OPEN ACCESS
تاريخ الاستلام: 16/ 08/ 2025م
تاريخ القبـول: 28/ 09/ 2025م

مجلة الآداب
للدراسات النفسية والتربوية

أثر حجم العينة وطول المقياس وتوازن المجموعات على أداء طريقة مانتل–هانزل العامة (GMH) في كشف الأداء التفاضلي للفقرات متدرجة الاستجابة: دراسة محاكاة

ماجد محمود شريف الجودة*

majed_jodeh@hotmail.com

الملخص

تتحقّق هذه الدراسة من كفاءة طريقة مانتل–هانزل العامة (GMH) في كشف الأداء التفاضلي للفقرة (DIF) في المقاييس متدرجة الاستجابة، من خلال الاعتماد على أسلوب محاكاة مونتِ كارلو لنموذج الاستجابة المتدرجة (الفقرات متدرجة الاستجابة) (GRM) مع تغيير عوامل تصميمية أساسية: حجم العينة، طول الاختبار، توازن أحجام المجموعتين، نسبة الفقرات ذات الأداء التفاضلي، ونوع الأداء التفاضلي وشدته. أظهرت النتائج أن قوة الكشف ترتفع كلما زادت شدّة الDIF ، وأن GMH أكثر حساسية للDIF المنتظم مقارنة بغير المنتظم، مع بقاء معدل الخطأ من النوع الأول قريبًا من المستوى الاسمي. كما أن عدم توازن المجموعات يضعف القوة، بينما يرفعها طول الاختبار وحجم العينة. عمليًا: عندما يكون الاختبار أطول والمجموعتان متقاربتين عددًا والحجم الكلي كافٍ، تعمل GMH جيدًا؛ وعند الاشتباه بوجود DIF غير منتظم يُستحسن زيادة العينة واستخدام أساليب داعمة (مثل IRT-LR أو MIMIC) لتعزيز عدالة القياس ودقته.

الكلمات المفتاحية: مانتل–هانزل العامة (GMH)؛ الأداء التفاضلي للفقرة (DIF)؛ نموذج الاستجابة المتدرجة (GRM)؛ مقاييس متدرجة الاستجابة

* استاذ القياس والتقويم المشارك، قسم التربية وعلم النفس، كلية التربية والآداب، جامعة تبوك، السعودية

للاقتباس: الجودة، ماجد محمود شريف. (2025). أثر حجم العينة وطول المقياس وتوازن المجموعات على أداء طريقة مانتل–هانزل العامة (GMH) في كشف الأداء التفاضلي للفقرات متدرجة الاستجابة: دراسة محاكاة ، *مجلة الآداب للدراسات النفسية والتربوية*، 7(4)، 9-33.

OPEN ACCESS
Received: 16-08-2025
Accepted: 28-09-2025

مجلة الآداب
للدراسات النفسية والتربوية

Effects of Sample Size, Test Length, and Group Balance on the Performance of the Generalized Mantel–Haenszel (GMH) Method in Detecting Differential Item Functioning for Graded Response Items: A Simulation Study

Majed Mahmoud Sharif Jodeh *

majed_jodeh@hotmail.com

Abstract

This study examines the effectiveness of the Generalized Mantel–Haenszel (GMH) method for detecting differential item functioning (DIF) in graded response items. Monte Carlo simulation under the graded response model (GRM) was employed, while systematically modifying factors of sample size, test length, group-size balance, the proportion of DIF items, and the type and magnitude of DIF. Results showed that statistical power increased with DIF magnitude and that GMH was more sensitive to uniform than to nonuniform DIF. At the same time, Type I error remained close to the nominal level. Power declined under group imbalance, whereas longer tests and larger samples improved performance. Practically, GMH performed well when tests were longer, groups were balanced, and the overall sample size was adequate. When nonuniform DIF was suspected, increasing the sample size and complementing GMH with additional methods, such as the IRT likelihood-ratio (IRT–LR) test or MIMIC models, can strengthen measurement fairness and accuracy.

Keywords: Generalized Mantel–Haenszel (GMH); Differential Item Functioning (DIF); Graded Response Model (GRM); polytomously scored scales.

* Associate Professor of Measurement and Evaluation, Department of Education & Psychology, College of Education and Arts, Tabuk University, Saudi Arabia.

Cite this article as:: Jodeh, Majed Mahmoud Sharif. (2025). Effects of Sample Size, Test Length, and Group Balance on the Performance of the Generalized Mantel–Haenszel (GMH) Method in Detecting Differential Item Functioning for Graded Response Items: A Simulation Study. *Journal of Arts for Psychological & Educational Studies 7*(4) 9-33

المقدمة:

يدرس الباحثون - بشكل تقليدي - السلوك التفاضلي لفقرات الاختبارات والمقاييس، والذي أطلق عليه مصطلح الأداء التفاضلي للفقرة Item Differential functioning (DIF) ويعرف الأداء التفاضلي للفقرة، بأنه: اختلاف احتمالية استجابة الأفراد على أحد خيارات الاستجابة للفقرة بشكل خاص (مثلاً الإجابة الصحيحة لها) باختلاف مجموعات المستجيبين الذين يملكون القدرة نفسها للسمة موضع القياس (Holland &Thyer,1988) .

ففي سياق الفقرات ثنائية الاستجابة فيمكن – بسهولة - تفسير الأداء التفاضلي للفقرة بأنه: اختلاف احتمالية إجابة الأفراد عند القدرة نفسها للسمة موضع القياس إجابة صحيحة باختلاف مجموعات المستجيبين، ولكن تصبح الأمور أكثر تعقيداً في تفسير الأداء التفاضلي للفقرة عندما يتعلق الأمر بالفقرة متعددة الاستجابات أو الفقرات ذات الاستجابات المتدرجة، فقد ميز الباحثون بين أسلوبين لتفسير الأداء التفاضلي لمثل هذه الفقرات، فبعضهم ينظر للأداء التفاضلي على أنه اختلاف متجه في الاستجابة على الفقرة بين الأفراد ذوي القدرة نفسها للسمة موضع القياس عبر جميع خيارات الاستجابة للفقرة باختلاف مجموعات المستجيبين، وبعضهم الآخر يرى أن الأداء التفاضلي للفقرة متدرجة الاستجابة؛ اختلاف في احتمالية اختيار أحد خيارات الاستجابة للفقرة بشكل خاص عندما يستجيب عليها أفراد من القدرة نفسها للسمة موضع القياس باختلاف مجموعة المستجيبين (Penfield et al.,2009; Penfield, 2010)

وبغض النظر عن نوع الفقرات وآليات تفسير الأداء التفاضلي لها، فأمر وجود الأداء التفاضلي بحد ذاته مقلق ويؤثر سلباً على خصائص أدوات القياس، التي قد تحابي مجموعة على أخرى في الأداء، رغم امتلاكهم القدرة نفسها للسمة موضع القياس في حالة وجوده، ويؤثر أيضاً على تكافؤ القياس عبر المجموعات المختلفة من المستجيبين. فالكشف عن الأداء التفاضلي في فقرات المقياس أمر ضروري ويسهم في تحقيق عدالة أدوات القياس، وزيادة دقة التنبؤات حول السمات موضع القياس (Jafari et al., 2013; Thissen, 2001)

ونظرًا لأن أداء الفقرات التفاضلي (DIF) يمكن أن يؤثر سلبًا على استنتاجات المقياس وتصنيف الأفراد. فقد قامت العديد من الدراسات بالتحقيق في ذلك في اختبارات التحصيل مثل TOEFL (اختبار اللغة الإنجليزية واختبار التقييم الدراسي)، MELAB (بطارية تقييم اللغة الإنجليزية في ميشيغان) (Park, 2008; Wagner,2004; Eom, 2008; Vahid et al., 2011).

