوفة المرتبطة بالحياة، وعلى النقيض كانت الفتيات أكثر عرضة من الفتيان للإجابة بشكل صحيح على المشكلات المألوفة والأقل صعوبة عديمة الصلة بالحياة.

هدفت دراسة Akcan & Kabasakal (2019) لتحديد الدالة التمييزية لمفردات اختبار اللغة الانجليزية لمرحلة البكالوريوس خلال الجنس ونوع المدرسة باستخدام اختبار مانتل هانزل واستخدام أسلوب Multiple Indicator and Multiple Causes (MIMIC). وقد تألفت عينة الدراسة من ٥٩٨١٨ طالب أجري الاختبار عليهم عام ٢٠١٦ وأسفرت نتائج التحليل للمفردات ٦٠ إلى وجود تحيز للمفردات للذكور في بعد الترجمة التحريرية. أما في ضوء تحليلات التحيز لنوع المدرسة فهناك ٩ مفردات في بعد المعرفة اللغوية والنحوية أبرزت التحيز، و ٦ مفردات في بعد الفهم القرائي.

وتحاول الدراسة إعادة توظيف الأساليب الاحصائية التقليدية (تحليل التغاير، وتحليل الانحدار) في الكشف عن التمييز والتحيز بين مفردات عبر مجموعتي الدراسة (TIMSS 2015)، (TIMSS 2007). كما أن الدراسة تختلف في طبيعتها حيث أن الدراسات السابقة كانت تهتم باختلاف الأداء بين الأساليب الاحصائية المتنوعة فحسب ولكن الدراسة الحالية تفسر تلك النتائج في ضوء ثلاثة مؤشرات احصائية منهجية هي الخطأ في النوع الأول الراجع لطبيعة البيانات أو انتقاء العينات وطرق المعاينة ومؤشر عدد المفردات المتحيزة وتقارب حجم التأثير بين تلك المداخل. وذلك بعد التحقق من إجراءات الضبط الاحصائي لطبيعة توزيع بيانات تلك المفردات المتصلة واستبعاد بعض الحالات التي تحتوي على قيم متطرفة Outliers والتي قد تؤدي إلى تفسيرات متحيزة والتي لم تتطرق الدراسات السابقة إلى عرضها في بحوث الدالة التمييزية للمفردات ومن ثم التعرف على اتساق المداخل التي تبناها الباحث في الدراسة لتحديد الدالة التمييزية لمفردات اختبار (TIMSS 2015)، (TIMSS 2007). ومن ثم يطرح الباحث اسئلة الدراسة على النحو التالي:

١. هل يختلف توزيع بيانات متغيرات الدراسة لكلا مجموعتي الدراسة (TIMSS 2015، (TIMSS 2007)؟

http://dx.doi.org/10.29009/ijres.2.4.11

٢. ما مدى اختلاف نتائج التحليل عبر مداخل تحديد الدالة التمييزية للمفردة المختلفة (الخطأ من النوع الأول، عدد المفردات المتحيزة، تقارب حجم التأثير)؟

أهداف الدراسة:

الكشف عن التمييز والتحيز بين مفردات عبر مجموعتي صورتي اختبار (TIMSS 2015)، (TIMSS 2007). وذلك عن طريق توظيف اختبار تحليل التغاير وتحليل الانحدار المتعدد. والمقارنة بين الدالة التمييزية للمفردات بكلا الأسلوبين الاحصائيين في ضوء بعض المؤشرات مثل الخطأ من النوع الأول، عدد المفردات المتحيزة في كل صورة، تقارب حجم التأثير.

فروض الدراسة:

١. تتأثر مداخل الكشف عن تحيز المفردة بطبيعة توزيع البيانات للمفردات المتصلة والقيم المتطرفة لبياناتها.

٢. تعتمد مطابقة أداء مداخل الكشف عن الدالة التمييزية على مؤشرات الخطأ من النوع الأول، وعدد المفردات المتحيزة، وحجم التأثير.

الطريقة والاجراءات

أولاً: المنهجية: اعتمدت الدراسة على المنهج الوصفي والسببي المقارن وذلك في المقارنة بين مجموعتي صورتي اختبار (TIMSS 2015)، (TIMSS 2007) من خلال استخدام مدخل تقدير الدالة التمييزية للمفردات (أسلوب تحليل التغاير وتحليل الانحدار المتعدد).

ثانياً: عينة الدراسة: استخدمت الدراسة بيانات الارشيف لاختبار الاتجاهات الدولية لتعلم الرياضيات TIMSS واعتمدت الدراسة على اختبار الرياضيات لعدد بلغ ١٣١٦٤ طالباً وطالبة بالصف الثاني الاعدادي لـ TIMSS 2007 وعدد بلغ ٧٠٩٥ طالباً وطالبة لـ TIMSS 2015 وقد بلغ بمتوسط عمري ١٤,٣٤ عاماً بانحراف معياري ٠,٨٩ عاماً.

ثالثاً: اختبار الاتجاهات الدولية في تعليم الرياضيات TIMSS: اعتمد الباحث على الدرجات الأرشيفية لمفردات الاختبار عبر دورتي التطبيق ٢٠٠٧ و ٢٠١٥ والتي طبقت

http://dx.doi.org/10.29009/ijres.2.4.11

على طلاب جمهورية مصر العربية. وتكون الاختبار من أربعة مجالات للمحتوى وهي: مجال الجبر وتكون من خمسة مفردات، ومجال البيانات والاستدلال وتكون من خمسة مفردات، ومجال الاحصاء والاحتمالات وتكون من خمسة مفردات، ومجال الهندسة وتكون من خمسة مفردات.

