

Effectiveness of Multiple Indicators and Multiple Causes Model (MIMIC) in Detecting the Differential Item Functioning of a Test in Terms of Sample Size and Test Length

فاعلية نموذج المؤشرات المتعددة والأسباب المتعددة (MIMIC) في الكشف عن الأداء التفاضلي في فقرات اختبار بدلالة حجم العينة وطول الاختبار

Reem Elyan^{1*}, Yousef Sawalmeh¹.

¹Yarmouk University, Irbid, Jordan.

ARTICLE INFO

Article history:

Received 20 May 2022

Accepted 28 Jun 2022

Published 01 Jul 2024

*Corresponding author:

Yarmouk University, Irbid, Jordan.

Email: reem.elyan@hotmail.com.

Abstract

This study aimed to examine the effectiveness of the Multiple Indicators Multiple Causes Model (MIMIC) in detecting differential item functioning of a test in relation to sample size and test length. Three random samples were used in the study, consisting of 10th-grade students in Jordan who took the PISA 2018 mathematics test, with varying sizes of 2000, 3000, and 4000 students. Additionally, three test lengths were used: 10 items, 20 items, and 30 items. The Mplus program was utilized for data analysis according to the MIMIC method to detect differential item functioning for gender. The results indicated that there was no statistically significant difference ($\alpha=0.05$) in the percentage of items with differential item functioning detected due to different sample sizes across the different test lengths. Similarly, there was no statistically significant difference ($\alpha=0.05$) in the percentage of items with differential item functioning detected due to different test lengths across the different sample sizes.

Keywords: Differential Item Functioning, PISA 2018 Test, Multiple Indicators Multiple Causes Model, MIMIC.

الملخص

هدفت هذه الدراسة إلى فحص فاعلية نموذج المؤشرات المتعددة والأسباب المتعددة (MIMIC) في الكشف عن الأداء التفاضلي في فقرات اختبار في ضوء اختلاف حجم العينة وطول الاختبار. ولأغراض الدراسة تم استخدام ثلاث عينات عشوائية من طلبة الصف العاشر الأساسي في الأردن الذين تقدموا لاختبار بيزا ٢٠١٨ في الرياضيات بحجوم مختلفة هي: ٢٠٠٠، ٣٠٠٠، ٤٠٠٠. كما تم استخدام ٣ أطوال للاختبار هي: ١٠ فقرات، ٢٠ فقرات، ٣٠ فقرات. واستخدم برنامج Mplus لتحليل البيانات وفق طريقة MIMIC للكشف عن الأداء التفاضلي للجنس. ولم تظهر النتائج وجود فروق دالة إحصائية ($\alpha=0.05$) في نسبة الفقرات ذات الأداء التفاضلي المكتشفة باختلاف حجم العينة في الأطوال المختلفة للاختبار.

كذلك لم تظهر النتائج وجود فروق دالة ($\alpha=0.05$) إحصائياً في نسبة الفقرات ذات الأداء التفاضلي المكتشفة باختلاف طول الاختبار في الأحجام المختلفة للعينة.

الكلمات المفتاحية: الأداء التفاضلي، اختبار بيزا ٢٠١٨، نموذج المؤشرات المتعددة والأسباب المتعددة.

١. المقدمة

١.١ الإطار النظري والدراسات السابقة

لا بد للاختبارات التي يبنى على نتائجها قرارات حاسمة أن تتصف بالعدالة والمساواة، وهذا يستدعي المراجعة المستمرة لفقرات الاختبار؛ لتحديد ما إذا كان الاختبار واستخدامه عادلاً بشكل متساوٍ لمجموعات مختلفة من الأفراد. من هنا كان لابد من الاهتمام ببعض القضايا ذات الصلة بالاختبار، مثل الأداء التفاضلي للفقرات (DIF) والذي يمكن أن يؤثر وجوده سلباً على استدلالات الاختبار وتصنيف الأفراد. لذلك قامت العديد من الدراسات بالتحقق من (DIF) في الاختبارات التحصيلية والمقاييس المتنوعة مثل اختبار اللغة الانجليزية TOEFL، واختبار القدرات (SAT) على سبيل المثال (Wall & Horak, 2008).

وتُعد الاختبارات الدولية من أبرز أدوات تقويم الأنظمة التعليمية والتربوية في دول العالم، وتُشرف عليها منظمات عالمية كمنظمة التعاون الاقتصادي والتنمية (OECD) والمنظمة الدولية لتقويم التحصيل التربوي (IEA). ومن الأمثلة على الاختبارات الدولية برنامج التقييم الدولي (PISA 2018)، والذي يقيس الأداء التعليمي على مقياس معياري عالمي. ويشارك الأردن في هذا البرنامج منذ عام ٢٠٠٦. وأظهرت نتائج الأردن في دورة ٢٠١٨ تحسناً جوهرياً في متوسط أداء الطلبة مقارنة بالدورات الماضية للاختبار.