ففي مجال بناء الاختبارات وتطوير أدوات القياس، ازدادت أهمية توفير أدلة على صدق هذه الأدوات ومن بينها مؤشرات الأداء التفاضلي للفقرات (DIF)، الذي عدّته أحدث نسخة لمعايير بناء الاختبارات التربوية والنفسية مؤشراً مهماً على الصدق البنائي للمقياس. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education,2014

ميز الباحثون بين نوعين من الأداء التفاضلي: الأداء التفاضلي المنتظم (UNDIF) وغير المنتظم (NUNDIF)، مع التركيز على العلاقة بين عضوية المجموعة ومستوى القدرة، وفقًا لأكرمان (1992)، فإن منحنيات خصائص الفقرات للمقياس (ICCs) متوازية مع UNDIF وغير متوازية مع NUNDIF (Ackerman, 1992; Mellenbergh, 1989; Millsap & Everson, 1993; Narayanon & Swaminathan, 1996)

طرق اكتشاف الأداء التفاضلي DIF

طور الباحثون عدداً من الطرق للكشف عن الأداء التفاضلي للفقرة، منها: نسبة الأرجحية-لمانتل هانزل Mantel-Haenszel –Log Odds Ratio (MH-LOR)، والمساحة والانحدار اللوجستي، ونموذج راش، واختبارات تحيز الفقرة المتزامن Simultaneous Item Bias (SIBTETSs)، والمجموعات المتعددة Multiple group في نظرية الاستجابة للفقرة IRT، والنماذج متعددة المؤشرات متعددة الأسباب Multiple Indicators Multiple Causes Models MIMIC

تقييم الأداء التفاضلي للفقرات متدرجة الاستجابة باستخدام طريقة مانتل هانزل العامة:

في سياق الفقرات متدرجة الاستجابة، التي هي محل اهتمام وتركيز هذه الدراسة، فإن من الطرق الشائعة والمثبتة لاكتشاف الأداء التفاضلي في هذا النوع من الفقرات؛ هي طريقة مانتل هانزل العامة General Mantel Haenszel Statistics (GMH)، وشاع استخدام طريقة مانتل هانزل في فحص الأداء التفاضلي للفقرات ثنائية الاستجابة، وتباعاً تم استخدامها في الفقرات متدرجة الاستجابة، وتبين أنها أداة جيدة لهذا الغرض.(Penfield,2001)

تُعد طريقة مانتل هانزل في صورتها الأصلية لاكتشاف الأداء التفاضلي امتداد لاختبار دلالة العلاقة باختبار كاي-تربيع χ^2، إذ تقوم الطريقة على مقارنة استجابات الأفراد بين مجموعتين الأولي تسمى المجموعة المرجعية Reference Group والمجموعة الثانية تسمى المجموعة البؤرية أو المركزية

Focal Group عبر مستويات مختلفة من قدرات الأفراد، وعندما نختبر الفرضية الصفرية المتعلقة بعدم وجود أداء تفاضلي للفقرة تظهر قيمة MH χ^2 في المعادلة رقم 1 ورقم 2:

$$\chi^2 = \frac{\left\{\left|\sum_{j=1}^{k}[A_j - E(A_j)]\right| - 0.5\right\}^2}{\sum_{j=1}^{k} Var(A_j)} \quad \text{.....(1)}$$

$$Var(A_j) = \frac{nR_j nF_j m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \quad \text{.....(2)}$$

حيث:

K: عدد مستويات القدرة للسمة موضع القياس.

j: مستوى القدرة رقم j

A_j : عدد الاستجابات الصحيحة الملاحظة في المجموعة المرجعية على الفقرة موضع دراسة الأداء التفاضلي.

$E(A\_j)$: عدد الاستجابات الصحيحة المتوقعة في المجموعة المرجعية على الفقرة موضع دراسة الأداء التفاضلي.

$Var(A\_j)$: تباين الاستجابات الصحيحة الملاحظة في المجموعة المرجعية على الفقرة موضع دراسة الأداء التفاضلي

nR_j: عدد الأفراد المستجيبين على الفقرة موضع اهتمام الأداء التفاضلي في المجموعة المرجعية عند مستوى القدرة *j*

nF_j: عدد الأفراد المستجيبين على الفقرة موضع اهتمام الأداء التفاضلي في المجموعة المركزية أو البؤرية عند مستوى القدرة *j*

T_j: حجم العينة الكلي عند مستوى القدرة j

m_1j: عدد الاستجابات الصحيحة على الفقرة موضع اهتمام الأداء التفاضلي عند مستوى القدرة j

m_0j: عدد الاستجابات الخاطئة على الفقرة موضع اهتمام الأداء التفاضلي عند مستوى القدرة j

(Penfield,2001)، وعادة يتم تقسيم مستويات القدرة حسب المجموع الكلي لاستجابات الأفراد على جميع فقرات المقياس.

ويمكن توسيع المعادلات رقم 1 ورقم 2 لتشمل الفقرات متدرجة الاستجابة. بحيث يتم ترتيب البيانات التي يتم الحصول عليها من استجابات الأفراد على الفقرة متدرجة الاستجابة بجدول توافقي (مصفوفة) من الرتبة 2xTxK ، إذ T تمثل عدد الاستجابات على الفقرة متدرجة الاستجابة، K تمثل عدد مستويات القدرة للأفراد المستجيبين وعادة تمثل القدرة مجموع درجات الأفراد على المقياس، وعند كل مستوى قدرة من مستويات القدرة K يوجد جدول توافقي 2xT، فلنفترض مثلاً أن $m_1, m_2, m_3, \ldots\ldots m_T$ تمثل الدرجات المخصصة لكل استجابة من استجابات الفقرة ، فيكون الجدول التوافقي لمستوى القدرة مثل k كما هو واضح في الجدول رقم 1 الآتي:

الجدول رقم 1: الجدول التوافقي لمستوى القدرة k لفقرة من فقرات مقياس ما.

الدرجات المخصصة لاستجابات الفقرة

المجموعة	m1	m2	m3	.....	m_T	المجموع
المرجعية	nR1k	nR2k	nR3k	.....	nR_Tk	nR+k
البؤرية	nF1k	nF2k	nF3k	.....	nF_Tk	nF+k
المجموع	n+1k	n+2k	n+3k	.....	n+Tk	n++k

حيث:

nRtk: تمثل عدد أفراد المجموعة المرجعية عند مستوى القدرة k الذين حصلوا على الدرجة المخصصة لاستجابة الفقرة وهي m_t

nFtk: تمثل عدد أفراد المجموعة البؤرية عند مستوى القدرة k الذين حصلوا على الدرجة المخصصة لاستجابة الفقرة وهي mt

"+": ترمز إلى المجموع على مستوى الصف أو العمود في المصفوفة التوافقية في الجدول رقم 1

"++": ترمز إلى المجموع الكلي للصفوف أو الأعمدة في المصفوفة التوافقية في الجدول رقم 1

تتعامل طريقة مانتل هنزل العامة مع فئات الاستجابة للفقرة وكأنها بيانات في المستوى الاسمي للقياس، ونظراً لتعدد الاستجابات وفئاتها عند المستويات المختلفة للقدرة لكل فقرة، فإننا سنتعامل مع البيانات على شكل مصفوفات، فتوسيع الصورة العامة للمعادلة رقم 1 في حالة الفقرات متعددة الاستجابة تعطى بالصورة الآتية:

$Q_{GMH}=[\sum D\_k - \sum E(D\_k)]'[\ /\!/\sum V(D\_k)]/\!/\ ^{\wedge}(-1)\ [\sum D\_k - \sum E(D\_k)]$ .....(3)

حيث:

$$D_k = \begin{bmatrix} nR1k \\ nR2k \\ nR3k \\ \dots \\ nR(T-1)k \end{bmatrix} \quad N_k = \begin{bmatrix} n+1k \\ n+2k \\ n+3k \\ \dots \\ n+(T-1)k \end{bmatrix} \quad E(D_k) = (nR+k)N_k/(n+$$

$$+k)$$

$$d_{Nk} = \begin{bmatrix} nR1k & 0 & 0 & 0 \\ 0 & nR2k & 0 & \dots \\ \dots & 0 & \dots & 0 \\ 0 & 0 & 0 & nR(T-1)k \end{bmatrix}$$

$$V(D_k) = (nR+k)(nF+k)\left[\frac{(n++k)d_{NK} - N_k N'_k}{(n++k)^2 - ((n++k)-1))}\right]$$

ويرى الباحثون أن طريقة مانتل هانزل العامة من الطرق الفعالة في الكشف عن الأداء التفاضلي للفقرات متعددة الاستجابة، ويمكن استخدامها – كذلك - في الفقرات ثنائية الاستجابة، وهي أكثر تطوراً وتوسعاً من طريقة مانتل هانزل التقليدية. (Fidalgo & madeira, 2008)