رابعاً: إجراءات الدراسة:

١. الحصول على نتائج اختبارات (TIMSS 2015)، (TIMSS 2007) وإعدادها للتحليل الاحصائي في برنامج SPSS.
٢. التحقق من مسلمات الاعتدالية لتوزيع بيانات المجموعتين، والخطية، والتحقق من وجود قيم متطرفة في بيانات المتغير التابع لكلا المجموعتين.
٣. اختيار متغير المجموعات (TIMSS 2015)، (TIMSS 2007) كمتغير مستقل والمفردة المراد حساب الدالة التمييزية لها كمتغير تابع.
٤. إنشاء متغير Rest score واعتباره المتغير المصاحب في اسلوب تحليل التغاير ومستقلاً في أسلوب تحليل الانحدار، ويحسب Rest score بالفرق بين درجة البعد الكلي للمحتوى ودرجة المفردة المراد حساب الدالة التمييزية لها.
٥. التحقق من وجود فروق بين المجموعتين باستخدام اختبار ت للعينات المستقلة للتحقق من وجود فروق بين المجموعتين على درجات ابعاد المحتوى. لتحديد نوع الدالة التمييزية المستخدمة (المنتظمة، غير المنتظمة، المختلطة).
٦. استخدام أسلوب تحليل التغاير لتحديد الدالة التمييزية للمفردات في ضوء المؤشرات التالية:
   أ. تكون المفردة مميزة لأداء المجموعة المرجعية إذا كانت دالة احصائياً $P \leq 0.05$.
   ب. يحسب مؤشر حجم التأثير بمربع ايتا بحيث تكون متوسط $0.035 < \eta^2 < 0.070$ ومرتفع $\eta^2 \geq 0.070$.

http://dx.doi.org/10.29009/ijres.2.4.11

٧. استخدام أسلوب تحليل الانحدار المتعدد بطريقة Stepwise لتحديد الدالة التمييزية للمفردات في ضوء المؤشرات التالية:

أ. تكون المفردة مميزة لأداء المجموعة المرجعية إذا كانت متغير المجموعات دالاً احصائياً $P \leq 0.01$.

ب. يحسب مؤشر حجم التأثير بمؤشر مربع معامل الانحدار المتعدد هي $R^2 \leq 0.130$. وتكون الدالة التمييزية للمفردة متوسطة إذا كان $0.035 < R^2 < 0.070$.

٨. المفاضلة بين طرق حساب الدالة التمييزية للمفردة عن طريق المؤشرات التالية:

أ. توزيع بيانات المفردات لكل مجموعة.

ب. اتساق النتائج للدالة التمييزية للمفردات على أبعاد المحتوي لأداء المجموعتين.

ج. الخطأ من النوع الأول.

د. عدد المفردات المميزة في كل طريقة للمجموعة المرجعية.

نتائج البحث وتفسيرها

الاجابة على السؤال الأول: هل يختلف توزيع بيانات متغيرات الدراسة لكلا مجموعتي الدراسة (TIMSS 2015)، (TIMSS 2007)؟

استخدم الباحث اختبار كولمجروف سيمرنوف للتحقق من اعتدالية بيانات المجموعتين على أبعاد المحتوى، والجدول (١) يوضح دلالة الاختبار:

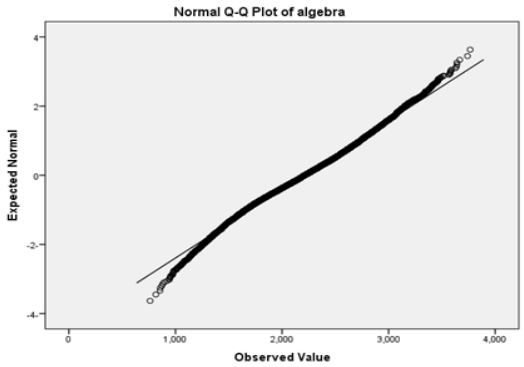
http://dx.doi.org/10.29009/ijres.2.4.11

جدول (١): دلالة الفروق لاختبارات اعتدالية بيانات المجموعتين.

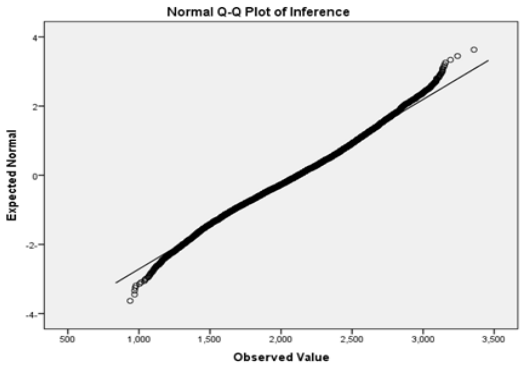
المجموعة	الاختبار	القيمة	د.ح	الدلالة
٢٠١٥	الجبر	٠,٠٢٧	٧٠٩٥	٠,٠٠٠
	البيانات والاستدلال	٠,٠٢٧	٧٠٩٥	٠,٠٠٠
	الكسور والارقام	٠,٠٢٦	٧٠٩٥	٠,٠٠٠
	الهندسة	٠,٠٢٨	٧٠٩٥	٠,٠٠٠
٢٠٠٧	الجبر	٠,٠٢٩	١٣١٦٤	٠,٠٠٠
	البيانات والاستدلال	٠,٠٢٠	١٣١٦٤	٠,٠٠٠
	الكسور والارقام	٠,٠٢٩	١٣١٦٤	٠,٠٠٠
	الهندسة	٠,٠٢٨	١٣١٦٤	٠,٠٠٠

توصلت نتائج التحليل إلى أن أبعاد المحتوى لمجموعتين التحليل كانت دالة احصائياً مما يعني عدم اعتدالية البيانات كما اسفرت نتائج مخرج كولمجروف سيمرنوف عن الأشكال الانتشارية لأبعاد المحتوى في المجموعتين كما يلي:

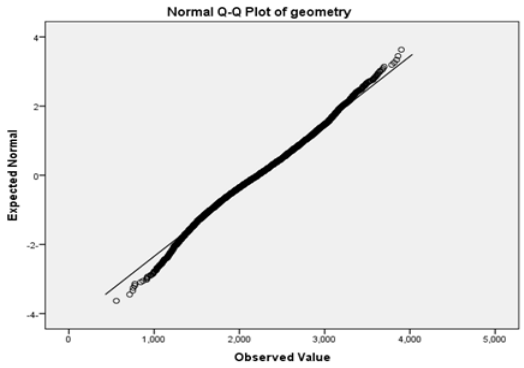
http://dx.doi.org/10.29009/ijres.2.4.11



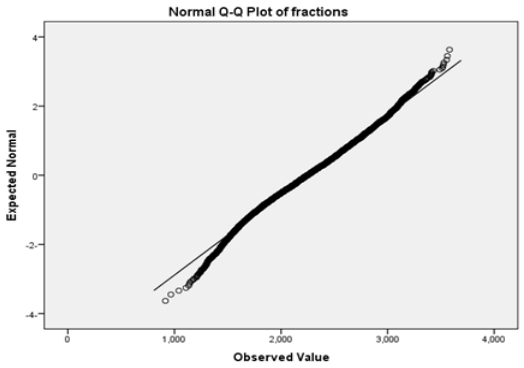
شكل (١-١): الانتشار لبعد الجبر TIMSS 2015.



شكل (١-٢): الانتشار لبعد البيانات TIMSS 2015



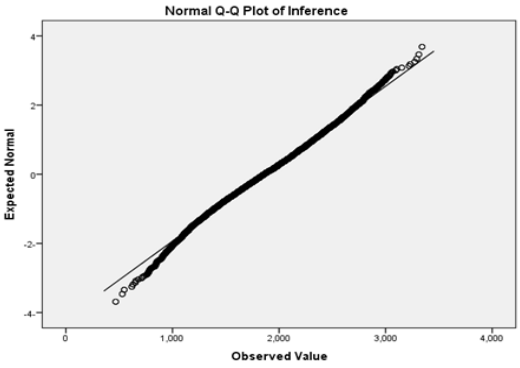
شكل (١-٣): الانتشار لبعد الكسور TIMSS 2015



شكل (١-٤): الانتشار لبعد الهندسة TIMSS 2015

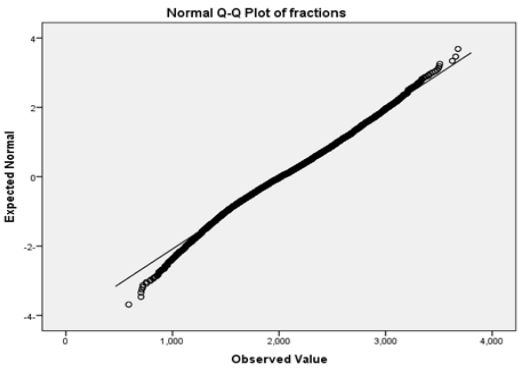
شكل (١): الرسوم البيانية لانتشار أبعاد الاختبار التحصيلي لعينة ٢٠١٥

http://dx.doi.org/10.29009/ijres.2.4.11

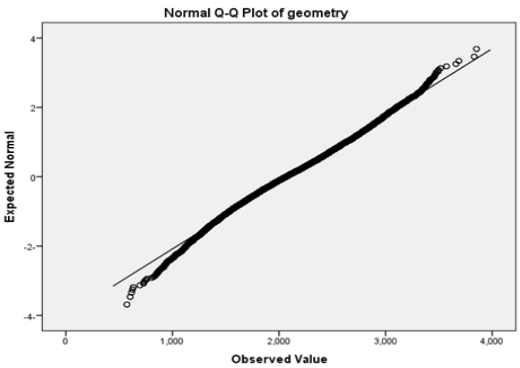


شكل (٢-١): الانتشار لبعد الجبر TIMSS 2007

شكل (٢-٢): الانتشار لبعد البيانات TIMSS 2007



شكل (٢-٣): شكل الانتشار لبعد الكسور TIMSS 2007



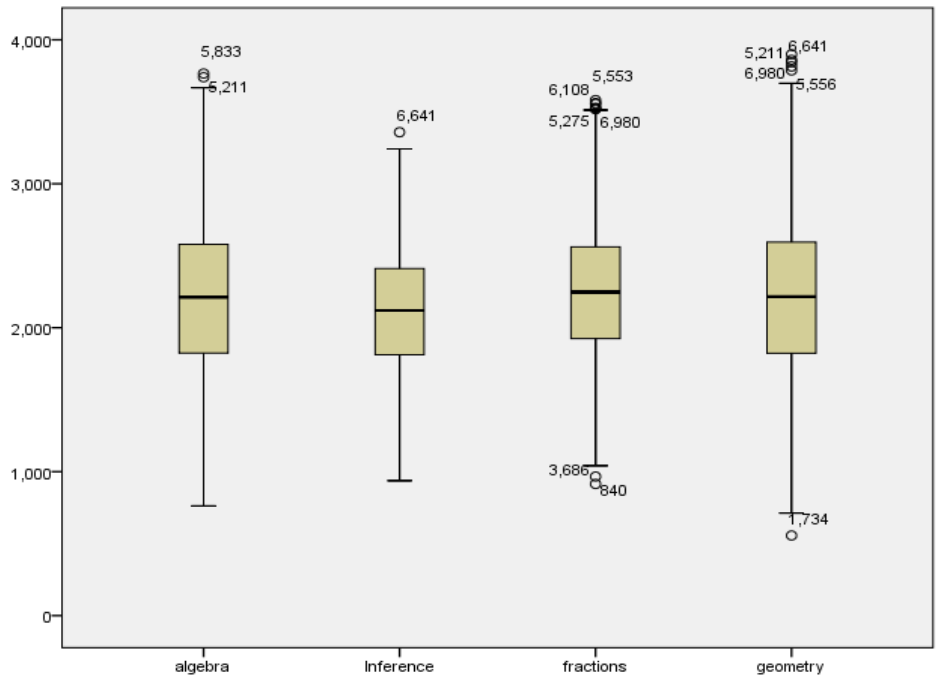
شكل (٢-٤): شكل لبعد الهندسة TIMSS 2007

شكل (٢): الرسوم البيانية لانتشار أبعاد اختبار TIMSS لعينة ٢٠٠٧

http://dx.doi.org/10.29009/ijres.2.4.11

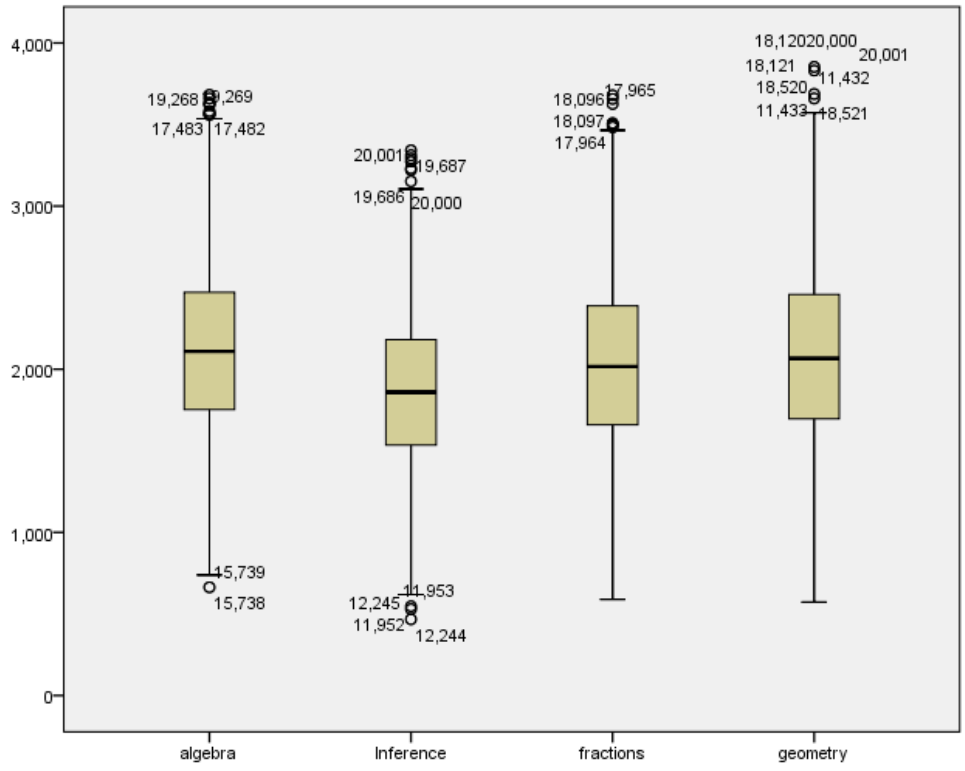
توصلت نتائج التحليل إلى أن أبعاد المحتوى لمجموعتين التحليل كانت دالة احصائياً مما يعني عدم اعتدالية البيانات وقد تتفق النتائج الحالية مع دراسة دانكن (Duncan, 2007) الذي يرى أن كبر حجم العينة بصورة متفاوتة بين العنتين قد يخلق خلل بتوزيعات العينة.

كما تتفق تماماً النتائج مع دراسة زويك وآخرون (Zwick et al., 1997) في طبيعة المفردات المتصلة فالمفردات تميل للتمييز بين المجموعات في استجابات الافراد عليها. وعلى الرغم من اتفاق نتائج توزيع بيانات متغير التحصيل في أنهما غير اعتداليان إلا أن نتائج اختبار "ت" ميز بين المجموعتين في أبعاد المنهج من ناحية.



شكل (٤): المربعات البيانية لأبعاد الاختبار التحصيلي لعينة ٢٠١٥

http://dx.doi.org/10.29009/ijres.2.4.11



شكل (٤): المربعات البيانية لأبعاد الاختبار التحصيلي لعينة ٢٠٠٧

بالتأمل في رسوم المربعات البيانية لعينة TIMSS 2007 يتضح وجود قيم متطرفة ايجاباً وسلباً لبعد الجبر وبعد البيانات والاستدلال مما يؤثر على توزيعات البيانات وأنها قد تعطي مؤشرات مضللة. وبالتأمل في شكل (٣) وشكل (٤) اتضح أن القيم المتطرفة ايجاباً عددها أكثر في TIMSS 2007 عنه في TIMSS 2015.

للإجابة على السؤال الثاني: ما مدى اختلاف نتائج التحليل عبر مداخل تحديد الدالة التمييزية للمفردة المختلفة (الخطأ من النوع الأول، عدد المفردات المتحيزة، تقارب حجم التأثير)؟

تم استخدام اختبار ت للعينات المستقلة للتحقق من دلالة الفروق بين دورتي تطبيق TIMSS لتحديد نوع الدالة التمييزية المستخدمة، والجدول (٢) بوضح دلالات الفروق بين المجموعتين:

http://dx.doi.org/10.29009/ijres.2.4.11

جدول (٢): دلالة الفروق بين المجموعات على أبعاد الاختبار والدرجة الكلية له.

بعد المحتوى	TIMSS 2015 (ن= ٧٠٩٥)		TIMSS 2007 (ن=١٣١٦٤)		قيمة ت	الدلالة
	المتوسط	الانحراف المعياري	المتوسط	الانحراف المعياري		
الجبر	٢٢٠٥,٧	٥٠٣,٩	٢١١٨	٤٨٧,٨	١١٢	٠,٠٠٠
البيانات والاستدلال	٢١٠٥,٨	٤٠٧,٧	١٨٦٣,٧	٤٤٦,٤	٣٧,٩	٠,٠٠٠
الكسور والارقام	٢٢٤٩,٩	٤٣٣,١	٢٠٣١,٨	٤٩٥,٢	٣١,٢	٠,٠٠٠
الهندسة	٢٢١٦,٢	٥١٨,٧	٢٠٨١,١	٥١٩,٤	١٧,٧	٠,٠٠٠
الدرجة الكلية	٨٧٧٧,٦	١٨١٨,٧	٨٠٩٤,٧	١٨٩٦,٤	٢٤,٨	٠,٠٠٠

أسفرت نتائج التحليل عن وجود فروق دالة احصائياً بين دورتين التطبيق لاختبار TIMSS في الرياضيات، وهذا يعني استخدام الدالة التمييزية المنتظمة حيث وجد اختلاف بين المجموعتين في درجاتهم على أبعاد المحتوى والدرجة الكلية للاختبار.