ولكل فقرة متدرجة الاستجابة توجد عتبات بين فئات الاستجابة لها، وتعرف العتبة بأنها عبارة عن حدود تفصل بين فئات الاستجابة المتجاورة، وبالتالي فإن لكل فقرة يوجد عدد عتبات مساوٍ لعدد فئات الاستجابة ناقصًا الواحد (T-1)، ولكل فقرة يوجد معامل تمييز a ويضبط انحدار دالة استجابة الفقرة للفروق في قدرات الطلبة، وحتى نربط مفهوم الأداء التفاضلي مع العتبات والتمييز فإن رفع العتبات أو خفضها مع بقاء الميل(التمييز ثابتاً) يحرك الفقرة أفقياً على متصل القدرة أي التغيير في صعوبة الفقرة، فلذلك تغير العتبات بين المجموعتين البؤرية والمرجعية يولد ما يسمى بالأداء التفاضلي المنتظم، في حين يؤدي تغيير معامل التمييز بين المجموعتين يؤدي إلى الأداء التفاضلي غير المنتظم، لأنه يعمل على خفض أو رفع التفاعل مع قدرات المفحوصين، وبالتالي تباين يعتمد على القدرة للمفحوص.(Cambridge Psychometrics Centre, 2014; Finch, 2022)

إن الكشف عن الأداء التفاضلي في الفقرات قضية مرتبطة بفقرات المقياس كالاختبار، لذا قد يتأثر ببعض المتغيرات مثل حجم العينة، وطول الاختبار، وخصائص الأفراد، وصيغة الفقرات، وغيرها. ومن هنا جاءت هذه الدراسة لتقصي أثر حجم العينة وطول المقياس وتوازن المجموعات على أداء طريقة مانتل–هانزل العامة (GMH) في كشف الأداء التفاضلي للفقرات متدرجة الاستجابة من خلال محاكاة بيانات توليدية تحت شروط معينة.

وعند تقصي الأدب النظري المتعلق بتأثير الظروف والعوامل المختلفة على طرق اكتشاف الأداء التفاضلي،وبالتحديد طريقة مانتل هانزل العامة لاكتشاف الأداء التفاضلي للفقرات متدرجة الاستجابة، نجد قلة في هذا النوع من الدراسات، فغالبية الدراسات كانت تتعلق بطريقة مانتل هانزل لاكتشاف الأداء التفاضلي في الفقرات ثنائية الاستجابة، وفي هذا الصدد أظهرت نتائج دراسة (Su & Wang, 2005) التي قارنت 3 طرق لاكتشاف الأداء التفاضلي في الفقرات متدرجة الاستجابة من خلال بيانات توليدية تحت ظروف مختلفة، منها نسبة وجود الأداء التفاضلي في الفقرات وحجمه على قوة الاختبار وتحديد الخطأ من النوع الأول، أن قيمة الأداء التفاضلي DIF له تأثير أعلى وأهم من نسبة وجوده في الفقرات، وتبين أن طريقة مانتل هانزل العامة هي الأضعف في قوة الاختبار لاكتشاف الأداء التفاضلي بين الطرق الأخرى المستخدمة في الدراسة.

وفي دراسات تأثير حجم العينة على طرق اكتشاف الأداء التفاضلي، نجد أيضاً أن غالبية الدراسات كان محور اهتمامها الفقرات ثنائية الاستجابة، ولا سيما أن معظم الدراسات استخدمت بيانات توليدية، وليست تجريبية حقيقة، فربما كان هذا وراء عدم دراسة تأثير حجوم العينات على طرق اكتشاف الأداء التفاضلي في الفقرات متدرجة الاستجابة، نظراً لصعوبة الحصول على بيانات توليدية تحاكي البيانات الحقيقية لمثل هذا النوع من الفقرات، ففي هذا الصدد نذكر بعض الدراسات التي تناولت فحص فعالية طرق اكتشاف الأداء التفاضلي تحت ظروف اختلاف حجوم العينات.

ففي دراسة توليدية قام بها كاباسكالا وآخرون (Kabasakala et al., 2014) من خلال مقارنة ثلاث طرق هي: طريقة مانتل هانزل، ونسبة الأرجحية لنظرية استجابة الفقرة (IRT - LR)، وطريقة (SIBTEST) تحت تأثير حجوم العينات وتوزيع القدرة، وطول الاختبار، وتأثير ذلك على الخطأ من النوع الأول وقوة الاختبار، وتبين أن هناك تأثير واضح لحجوم العينة خصوصاً على طريقة مانتل هانزل؛ إذ ظهر تغيير واضح في معدل الخطأ من النوع الأول وقوة الاختبار، فبزيادة حجم العينة يقل معدل الخطأ من النوع الأول وتظهر أعلى قوة اختبار.

وفي دراسة الجودة (Aljoudeh, 2021) التي استخدمت بيانات توليدية في الفقرات ثنائية الاستجابة لدراسة أداء طريقة نسبة الأرجحية للنظرية الحديثة في القياس IRT-LR في ظروف مختلفة لحجم العينة وحجم الأداء التفاضلي للفقرات، إذ تم توليد أربعة مستويات لحجم العينة: 250، 500، 750، و1000 تمثل الاستجابات على 40 فقرة ثنائية الاستجابة، وبعض الفقرات أجبرت

على أن تكون فقرات أداء تفاضلي في مستويات مختلفة له وفي حالتين من الأداء التفاضلي المنتظم والأداء التفاضلي غير المنتظم، وتوصلت الدراسة إلى أنه عند عينة حجمها (1000) فرد أظهرت الطريقة نسبة عالية من الأداء في الكشف عن الفقرات ذات الأداء التفاضلي المنتظم لجميع المستويات، في حين انخفض الأداء في الفقرات ذات الأداء التفاضلي غير المنتظم لجميع المستويات وفي جميع أحجام العينات التي تم تناولها.

وأجرى فينش (Finch, 2005) مقارنة بين عدة طرق للكشف عن الأداء التفاضلي هي: طريقة نموذج متعدد المؤشرات ومتعدد الأسباب MIMIC، مانتل هانزل MH، طريقة SIBTEST، ونسبة الأرجحية لنظرية الاستجابة للفقرة (IRT- LR)، في ضوء طول الاختبار، وحجم العينة، وتوزيع القدرة للمفحوصين، ومستوى الأداء التفاضلي وحجمه في الفقرات. وأشارت النتائج إلى أن طرق اكتشاف الأداء التفاضلي تكون أكثر عند أطول اختبار، وحجوم عينة مرتفعة، أو في حالات فقرات ثنائية المعلم، وتقل الفاعلية عندما يقل عدد الفقرات وخصوصاً عندما يكون عدد الفقرات 20 فقرة ثلاثية المعلم.

وأظهرت دراسة وودز (Woods,2009) أفضلية لطريقة نموذج متعدد المؤشرات ومتعدد الأسباب MIMIC على نماذج نظرية استجابة الفقرة في الكشف عن الأداء التفاضلي؛ خصوصًا عند استخدام حجوم العينات الصغيرة وفقرات ثنائية الاستجابة، من خلال دراسة لبيانات توليدية بحجوم عينات وأطوال مختلفة من الاختبار.

وفي دراسة قام بها أوقورلو وأتار (Ugurlu & Atar, 2020) قام بها الباحثان بمقارنة طريقتين في الكشف عن الأداء التفاضلي للفقرات ثنائية الاستجابة، الطريقة الأولى هي طريقة نموذج متعدد المؤشرات ومتعدد الأسباب MIMIC والثانية طريقة الانحدار اللوجستي، من خلال بيانات مولدة في ظروف مختلفة لحجم العينة، وتأثير ذلك على الخطأ من النوع الأول ونسبة الفقرات ذات الأداء التفاضلي، وتوصلت الدراسة إلى أن نسبة الفقرات ذات الأداء التفاضلي تغيرت من 20% إلى 40% عند تغيير حجم العينة من 2000 إلى 4000، إذ إن تأثير حجم العينة كان له أثر واضح في فاعلية الطريقتين في الكشف عن الأداء التفاضلي DIF من خلال انخفاض معدلات الخطأ من النوع الأول.