اعتمدت الدراسة على طريقتين لتقدير الدالة التمييزية للمفردات على كل بعد بحيث تكون درجة المفردة هي متغيراً تابعاً بينما تكون المجموعة متغيراً تصنيفياً مستقلاً، كما يكون متغير التفاعل Rest score هي متغيراً مصاحباً مرة في أسلوب تحليل التغاير مرة، ومتغيراً مستقلاً في تحليل الانحدار بطريقة Stepwise. والجدول (٣) يوضح الدوال التمييزية لمفردات وحجم التأثير في كل طريقة:

جدول (٣): نتائج الدالة التمييزية لمفردات اختبار TIMSS 2007, 2015.

بعد المحتوى	المفردة	تحليل التغاير			تحليل الانحدار	
		F	الدلالة	Eta square	الدلالة	R^2
الجبر	Alg1	٣,٢٢	٠,٠٧٣	–	٠,٠٧٣	–
	Alg2	٤,٤٤	٠,٠٣٥	٠,٠٠٠	٠,٠٧٣	–
	Alg3	٤,١٣	٠,٠٤٢	٠,٠٠٠	٠,٠٣٥	٠,٨١
	Alg4	١,١٨	٠,٢٧٨	–	٠,٢٧٨	–

http://dx.doi.org/10.29009/ijres.2.4.11

بعد المحتوى	المفردة	تحليل التغاير			تحليل الانحدار	
		F	الدلالة	Eta square	الدلالة	R^2
	Alg5	٥,١٣	٠,٠٢٤	٠,٠٠٠	٠,٠٢٤	٠,٨١
البيانات والاستدلال	Dap1	٣٥,١١	٠,٠٠٠	٠,٠٠٢	٠,٠٠٠	٠,٧٧
	Dap2	٧,١٢	٠,٠٠٨	٠,٠٠٠	٠,٠٠٨	٠,٧٧
	Dap3	٧,١١	٠,٠٠٨	٠,٠٠٠	٠,٠٠٨	٠,٧٦
	Dap4	٤٦,٨٢	٠,٠٠٠	٠,٠٠٢	٠,٠٠٠	٠,٧٧
	Dap5	٢,٥٩	٠,١٠٨	–	٠,١٠٨	–
الكسور والارقام	Frac1	١٤,٨٢	٠,٠٠٠	٠,٠٠١	٠,٠٠٠	٠,٨٣
	Frac2	٠,١٢٠	٠,٧٢٧	–	٠,٧٢٧	–
	Frac3	٣٧,٤٩	٠,٠٠٠	٠,٠٠٢	٠,٠٠٠	٠,٨٢
	Frac4	٢٢,٥٠	٠,٠٠٠	٠,٠٠١	٠,٠٠٠	٠,٨٣
	Frac5	٠,٣١	٠,٥٧٥	–	٠,٥٧٥	–
الهندسة	Geo1	١٠٢,٦١	٠,٠٠٠	٠,٠٠٥	٠,٠٠٠	٠,٨٢
	Geo2	٧١,٢٩	٠,٠٠٠	٠,٠٠٤	٠,٠٠٠	٠,٨٢
	Geo3	٤٢,٣٣	٠,٠٠٠	٠,٠٠٢	٠,٠٠٠	٠,٨١
	Geo4	٩١,٠٨	٠,٠٠٠	٠,٠٠٤	٠,٠٠٠	٠,٨١
	Geo5	٢٦,٢٨	٠,٠٠٠	٠,٠٠١	٠,٠٠٠	٠,٨٢

اتفقت النتائج مع (Cole et al., 2000) على اختلاف دورتي التقييم لعـامي ٢٠٠٧ و ٢٠١٥ في النتائج في التفريق بين مجموعتي الدراسة. إلا أنه بالرغم من هذا الاتفاق فقد تختلف النتـائج جزئيـاً في طبيعة الثقافة لكلا الدراستين كما أن الصفة المقاسة في دراسة كوول وآخرون (Cole et al., 2000) هـي سمة انفعالية بينما في الدراسة الحالية هي سمة معرفية.

وجاءت نتائج الدراسة في وجود اتفاق المفردة Alg3 و Alg4 و Alg5 في بعد الجبر أنها متحيـزة للمجموعة المرجعية TIMSS 2015. وهـذا يعنـي أن المفـردة قـد تكـون تكـررت في وجودهـا داخـل

http://dx.doi.org/10.29009/ijres.2.4.11

الاختبار في نسختي التطبيق مع تغيير طفيف في نتائج المفردة ومن ثم فقد أشار حجم التأثير إلى وجود تأثير ضعيف على مؤشري مربع ايتا ومربع معامل الارتباط المتعدد.

وتعارضت النتائج فيما يخص المفردة Alg2 ببعد الجبر بين طريقتي التقدير في تقييم التمييز بين المجموعتين فقد كانت نتائج اسلوب تحليل التغاير دالة وذات حجم أثر ضعيف، بينما في أسلوب تحليل الانحدار كانت النتيجة عدم الدلالة الاحصائية. وكانت المفردة Alg1 في بعد الجبر غير دالة احصائياً في طريقتي تقدير تمييز المفردة مما يعني تحرر المفردة عبر المجموعات من التمييز. وهذا يعني ملائمة المفردة لكلا العينتين سواء من حيث الوضوح وطريقة التعلم.