وأجرى أريكان وآخرون دراسة (Arikan et al., 2016) قارنت أربع طرق للكشف عن الأداء التفاضلي للفقرات، وهي: MIMIC، وSIBTEST، والانحدار اللوجستي LR، ومانتل هانزل MH. تحت ظروف تغيير حجم العينة، إذ تم اختيار عينات فرعية: 300، 600، 1000، 1200، 2000 من

مجموعة بيانات إجمالية بلغ عددها 340000، وتوصلت الدراسة إلى أنه في حجوم العينات المرتفعة مثل 2000 أو أكثر تكون النتائج أكثر فاعلية في اكتشاف الأداء التفاضلي، وتكون الطرق أكثر اتساقاً؛ إذ تكون الطرق الأربع قادرة على اكتشاف الفقرات ذات الأداء التفاضلي نفسها، على خلاف حجوم العينات الصغيرة التي أظهرت اختلافاً في الفقرات ذات الأداء التفاضلي المكتشفة في كل طريقة.

وحديثاً قام عليان، والجودة (Elyan & Aljoudeh,2024) بدراسة هدفت لفحص فعالية طريقة نسبة الأرجحية لاكتشاف الأداء التفاضلي في الفقرات ثنائية الاستجابة لمتغير الجنس في ظروف مختلفة لحجوم العينات، وأطوال الاختبار من خلال بيانات حقيقية تم الحصول عليها من نتائج طلبة الصف العاشر الأساسي في الأردن على اختبار بيزا الدولي عام 2018، وقام الباحثان باختيار 3 مستويات لحجم العينة هي: 342، 200، 100، وثلاثة مستويات لطول الاختبار هي: 30، 20، 10 وتوصل الباحثان إلى أن فعالية طريقة نسبة الأرجحية لمانتل هانزل تزداد في الكشف عن الأداء التفاضلي للفقرات بزيادة حجم العينة مع ثبوت طول الاختبار عند مستوى معين، وتقل فعاليتها بزيادة طول الاختبار عند ثبوت حجم العينة عند مستوى معين، وخلصت الدراسة إلى أن استخدام حجم عينة كبير وطول اختبار قصير تكون فعالية الطريقة أكبر ما يمكن.

مشكلة الدراسة وأهميتها:

تتعدد أدوات القياس والمقاييس النفسية والشخصية والاختبارات تبعاً لاختلاف الغرض الذي تم تطويرها لأجله، فهي تلعب دوراً مهماً في مختلف المجالات، كتصنيف الأفراد، وقبولهم في الجامعات والمؤسسات التعليمية، وتوزيعهم على التخصصات المختلفة، وكذلك تقصي رضاهم عن الخدمات والأنشطة المختلفة وجودتها، وميولهم واتجاهاتهم، واستعداداتهم، مما ينبني عليها قرارات مهمة تساعد في تشخيص أماكن القوة والضعف، وبدورها تساعد في رفع جودة العمل الأكاديمي وعمل المؤسسات التعليمية.

وحتى تؤدي هذه المقاييس وظيفتها على الوجه التي أعدت لأجله، لا بد أن تتوفر فيها الخصائص الضرورية التي تمنحها طابعًا من الصدق والثبات والدقة والعدالة بين المستجيبين، ومن القضايا المهمة والبارزة، التي أصبحت حديثاً محل اهتمام الباحثين والعاملين في تطوير المقاييس هي تحيز الفقرات والأداء التفاضلي لها، وزيادة على ذلك فإن أحدث نسخة لمعايير تطوير الاختبارات النفسية والتربوية عدّت دراسة الأداء التفاضلي للفقرات دليلاً ومؤشراً على الصدق البنائي لها كما ذكرنا سابقاً.

والأداء التفاضلي للفقرات تبعاً لمتغيرات مختلفة كالجنس، والعرق والتخصص، واللغة، والثقافة، وغيرها في المقاييس؛ قد يتأثر بعوامل مختلفة مثل ما تتأثر الخصائص الأخرى كالصدق والثبات ومعالم الفقرات، وغيرها. ومن بين العوامل التي قد تلعب دوراً مهماً في التأثير على الأداء التفاضلي للفقرة هو اختلاف حجوم العينات، وطول المقياس، وعدم التوازن في حجوم العينات بين المجموعة البؤرية والمرجعية، وهذا ما تم ملاحظته من الدراسات التي تم عرضها سابقاً، فقد يلزم حجم عينة مناسب لزيادة قدرة طريقة اكتشاف الأداء التفاضلي للفقرة. إن وجود أداء تفاضلي في فقرات المقياس تبعاً لمتغير ما ينعكس سلباً على المقياس وخصائصه، فوجود المحاباة في فقرات المقياس لفئة دون غيرها يشكك في صحة النتائج لهذا المقياس، وبالتالي صعوبة الثقة في القرارات التي تنطوي عليها. ومن هنا فإن دراسة العوامل التي تؤثر على أداء طرق اكتشاف الأداء التفاضلي لفقرات المقياس تبعاً لمتغير ما، تُعد غاية في الأهمية نظراً لما تعكسه نتائج هذه الدراسة من إيجابيات وتوصيات وإرشادات للباحثين في قضية اختيار أفضل الظروف والمتغيرات التي تناسب الطريقة المستخدمة لأداء عملها بالشكل الدقيق.

ولما كانت طريقة مانتل هانزل العامة GMH طريقة شائعة لدى الباحثين في الكشف عن الأداء التفاضلي للفقرات وخصوصاً في الفقرات ذات الاستجابة المتدرجة، فإن لدراسة هذه الطريقة في ظروف مختلفة قيمة مهمة في الأدب النظري المتعلق بالكشف عن الأداء التفاضلي لفقرات المقاييس.

يفتقر الباحث والممارس إلى دليل إجرائي موجز يبيّن –في سياق متعدد الاستجابات– أثر العوامل التصميمية الأساسية مثل حجم العينة وطول المقياس، التوازن بين حجم العينة بين المجموعات، نسبة الفقرات ذات الأداء التفاضلي، نوع/شدة الأداء التفاضلي على أداء GMH من حيث قوة الاختبار وضبط الخطأ من النوع الأول. إلى جانب ذلك فإن هذه الدراسة تعمل على تعزيز عدالة القياس وصحته عبر تحسين كشف الفقرات ذات الأداء التفاضلي، وتوفير مرجع تطبيقي لتخطيط العينات والأطوال المناسبة للمقاييس، وتبسيط تفسير نتائج GMH في سياق الفقرات متدرجة الاستجابة.

وبالتحديد فإن هذه الدراسة ستجيب عن الأسئلة الآتية:

1. ما أثر نوع الأداء التفاضلي وشدته للفقرة على الخطأ من النوع الأول، وقوة الاختبار الإحصائي لطريقة مانتل هانزل العامة عند ثبوت المتغيرات الأخرى (حجم العينة=600، توازن حجوم

العينات بنسبة 1:1، طول المقياس=20، ونسبة الفقرات التي تحتوي على أداء تفاضلي=20%)؟

2. ما أثر توزان حجوم العينات بين المجموعة البؤرية والمجموعة المرجعية 1:1 مقابل 3:1 عند ثبوت المتغيرات الأخرى (حجم العينة ككل=600، طول المقياس=20، نسبة الفقرات التي تحتوي على أداء تفاضلي=20%، شدة متوسطة من الأداء التفاضلي في كلا النوعين المنتظم وغير المنتظم)؟
3. ما أثر نسبة وجود فقرات ذات أداء تفاضلي (10%، 20%، 30%) عند ثبوت المتغيرات الأخرى (شدة متوسطة وتوازن 1:1 في حجوم العينات لكل من النوعين، حجم عينة 600، وطول مقياس=20)؟
4. كيف يتغير أداء الاختبار للطريقة عند تغيير طول المقياس (10، 40) مع ثبات باقي المتغيرات والعوامل؟
5. ما أثر زيادة حجم العينة من 600 إلى 1000 على متوسطات قوة الاختبار وحجم الخطأ؟

هدف الدراسة:

توفير دليل عملي يوجّه الباحثين في كشف الأداء التفاضلي للفقرات متعددة الاستجابة، من خلال بناء "خريطة أداء" منهجية لطريقة مانتل–هانزل العامة (GMH) عبر محاكاةٍ توليدية مضبوطة تفترض تكافؤ توزيع القدرة بين المجموعتين البؤرية والمرجعية. وتتمثل من خلال توصيف كيف تتغيّر القوة الإحصائية والخطأ من النوع الأول عند تغيير بعض العوامل مثل: حجم العينة، طول المقياس، توازن المجموعات، نسبة الفقرات ذات الأداء التفاضلي، ونوع الأداء التفاضلي وشدته ضمن إطار فقرات متعددة الاستجابة مُولَّدة وفق نموذج الاستجابة المتدرجة (Graded (GRM) Response Model).

إجراءات الدراسة:

اعتمدت هذه الدراسة بيانات توليدية باستخدام برنامج R، بأسلوب محاكاة مونتِ كارلو Monte Carlo method لنموذج الاستجابة المتدرجة (GRM) مع تغيير عوامل تصميمية أساسية: حجم العينة، طول الاختبار، توازن أحجام المجموعتين البؤرية والمرجعية، نسبة الفقرات ذات الأداء التفاضلي، ونوع الأداء التفاضلي وشدّته بهدف قياس أداء GMH تحت شروط يمكن التحكم بها بدقة. يتيح لنا التوليد ضبط العوامل الأساسية (حجم العينة، طول المقياس، توازن المجموعات،

نسبة الفقرات ذات الأداء التفاضلي، ونوع الأداء التفاضلي وشدته) وتغيير كل عامل على حدة، مع امتلاكنا لحقيقة معروفة مسبقاً (أي نعرف الفقرات ذات الأداء التفاضلي مسبقاً) لقياس القوة والخطأ من النوع الأول بلا تحيز. كما أنه يوفر الوقت والتكلفة، ويسهّل إعادة الإنتاج وتكرار التجارب، وبهذه الطريقة نقدّم خريطة أداء دقيقة يمكن الاستناد إليها عند الانتقال لاحقًا إلى البيانات الواقعية.

وصف البيانات المولدة:

في ضوء هدف الدراسة التي تسعى إلى تقييم أداء طريقة مانتل–هانزل العامة (GMH) في كشف الأداء التفاضلي للفقرات متعددة الاستجابة تحت شروط تصميمية متعدّدة، فقد اعتمد الباحث بيانات توليدية (محاكاة احتمالية) صُمِّمت بعناية لتوفير ضبطٍ كامل لعوامل التصميم وعزل أثرها عن أي فروق حقيقية في القدرة بين المجموعتين عمليًا، جرى أولًا تحديد هيكل المقياس على هيئة فقرات متدرجة الاستجابة بخمس فئات مرتّبة (0–4) (مقياس اتجاهات مثلاً)، ثم توليد القدرة الكامنة لكل مفحوص في كل مجموعة من توزيع طبيعي معياري بمتوسط صفر وانحراف معياري 1، بحيث تفترض الدراسة تكافؤ توزيع القدرة بين المجموعتين البؤرية والمرجعية من البداية. ثم تم توليد معالم التمييز للفقرات لتحقيق توزيع طبيعي لوجستي بمتوسط صفر وانحراف معياري قريب من الواحد، لضمان قيم موجبة وتمييز واقعي.

ثم توليد عتبات فئات الاستجابة بتوزيع طبيعي بمتوسط صفر وانحراف معياري واحد تُرتَّب تصاعديًا لضمان الرتابة بين حدود فئات الاستجابة للفقرة الواحدة. وتم زرع فقرات بأداء تفاضلي بنسب (10%، 20%، 30%) وبنوعين مختلفين منتظم وغير منتظم، إذ تم إجراء إزاحة ثابتة (0.25، 0.5، 1) في عتبات الاستجابة للفقرات ذات الأداء التفاضلي بالنسب المذكورة في المجموعة البؤرية وذلك بهدف توليد أداء تفاضلي منتظم، وذلك للتعبير عن شدته، إذ إن الإزاحة 0.25 تمثل أداء تفاضليًا ضعيفًا، و0.5 أداء تفاضليًا متوسطًا، و1 أداء تفاضليًا كبيرًا)، وذلك مع بقاء معاملات التمييز للفقرات ثابتة، ولخلق الأداء التفاضلي غير المنتظم تم زيادة معاملات التمييز للفقرات ذات الأداء التفاضلي المحددة بالنسب مسبقاً بزيادات مقدارها، 0.2، 0.5، 0.8 وذلك للتعبير عن شدة الأداء التفاضلي مع بقاء عتبات فئات الاستجابة للفقرات ثابتة.

بعد توليد القدرة الكامنة للأفراد ومعاملات الفقرات، تم تحويل هذه القيم إلى احتمالات اختيار فئات الاستجابة وفق نموذج الاستجابة المتدرجة (GRM): وبحسب احتمال أن تكون إجابة

الشخص في فئةٍ ما أو أعلى، ثم تستخلص احتمال كل فئة من فروق تلك الاحتمالات. وبذلك نحصل على توزيعٍ احتمالي لخمس فئات لكل زوج (مفحوص، فقرة)، ومن هذا التوزيع تُسحب الاستجابة عشوائيًا. تُكرَّر الخطوات عبر جميع خلايا التصميم: طول المقياس (10، 20، 40) وتوازن المجموعات (1:1، 3:1) ونسبة الفقرات ذات الأداء التفاضلي (10%، 20%، 30%) ونوع الأداء التفاضلي (منتظم، غير منتظم) وشدة الأداء التفاضلي (ضعيف، متوسط، كبير) في ظل اعتماد حجم عينة 600، ومن ثم توسيعه إلى 1000 مفحوص. وتم إعادة التوليد عدة مرات (50 مرة)، وذلك لضمان الحصول على متوسطات وانحرافات معيارية دقيقة لقوة الاختبار والخطأ. وبهذا تم الحصول على قاعدة بيانات طويلة تحتوي على رقم الطالب، ومجموعته، والفقرة، والاستجابة، من خلال دمج استجابات المجموعتين، ومن دون وجود قيم مفقودة، تمهيدًا لتكوين جداول (مجموعة × فئات) لكل فقرة وإجراء اختبار كاي-تربيع كما سَتُفصّل في إجراءات التحليل الآتية.

إجراءات التحليل:

تم استخدام اختبار كاي-تربيع للاستقلالية لاختبار GMH في برنامج R من حزمة difR في R عبر(*stats::chisq.test*) (R Core Team, 2025) ثم جمعت قيم الدلالة الإحصائية للفقرات بعد تصحيحها بطريقة Benjamini–Hochberg (Benjamini, & Hochberg,1995) ويتم مقارنة قيم الدلالة بمستوى الدلالة الإحصائية 0.05، فإذا كانت أقل تكون الفقرة ذات أداء تفاضلي، ومن خلال التكرارات تكون نسبة الفقرات ذات الأداء التفاضلي الفعلي(التي تم زرعها مسبقاً) هي قوة الاختبار، ونسبة الفقرات التي تم اعتبارها ذات أداء تفاضلي بناءً على الدلالة الإحصائية، وهي في الواقع لم تكن من ضمن الفقرات الفعلية(التي تم زرعها مسبقاً) تمثل الخطأ من النوع الأول.

النتائج:

النتائج المتعلق بسؤال الدراسة الأول:

ويبحث السؤال الأول في تقصي أثر نوع الأداء التفاضلي وشدته في الفقرة على الخطأ من النوع الأول وقوة الاختبار الإحصائي لطريقة مانتل هانزل العامة (GMH) عند ثبوت المتغيرات الأخرى (حجم العينة=600، توازن حجوم العينات بنسبة 1:1، طول المقياس=20، ونسبة الفقرات التي تحتوي على أداء تفاضلي=20%)، وقد تبين من نتائج التحليل أنه كلما زادت شدة الأداء التفاضلي يتحسن كشفه، كما تظهر GMH حساسية أعلى في المنتظم مقارنةً بـ غير المنتظم لأن الإزاحة الأفقية للعتبات تُنتج فرقًا منتظمًا عبر القدرة، بينما يتطلّب غير المنتظم تفاعلًا مع القدرة قد لا يتم كشفه

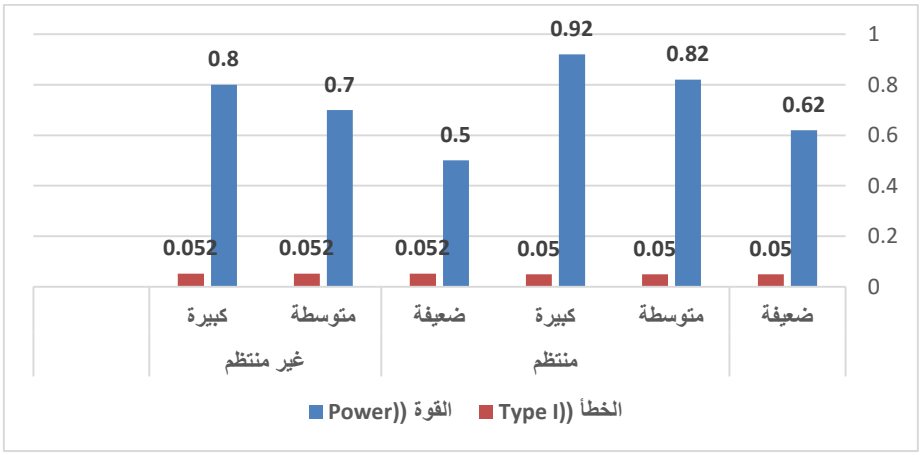
بالقوة نفسها. والجدول رقم 1 يظهر متوسط القوة والخطأ حسب النوع والشدة للأداء التفاضلي عند حجم عينة 600 وطول مقياس 20.