وجاءت النتائج متفقة لبعد البيانات والاستدلال في المفردات DAP1 و DAP2 و DAP3 و DAP4 وجود فروق بين المجموعتين في تمييز المفردة لاستجابات المجموعتين (TIMSS 2015)، (TIMSS 2007)، بينما كان حجم التأثير في كلا الطريقتين ضعيفاً.

وأسفرت النتائج عن عدم وجود فروق دالة في بعد البيانات والاستدلال على المفردة DAP5 باستخدام طريقتي تقدير التمييز. كذلك اتفقت النتائج على أن المفردات FRAC2 و FRAC5 كانت غير متحيزة إذ لم تظهر تمييزاً بين المجموعتين في الأداء التحصيلي ببعد الهندسة. وتوصلت النتائج إلى وجود فروق دالة احصائياً على جميع مفردات بعد الهندسة الخمسة والتي بدت تحيزاً بين المجموعتين على مفردات البعد ولكن قيمة حجم التأثير كانت ضعيفة في كلا الطريقتين.

وبالتأمل في النتائج فلا يوجد تحيز للمفردات عبر الزمن مما يعني أن المنهج المدروس الذي تعلمه الطلاب يتوافق مع مواقفهم وإنجازاتهم ودراسة المفاهيم والعمليات والمواقف التي تعلمها في الرياضيات. كما أن دراسة العوامل المرتبطة بفرصة تعلم الطلاب لم تتغير عبر الزمن وهذا يعني أن المنهج يتحرر من العوامل الثقافية والتكنولوجية لتعلم الرياضيات وهذا يتفق مع (Thomson et al., 2016).

http://dx.doi.org/10.29009/ijres.2.4.11

واتفقت نتائج الدراسة عبر مرات التطبيق فقد كان متوسط معاملات الارتباط بين قياسات التطبيقين بلغ ٠,٧٦ وهذا يعني اتساق القياسات عبر المجموعتين، علاوة على هذا فلم يوجد تحيز للمفردات إلا المفردة Alg2 في بعد الجبر وهذا اتفق مع نتائج دراسات جونسون، وساشي وهاغ (Johansone, 2015; Sachse & Haag, 2017) التي أكد أن تحيزات المفردات في البيئة الواحدة تكون أقل ما يمكن عبر الزمن في ضوء الدالة التمييزية.

واختلفت الدراسة مع نتائج زويك وآخرون (Zwick et al., 1997) في أن انتهاك البيانات لشرط الاعتدالية لم يسبب التمييز بين المجموعات. وهذا يبرره أن اختبار تحليل الانحدار اختباراً منيعاً لشرط الاعتدالية.

ويعزي الباحث حدوث اختلاف في تحيز المفردة في بعد الجبر إلى ارتفاع عدد أفراد عينة اختبار TIMSS 2007 من ناحية كما أن القيم المتطرفة ايجاباً وسلباً مما سبب فروقاً في المجموعة المرجعية خصوصاً بعد ضبط متغير التفاعل الذي أدى لتشوه الاختلافات بين المجموعات في أسلوب تحليل التغاير، وهذا يؤكده قيمة الاعتمادية الخطية الطفيفة لمتغير التفاعل في اختبار تحليل الانحدار المتعدد وهذا يؤكده دراسات هارتر وآجراوال، وروث وويلسون (Harter & Agrawal, 2011; Ross & Willson, 2017).

وقد تمثل الخطأ من النوع الأول في تحيز المفردة ٢ في بعد الجبر الناتج عن تفاوت حجوم العينة في اختباري الدراسة، انتهاك شرط الاعتدالية، وتطرف القيم إيجاباً وسلباً في كلا الاختبارين (TIMSS 2015، TIMSS 2007). وهذا التحيز نشأ في أحد طريقتي التقدير دون الأخرى وهو تحيز طفيف يمكن التغاضي. واتفقت الدراسة مع نتائج فينش (Finch, 2016) في تغلب على الخطأ من النوع الأول (تفاوت الزيادة في أحجام أحد العينات الداخلة للتحليل) أثناء حساب تحيز مفردات الاختبار باستخدام اختبار تحليل الانحدار المتعدد. وبالرغم من هذا الاتفاق وجد بعض الاختلافات بين الدراستين فبيانات الدراسة الحالية بيانات تجريبية، كما أن الدراسة الحالية اخفقت فيما يتعلق بالمفردة الثانية في بعد الجبر.

http://dx.doi.org/10.29009/ijres.2.4.11

واتفقت الدراسة جزئياً مع نتائج دراسة المحاكاة لسوامنسان وروجرز (Swaminathan & Rogers, 1990) والتي توصلت إلى عدم وجود تحيز أو تمييز في مفردات الاختبار باستخدام تحليل الانحدار. وبالرغم من هذا الاتفاق بين الدراستين فقد اختلفت طبيعة الاختبارات في كلا الدراستين ففي الدراسة الحالية المفردات متصلة تقع في مستوى القياس الفترى أما في دراسة المحاكاة كانت المفردات تقع ضمن مستوى القياس الترتيبي.

واختلف الباحث مع دراسة إلينا-أوليفيري وآخرون (Elena-Oliveri et al., 2018) التي أكدت على وجود تحيز في المفردات في الولايات المتحدة الأمريكية، وهذا يرجع إلى اختلاف الطبيعة الثقافية للطلاب الذين أجري عليهم الاختبار. إلا أن الدراسة الحالية لم توجد فيها إلا تحيزاً في مفردة واحدة في بعد الجبر وهذا يبرر أن تعلم الرياضيات أقرب إلى المسلمات وأن المفردات تعتمد في جوهرها على التجهيز المعرفي للمعلومات والأساليب المعرفية لدى المتعلم، أو قد تكون بسبب توقع أفكار المفردات التي وردت في صورة ٢٠١٥ في ضوء تسلسل الصور السابقة.