جدول(1) .

متوسط القوة والخطأ حسب النوع والشدة للأداء التفاضلي عند حجم عينة 600 وطول مقياس 20 فقرة.

النوع	الشدة	القوة (Power)	الخطأ (Type I)
	ضعيفة	0.62	0.050
منتظم	متوسطة	0.82	0.050
	كبيرة	0.92	0.050
	ضعيفة	0.50	0.052
غير منتظم	متوسطة	0.70	0.052
	كبيرة	0.80	0.052

والشكل رقم1 يوضح كيفية تغير مستوى القوة والخطأ في الأداء التفاضلي باختلاف شدة الأداء التفاضلي في الأداء التفاضلي المنتظم وغير المنتظم.



الشكل رقم (1): متوسطات القوة والخطأ باختلاف شدة الأداء التفاضلي المنتظم

من الجدول والشكل رقم 1 يتبين لنا أن متوسط القوة يرتفع مع تغيير شدة الأداء التفاضلي في كلا النوعين المنتظم وغير المنتظم، مع ثبات في متوسط الخطأ من النوع الأول تقريباً، إلا أنه يلاحظ أن الأداء التفاضلي المنتظم أكثر حساسية للشدة منه في غير المنتظم.

النتائج المتعلق بسؤال الدراسة الثاني:

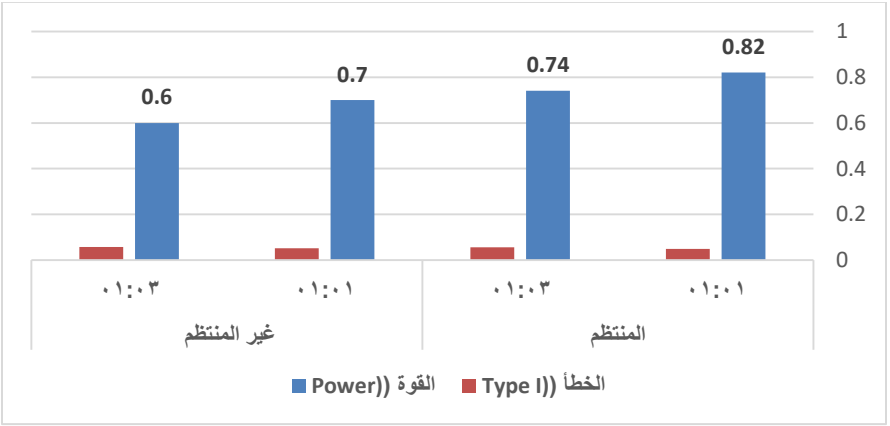
ويتعلق سؤال الدراسة الثاني بتقصي أثر توزان حجوم العينات بين المجموعة البؤرية والمجموعة المرجعية 1:1 مقابل 1:3 عند ثبوت المتغيرات الأخرى(حجم العينة ككل=600، طول المقياس=20، نسبة الفقرات التي تحتوي على أداء تفاضلي=20%، شدة متوسطة من الأداء التفاضلي في كلا النوعين المنتظم وغير المنتظم)، وتبين أن عدم التوازن يقلّل مستوى القوة وخاصة في غير المنتظم مع زيادة طفيفة في الخطأ من النوع الأول، والجدول رقم 2 يظهر متوسطات القوة والخطأ من النوع الأول باختلاف التوازن في حجم العينة بين المجموعتين البؤرية والمرجعية عند حجم عينة 600 ككل وطول مقياس 20.

جدول(2).

متوسط القوة والخطأ باختلاف التوازن في حجم العينة بين المجموعتين البؤرية والمرجعية عند حجم عينة 600 ككل وطول مقياس 20.

النوع	التوازن	القوة (Power)	الخطأ (Type I)
المنتظم	1:1	0.82	0.050
	1:3	0.74	0.056
غير المنتظم	1:1	0.70	0.052
	1:3	0.60	0.058

والشكل رقم 2 يوضح كيفية تأثير اختلاف التوازن في حجم العينة بين المجموعتين البؤرية والمرجعية على متوسطات القوة والخطأ.



الشكل رقم 2: اختلاف التوازن في حجم العينة بين المجموعتين البؤرية والمرجعية على متوسطات القوة والخطأ.

يتبين لنا أن عدم التوازن يضعف القوة بوضوح ويؤثر أكثر في غير المنتظم؛ ويظهر ارتفاع طفيف في الخطأ من النوع الأول Type I

النتائج المتعلق بسؤال الدراسة الثالث:

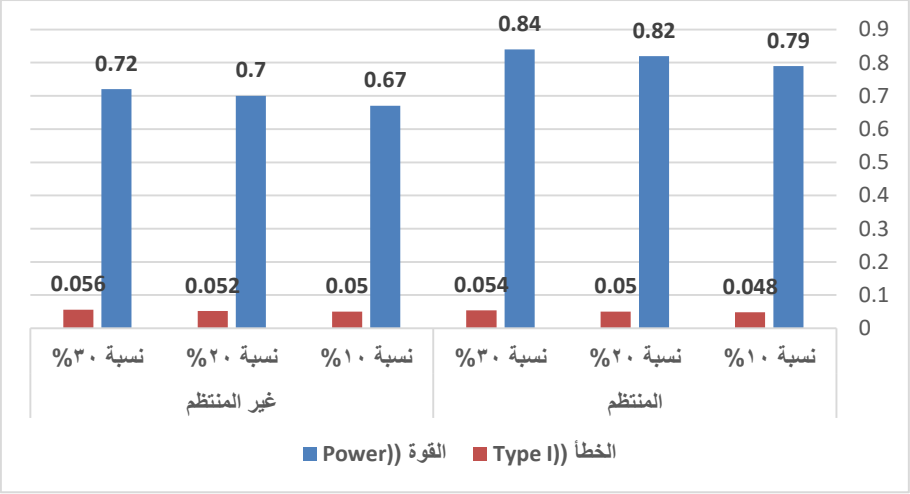
ويبحث السؤال الثالث عن أثر نسبة وجود الفقرات ذات الأداء التفاضلي (10%، 20%، 30%) عند شدة أداء تفاضلي متوسطة وتوازن حجم العينتين.

جدول(3)

يظهر متوسطات القوة والخطأ حسب نسبة الأداء التفاضلي عند حجم عينة 600 وطول مقياس 20.

النوع	نسبة الأداء التفاضلي	القوة (Power)	الخطأ (Type I)
المنتظم	10%	0.79	0.048
	20%	0.82	0.050
	30%	0.84	0.054
غير المنتظم	10%	0.67	0.050
	20%	0.70	0.052
	30%	0.72	0.056

والشكل رقم 3 يوضح تأثير اختلاف نسبة الفقرات ذات الأداء التفاضلي على متوسطات القوة والخطأ من النوع الأول.



الشكل رقم 3 اختلاف نسبة الأداء التفاضلي على متوسطات القوة والخطأ من النوع الأول.

من الجدول والشكل رقم 3 يتبين لنا أن القوة تتحسّن تدريجيًا مع زيادة نسبة الفقرات ذات الأداء التفاضلي، بينما يظل الخطأ من النوع الأول Type I قريبًا من 0.05 مع زيادة طفيفة عند 30%.

النتائج المتعلقة بسؤال الدراسة الرابع:

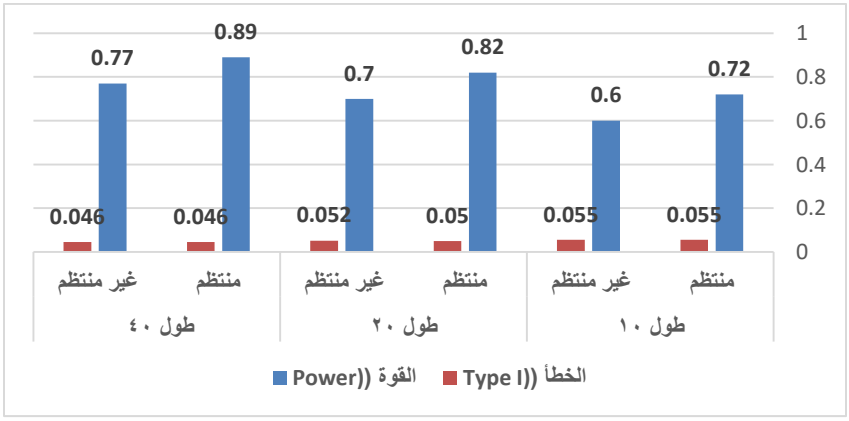
ويتعلق السؤال الرابع بكيفية التغير في متوسطات القوة والخطأ تبعاً لاختلاف طول المقياس أي بزيادة طول المقياس من10 ثم إلى 20 فقرة ومن ثم إلى 40 فقرة عند ثبوت العوامل والمتغيرات الأخرى.

تبين أن المقياس الأطول يوفّر معلومات أكثر واستقرارًا أعلى في الجداول (تكرارات أكبر)، فتزداد القوة ويقترب الخطأ Type I من 0.05 أو ينخفض قليلًا، والجدول والشكل 4 يوضحان ذلك.

جدول(4)

اختلاف طول الاختبار على متوسطات القوة والخطأ عند ثبوت العوامل الأخرى عند حجم عينة 600 وتوازن حجم العينة ونسبة أداء تفاضلي 20%

الطول	النوع	القوة (Power)	الخطأ (Type I)
10	منتظم	0.72	0.055
	غير منتظم	0.60	0.055
20	منتظم	0.82	0.050
	غير منتظم	0.70	0.052
40	منتظم	0.89	0.046
	غير منتظم	0.77	0.046



الشكل رقم 4: اختلاف طول الاختبار على متوسطات القوة والخطأ من النوع الأول.

يلاحظ أن إطالة المقياس ترفع القوة لكلا النوعين ومردّه زيادة المعلومات وتثبيت تقديرات الجداول مع خطأ Type I أدنى قليلًا عند طول اختبار 40.

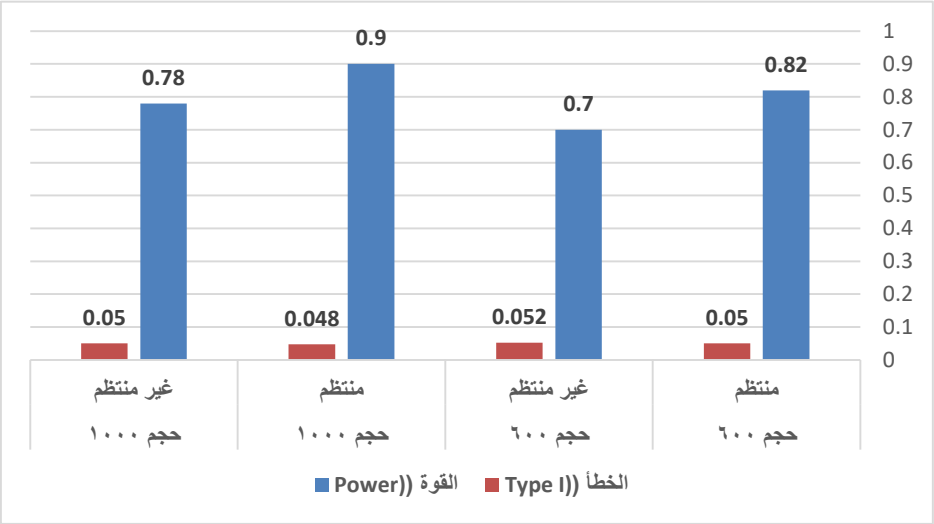
النتائج المتعلقة بسؤال الدراسة الخامس:

ويتعلق سؤال الدراسة الخامس بأثر زيادة حجم العينة من 600 إلى 1000 على متوسطات قوة الاختبار وحجم الخطأ، فقد أشارت النتائج أن توسيع العيّنة يزيد دقّة التكرارات في الجداول ويُحسّن قوة الاختبار، مع إبقاء الخطأ Type I تحت السيطرة، كما يظهر في الجدول والشكل رقم 5.

الجدول (5)

مقارنة اختلاف حجم العينة على متوسطات القوة والخطأ عند ثبوت العوامل الأخرى عند طول اختبار 20، وتوازن حجم العينة وشدة متوسطة، ونسبة أداء تفاضلي 20%.

حجم العينة N	النوع	القوة (Power)	الخطأ (Type I)
600	منتظم	0.82	0.050
600	غير منتظم	0.70	0.052
1000	منتظم	0.90	0.048
1000	غير منتظم	0.78	0.050



الشكل رقم 5: تأثير اختلاف حجم العينة على متوسطات القوة والخطأ من النوع الأول.

يتبين أن زيادة حجم العينة يزيد من قوة الاختبار وخصوصاً عند الأداء التفاضلي المنتظم، مع سيطرة - نوعاً ما - على الخطأ من النوع الأول.

مناقشة النتائج والاستنتاجات والتوصيات:

مناقشة النتائج المتعلقة بسؤال الدراسة الأول:

ويهدف السؤال الأول إلى تقصي أثر نوع الأداء التفاضلي وشدته على متوسطات القوة والخطأ من النوع الأول، فقد أظهرت النتائج أن قوة الاختبار ترتفع مع ازدياد شدة الأداء التفاضلي في كلا النوعين، مع تفوّق مستمر للمنتظم على غير المنتظم. يتسق ذلك مع التمييز المفاهيمي الكلاسيكي بين المنتظم وغير المنتظم؛ إذ يُعزى الأول غالبًا إلى إزاحة العتبات، ويُنتج فروقًا شبه ثابتة عبر القدرة، بينما يعكس الثاني اختلاف التمييز وتفاعلًا مع القدرة (Ackerman, 1992; Mellenbergh, 1989; Millsap & Everson, 1993; Narayanan & Swaminathan, 1996). كما أن استخدام GMH للفقرات متدرجة الاستجابة مدعوم تطبيقياً ومنهجيًا (Penfield, 2001; Fidalgo & Madeira, 2008).

ورغم أن مقارنة ثلاث طرق في سياق الفقرات متعددة الاستجابة أشارت إلى أن GMH كانت الأضعف نسبيًا في القوة ضد بدائل أخرى (Su & Wang, 2005)، فإن نمط التأثر بالشدة والتمييز الذي رصدته نتائج هذه الدراسة يظل متسقًا مع هذا الإطار النظري. بقاء الخطأ من النوع الأول Type I حول 0.05 يعكس جدوى تصحيح قيم الدلالة الإحصائية في ظل تعدد الفقرات (Benjamini & Hochberg, 1995).

مناقشة النتائج المتعلقة بسؤال الدراسة الثاني:

والذي يبحث في تأثير اختلاف التوازن في حجم العينة بين المجموعة البؤرية والمرجعية، فقد توصلت الدراسة إلى أن اختلاف التوازن في حجم العينة هبوطًا ملحوظًا في متوسط القوة مع عدم التوازن، وبصورة أشد في غير المنتظم، مع ارتفاع طفيف في الخطأ من النوع الأول، هذا الاتجاه متوافق مع الأدب الذي أبرز حساسية إجراءات الكشف لحجم العيّنة وتوزيع القدرة وطول الاختبار (Kabasakala et al., 2014؛ Finch, 2005). كما أن شواهد أخرى على طرق مختلفة تؤكد تحسن الأداء والاتساق مع العينات الأكبر (Arikan et al., 2016, Aljoudeh, 2021؛ Ugurlu & Atar, 2020).

مناقشة النتائج المتعلقة بسؤال الدراسة الثالث:

ويتعلق بتقصي تأثير وجود فقرات ذات أداء تفاضلي بنسب مختلفة على متوسط قوة الاختبار والخطأ من النوع الأول، فقد توصلت الدراسة إلى أن متوسط القوة يتحسن تدريجياً بزيادة نسبة الفقرات ذات الأداء التفاضلي، مع بقاء الخطأ من النوع الأول حول 0.05 مع ملاحظة زيادة طفيفة عليه عند وجود أعلى نسبة أداء تفاضلي للفقرات، وهذا متوقع بسبب تعدد وتكرار الاختبارات(Benjamini & Hochberg, 1995).