وتعاني الدراسة من بعض المحددات منها التفاوت في حجم العينات ففي مجموعتين اختبارات (TIMSS 2015)، (TIMSS 2007) كما أن الدراسة لم تستبعد الحالات التي سببت البيانات المتطرفة ايجاباً وسلباً، قبل تقدير التحيز من خلال الدالة التمييزية. كما أن عدم اعتدالية التوزيع أدت إلى وجود تحيز طفيف في بعض مفردات الاختبار وأدت لتعارض نتائج مؤشرات الدالة التمييزية.

مناقشة النتائج والتعليق عليها:

يرجع تباين قياس طبيعة الظاهرة النفسية إلى طبيعة المرحلة العمرية، اختلاف الثقافات، الجنس، والزمن. ومن امثلة تلك المقاييس ذات الحساسية العالية للطبيعة الثقافية مقياس العوامل الخمسة الكبرى للشخصية ومقياس روسنبرج لتقدير الذات وهذا مغايراً إلى حد ما لأهداف الدراسة إذ أن هذه الاختبارات تتمايز مفرداتها ثقافياً، بينما يتأثر مقياس وكسلر بلفيو ومقياس ستانفورد بينيه بالمرحلة العمرية. ويختلف الأداء التفاضلي لبعض الاختبارات المعرفية مثل TOEFL iBT لدى الطلاب الدوليين من مجتمعات تتحدث الانجليزية عن نظيرتها من المجتمعات. وبمراجعة التراث

http://dx.doi.org/10.29009/ijres.2.4.11

النفسي وجد الباحث اهتمام الدراسات النفسية بتباين الخصائص السيكومترية كالصدق البنائي والثبات عبر اختلاف الثقافات والمراحل العمرية، وإهمال مدى تحيز مفردات هذه المقاييس لهذه الثقافات أو المراحل والذي يقاس بالدالة التمييزية. كما أن أداة الدراسة ذات طبيعة معرفية ترتبط إلى حد كبير بالمسلمات ويرتبط تحيز مفرداتها إما بالمرحلة العمرية أو القدرات العقلية وتجهيز المعلومات.

ويتمثل جوهر الدالة التمييزية في قياس التحيز الثقافي أو التحيز للجنس أو للعرق أو العمر أو عبر الزمن بين مفردات المقياس عبر مجموعات القياس ويصبح الاختبار أو المقياس غير متحيزاً عند تساوى السمة المقاسة عبر مجموعات القياس. وتتميز طبيعة الدراسة الحالية عن مثيلتها من الدراسات السابقة في السعي إلى دراسة تحصيل الرياضيات في البيئة المصرية كدراسة طولية اعتمدت على بيانات أرشيفية لاختبار (TIMSS 2015)، (TIMSS 2007) لدراسة أثر التحيز الزمني لمفردات الاختبار. ويوصي الباحث ببعض المحكات كمحدد لدراسة التحيز لمفردات اختبار تحصيلي ذو متغيرات تابعة متصلة:

١. دراسة اعتدالية المفردات باستخدام اختبار كولمجروف سيمرنوف، فإنتهاك شرط الاعتدالية قد يؤدي في تحيز المفردة للمجموعة المرجعية.
٢. دراسة خطية البيانات الداخلة للتحليل.
٣. مدى تطرف بيانات المفردات الداخلة للتحليل. ويفضل استبعاد المفردات ذات التطرف الايجابي بالزيادة أو التطرف السلبي بالنقص، والتطرف المزدوج ذوي الحالات السلبية والايجابية.
٤. ضرورة تقارب العينات الداخلة للتحليل في العدد حيث أن الفرق الشاسع بين كلا المجموعتين في العدد يولد خطأ من النوع الأول ويولد تحيز في بعض المفردات مما يجعل اتخاذ القرار مضللاً إلى حد ما.
٥. تحليل نوع الدالة التمييزية المستخدمة لدراسة التحيز على النحو التالي:

http://dx.doi.org/10.29009/ijres.2.4.11

- الدالة التمييزية المنتظمة ويتم فيها استخدام اختبار ت للعينات المستقلة أو ت المرتبطة لدراسة دلالة الفروق بين الدرجة الكلية للاختبار، ودرجات أبعاد المقياس وتتفق نتائج التحليل على وجود فروق دالة احصائياً.
- الدالة التمييزية غير المنتظمة ويتم فيها استخدام اختبار ت للعينات المستقلة أو ت المرتبطة لدراسة دلالة الفروق بين الدرجة الكلية للاختبار، ودرجات أبعاد المقياس وتتفق نتائج التحليل على وجود فروق دالة احصائياً على الدرجة الكلية وبعض أبعاد المقياس دالة والبعض عديم الدلالة الاحصائية.

٦. إنشاء متغير rest score وهو عبارة عن الدرجة الكلية للبعد مطروحاً منها درجة المفردة المراد دراسة التحيز لها.

٧. اختيار الاسلوب الاحصائي الأمثل لطبيعة المفردة المراد دراسة التحيز لها على النحو التالي:

- إذا كانت المفردة من ذات درجات متصلة أو متغيراً فترياً يستخدم اختبار تحليل الانحدار المتعدد بطريقة stepwise. أو اختبار تحليل التغاير.
- إذا كانت المفردة تتبع مستوى القياس الترتيبي يستخدم اختبار تحليل الانحدار اللوجستي أو تحليل الانحدار بطريقة بواسون.
- إذا كانت المفردة تتبع مستوى القياس الاسمي يستخدم اختبار مربع كاي أو اختبار مانتل هانزل.
- إذا كانت المفردة ثنائية الاستجابة (0, 1) Binary يستخدم اختبار تحليل الانحدار اللوجستي، أو التحليل التمييزي المتدرج.