وعلى الرغم من أن الدراسات المعروضة ركّزت أكثر على طول الاختبار وحجم العينة وتوزيع القدرة، فإن النتيجة هنا منطقية منهجيًا وتنسجم مع فكرة تحسّن قابلية الطريقة في كشف الفقرات ذات الأداء التفاضلي كلما زادت نسبتها ضمن المقياس.

مناقشة النتائج المتعلقة بسؤال الدراسة الرابع:

ويدرس السؤال الرابع اختلاف طول المقياس وتأثير ذلك على متوسط القوة والخطأ من النوع الأول، فلقد توصلت النتائج إلى أن إطالة المقياس ترفع القوة وتُحسّن الاستقرار في الخطأ من النوع الأول وهذا يتوافق مع ما توصلت إليه مقارنات طرق متعددة بشأن أفضلية الاختبارات الأطول للحساسية (Finch, 2005). بالمقابل، أفادت دراسة حديثة ببيانات حقيقية وقياس ثنائي الاستجابة باستخدام MH-LOR بأن الفاعلية تكون أكبر مع عينة كبيرة وطول اختبار قصير (Elyan & Aljoudeh, 2024). يمكن تفسير التباين باختلاف الإجراء (GMH مقابل MH-LOR)، صيغة الاستجابة (متعددة الفئات مقابل ثنائية)، وطبيعة البيانات (توليدية مضبوطة مقابل بيانات ميدانية). عمليًا، تبدو إطالة المقياس مفيدة في سياق GMH متعدد الفئات، مع التنبه لإمكان اختلاف السلوك عند تبنّي إجراءات وطرق مختلفة.

مناقشة النتائج المتعلقة بسؤال الدراسة الخامس:

ويبحث السؤال الخامس في تأثير رفع حجم العينة على متوسط قوة الاختبار ومتوسط الخطأ من النوع الأول، فلقد تبين أن متوسط القوة يتحسن في كلا النوعين (المنتظم، غير المنتظم) مع بقاء الخطأ من النوع الأول Type I قريبًا من 0.05، وهو اتجاه موثّق على نطاق واسع: العينات الأكبر ترفع الحساسية وتُحسّن الاتساق بين الطرق (Finch, 2005؛ Kabasakala et al., 2014؛ Arikan et al., 2016؛ Ugurlu & Atar, 2020)، وتؤكد البيانات الواقعية بالـ MH-LOR تحسّن الأداء مع تكبير العينة (Elyan & Aljoudeh, 2024).

تؤكد المعايير الحديثة للاختبارات (AERA/APA/NCME, 2014) على أهمية وضرورة فحص الأداء التفاضلي كدليل على الصدق البنائي والعدالة للمقاييس؛ وتُظهر نتائج هذه الدراسة أن ضبط العوامل التصميمية (حجم العينة، الطول، التوازن) تسهم في قراءات أكثر موثوقية لنتائج طريقة GMH.

الاستنتاجات:

من خلال نتائج هذه الدراسة، فإنه يمكن التوصل إلى ما يأتي:

1. شدة الأداء التفاضلي في الفقرات متعددة الاستجابة تلعب دوراً مهماً في رفع متوسط قوة الاختبار في كلا نوعي الأداء التفاضلي المنتظم وغير المنتظم في طريقة مانتل هانزل العامة GMH
2. عدم توازن حجوم العينات بين المجموعة البؤرية والمجموعة المرجعية يقلل متوسط قوة الاختبار ويرفع من تذبذب الخطأ من النوع الأول في طريقة مانتل هانزل العامة GMH
3. إطالة عدد فقرات المقياس يحسن من متوسط القوة للاختبار ويساعد على استقرار الخطأ من النوع الأول في طريقة مانتل هانزل العامة GMH
4. تكبير العينة إلى 1000 يعزز الحساسية دون تكلفة تُذكر على Type I

التوصيات:

1. إذا كان من المتوقع وجود أداء تفاضلي غير منتظم في فقرات المقياس، فيستحسن استخدام حجوم عينات أكبر أو يساوي 1000
2. من الأفضل تجنب عدم التوازن في حجم العينة بين المجموعة البؤرية والمجموعة المرجعية عند إجراء دراسة كشف الأداء التفاضلي.
3. يفضل استخدام طول مقياس أكبر من 20 في الفقرات متعددة الاستجابة.
4. عند الاشتباه بغير المنتظم أو في العينات الصغيرة/غير المتوازنة، يُستحسن تدعيم GMH بطريقة بديلة في الأدب (مثل IRT-LR أو MIMIC) وفق ما أظهرته المقارنات (Finch, 2005؛ Su & Wang, 2005؛ Woods, 2009).

المراجع: References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Aljodudeh, M. (2021). Item response theory likelihood ratio test performance for deducting DIF items in different levels in samples sizes and different levels of DIF items. *Vidyabharati International Interdisciplinary Research Journal 13* (1), 392-399

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing. Washington,* DC: American Educational Research Association.

Arikan, C. A., Ugurlu S. , & Atar, B. (2016). A DIF and Bias Study by using MIMIC, SIBTEST, Logistic *Regression and Mantel-Haenszel Methods. Journal of Education, 31*(1), 34-52.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: *Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cambridge Psychometrics Centre. (2014). Session 4: Overview of polytomous IRT models (GRM thresholds & discrimination).

Elyan, R. M. ., & Al jodeh, M. M. . (2024). The Effectiveness of Mantel Haenszel Log Odds Ratio Method in Detecting Differential Item Functioning Across Different Sample Sizes and Test Lengths Using Real Data Analysis. Dirasat: *Educational Sciences, 51*(3), 37–46. https://doi.org/10.35516/edu.v51i3.6755

Eom, M. (2008). Underlying factors of MELAB listening construct. S*paan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 77–94.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*(6), 940-958

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Finch, W. H. (2022). *The Impact and Detection of Uniform Differential Item Functioning*. Frontiers in Education (PMC).

Holland, P. W., & Thayer, D. T. (1988). *Differential item performance and the Mantel–Haenszel procedure*. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056905-12

Jafari, P., Bagheri, Z., Hashemi, S. Z., & Shalileh, K. (2013). Assessing whether parents and children perceive the meaning of the items in the PedsQLTM 4.0 quality of life instrument consistently: a differential item functioning analysis. *Global Journal of Health Science, 5*(5), 80 – 88.

Kabasakala, K., Arsan, N., Gok, B., & Kelecooglu, H. (2014). Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. Educational Sciences: *Theory & Practice, 14*(6), 2186-2193.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research, 13*(2), 127-143.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement, 17*(4), 297-334. https://doi.org/10.1177/014662169301700401

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied psychological measurement, 20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Park, G.(2008). Differential Item Functioning on an English Listening Test across Gender. *TESOL Quarterly,42*(1), pp. 115-123

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: a comparison of three Mantel-Haenszel procedures. *Appl. Meas. Educ. 14,*(3) 235–259. doi: 10.1207/ S15324818AME1403_3

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomously scored items. *Journal of Educational Measurement, 47*(1), 129–149. https://doi.org/10.1111/j.1745-3984.2010.00105.x

Penfield, R. D., Gattamorta, K. A., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38–49. https://doi.org/10.1111/j.1745-3992.2009.01135.x

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.4.3). R Foundation for Statistical Computing. https://www.r-project.org/

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, Generalized Mantel–Haenszel, and Logistic Discriminant Function Analysis Methods in Detecting Differential Item Functioning for Polytomous Items. *Applied Measurement in Education, 18*(4), 313–350. https://doi.org/10.1207/s15324818ame1804_1

Thissen, D. (2001). IRTLRDIF v.2.0b: *Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for Differential Item Functioni*ng. L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC.

Ugurlu, S. & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 1-12.

Vahid A., Christine C. & Lee O. (2011). *An Investigation of Differential Item Functioning in the MELAB Listening Test.* Language Assessment Quarterly, 8, 361–385. DOI:10.1080/15434303.2011.628632

Wagner, A. (2004). *A construct validation study of the extended listening sections of the ECRE and MELAB*. Spaan Fellow Working Papers in Second or Foreign Language Assessment, 2, 1–23.

Woods, C. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.