٨. عند استخدام اختبار تحليل الانحدار المتعدد يراعي التحقق من الاعتمادية الخطية المؤشرات التالية (Ross & Willson, 2017):

http://dx.doi.org/10.29009/ijres.2.4.11

- مؤشر VIF للاعتمادية الخطية المتعددة multicollinearity لا يتخطى القيمة ١٠.
- مؤشرات الحالة Condition Indices للنموذج لا تتخطى القيمة ٣٠.
- استبعاد تفاعل المتغيرات من النموذج من جدول المتغيرات المستبعدة Excluded Variables وذلك بأن تكون قيمة $P \leq 0.05$.

٩. الحكم على مدى تحيز المفردة في ضوء مؤشرات حجم التأثير، إذ تكون الدلالة الاحصائية ناجمة عن الصدفة، وذلك في ضوء نقاط القطع التالية:

جدول (٤): نقاط القطع لمؤشرات الدالة التمييزية للحكم على تحيز مفردات اختبار.

الاختبار المستخدم	تحيز مقبول	التحيز الضعيف	التحيز المتوسط	التحيز المرتفع
ANCOVA	$P \leq 0.05$	$\eta^2 \leq 0.035$	$0.035 < \eta^2 < 0.070$	$\eta^2 \geq 0.070$
Multiple linear regression	$P \leq 0.01$ $R^2 \leq 0.130$		$0.035 < R^2 < 0.070$	

http://dx.doi.org/10.29009/ijres.2.4.11

References:

- Abah, J. (2018). The quest for statistical significance: Ignorance, bias and malpractice of research practitioners. International Journal of Research and Review, 5(3), 112-129.
- Akcan, R., & Kabasakal, K. A. (2019). An Investigation of Item Bias of English Test: The Case of 2016 Year Undergraduate Placement Exam in Turkey. International Journal of Assessment Tools in Education, 6(1), 48-62.
- Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. Journal of Educational and Behavioral Statistics, 41(6), 559-592.
- Carnoy, M., Khavenson, T., & Ivanova, A. (2015). Using TIMSS and PISA results to inform educational policy: a study of Russia and its neighbours. Compare: A Journal of Comparative and International Education, 45(2), 248-271.
- Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: experience from the New Haven EPESE study. Journal of clinical epidemiology, 53(3), 285-289.
- Duncan, S. C. (2007). Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel-Haenszel DIF methods(Doctoral dissertation, Texas A&M University).
- Elena-Oliveri, M., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. Applied Measurement in Education, 31(1), 1-16.

http://dx.doi.org/10.29009/ijres.2.4.11

- Engelhard Jr, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied measurement in education, 3(4), 347-360.
- Feingold, A. (2013). A regression framework for effect size assessments in longitudinal modeling of group differences. Review of General Psychology, 17(1), 111.
- Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. Applied Measurement in Education, 29(1), 30-45.
- Gesicki, A. (2015). Decision rules based on hypothesis tests and effect sizes for logistic regression differential item functioning (Doctoral dissertation, University of British Columbia).
- Grønmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. (2015). TIMSS 2015 mathematics framework. TIMSS, 11-27.
- Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. Applied measurement in Education, 12(3), 211-235.
- Harter, J. K., & Agrawal, S. (2011). Cross-cultural analysis of Gallup's Q12 employee engagement instrument. Omaha, NE: Gallup.
- Innabi, H., & Dodeen, H. (2018). Gender differences in mathematics achievement in Jordan: A differential item functioning analysis of the 2015 TIMSS. School Science and Mathematics, 118(3), 127-137.
- Johansone, I. (2015). Survey operations procedures in TIMSS 2015. Methods and procedures in TIMSS.

http://dx.doi.org/10.29009/ijres.2.4.11

- Karami, H. (2012). An introduction to differential item functioning. The International Journal of Educational and Psychological Assessment.
- Karpen, S. C. (2017). Misuses of Regression and ANCOVA in Educational Research. American Journal of Pharmaceutical Education, 81(8), 65-101.
- Liou, P. Y., & Hung, Y. C. (2015). Statistical techniques utilized in analyzing PISA and TIMSS data in science education from 1996 to 2013: A methodological review. International Journal of Science and Mathematics Education, 13(6), 1449-1468.
- Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL® iBT reading performance? A confirmatory approach to differential item functioning. Applied Measurement in Education, 24(3), 235-255.
- Martin, M. O., Mullis, I. V., & Foy, P. (2015). TIMSS 2015 assessment design. TIMSS, 85-99.
- Ross, A., & Willson, V. L. (2017). Multiple Regression with Two Continuous Predictors and the Interactions Betweenbetween Them. In Basic and Advanced Statistical Tests (pp. 75-86). SensePublishersSense Publishers, Rotterdam.
- Sachse, K. A., & Haag, N. (2017). Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. Applied Measurement in Education, 30(2), 102-116.
- Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. International Journal of Testing, 3, 129–150.

http://dx.doi.org/10.29009/ijres.2.4.11

- Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations. Journal of Cross-Cultural Psychology, 37, 557–567.
- Stephens, M., Landeros, K., Perkins, R., & Tang, J. H. (2016). Highlights from TIMSS and TIMSS Advanced 2015: Mathematics and Science Achievement of US Students in Grades 4 and 8 and in Advanced Courses at the End of High School in an International Context. NCES 2017-002. National Center for Education Statistics.
- Stone, E., Cook, L., Laitusis, C. C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. Applied Measurement in Education, 23(2), 132-152.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.
- Sweeney, K. P. (1996). A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning. Unpublished doctoral dissertation, Fordham University, New York, NY
- Thomson, S., Wernert, N., O'Grady, E., & Rodrigues, S. (2016). TIMSS 2015: A first look at Australia's results.
- Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. Applied Measurement in Education, 9(2), 175-199.

http://dx.doi.org/10.29009/ijres.2.4.11

- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). Ottawa: National Defense Headquarters.
- Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.
- Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. Applied Measurement in Education, 10(4), 321-344.

http://dx.doi.org/10.29009/ijres.2.4.11