وبالرغم من التحسن الذي أظهرته نتائج الأردن في دورة (PISA 2018) إلا أنها ما زالت ذات ترتيب متدني مقارنة بغيرها وتقل عن الوسط الحسابي العالمي. وتجدر الإشارة إلى أنَّ هذا الاختبار يتم إعداده من قبل مختصين أجانب من الدول المنتمية للمنظمة العالمية (OECD). وتكون الصورة الأصلية للاختبار باللغة الانجليزية، ومنها تُترجم إلى اللغات الأخرى. وأمام هذا الوضع فإنه من الممكن أن تتأثر الأسئلة بالمناهج الدراسية للدول التي ينتمي لها المختصون وواضعو الأسئلة، وقد يتأثرون بالثقافات السائدة في دولهم، وبالتالي من المحتمل أن تكون هذه الفقرات محابية لفئات معينة من الطلبة، وتُظهر أداءً تفاضلياً عبر المجموعات المختلفة من الطلبة (المركز الوطني للموارد البشرية، ٢٠٢٠). ومن المحتمل أن يتأثر الأداء التفاضلي في الفقرات بعوامل متعددة، مثل: حجم العينة، وطول الاختبار التي هي محور اهتمام الدراسة الحالية.

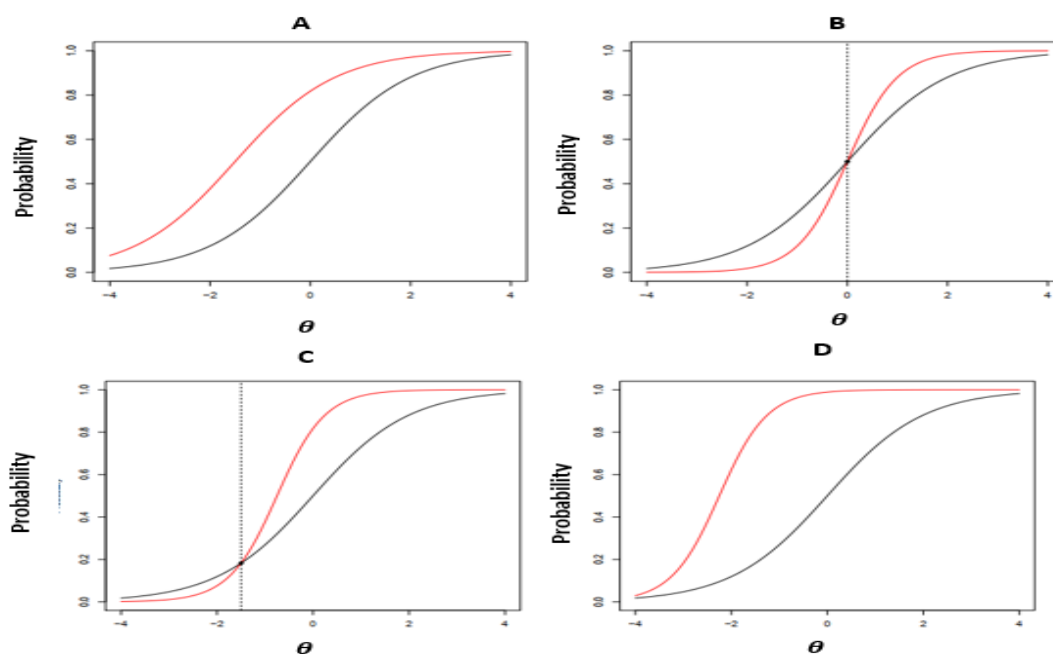
وغالباً ما تكون تعريفات الأداء التفاضلي للفقرة (DIF) متماثلة؛ فقد عرّف كاميلي وشيبارد (Camilli & Shepard, 1994) الأداء التفاضلي على أنه دالة مشتقة إحصائياً تعبر عن الفرق في استجابة فقرة بين مجموعتين عند مستوى القدرة نفسها. ويعرفه أوسترلند (Osterlind, 1983) على أنه الفرق في الأداء بين المجموعتين المستهدفة (FG)، والمعيارية (RG) عند مستويات القدرة نفسها. ويعرفه هاملتون (Hamilton, 1999) على أنه الفروق في احتمالات الإجابة الصحيحة للفقرة في المجموعات المختلفة والمتساوية في القدرة.

وميز ويليمز (Williams, 1997) بين الأداء التفاضلي DIF والتحيز، فبين أن الفقرة تحتوي على DIF إذا كان لدى المفحوصين ذوي القدرة المتساوية من مجموعات فرعية مختلفة قيم احتمالية غير متكافئة للاستجابة بشكل صحيح لفقرة ما، بينما تكون الفقرة متحيزة إذا كان من الممكن القول بأن الفقرة تقيس سمات غير ذات صلة بما أُعد لأجله الاختبار.

وبناءً على التعريفات السابقة للأداء التفاضلي وانطلاقاً من أهمية بناء فقرات للاختبارات والمقاييس بحيث تحقق خاصية تكافؤ القياس Measurement Invariance؛ أي يتمتع المقياس بخصائص قياس ثابتة عبر مجموعات الدراسة، ويعمل بالطريقة نفسها وبدون تحيز، فإن التعريف المعتمد للأداء التفاضلي للفقرة في هذه الدراسة يتلخص بأنه مؤشر إحصائي يُعبر عن اختلاف في الخصائص السيكومترية للفقرة عبر المجموعات المختلفة عند المستويات المتساوية في القدرة.

وهناك نوعان للأداء التفاضلي للفقرة: المنتظم (UNDIF)، وغير المنتظم (Non-Uniform DIF, NUNDIF) (Mellenbergh, ١٩٨٢). وتحديد ما إذا كان الأداء التفاضلي منتظماً أو غير منتظم يتوقف على طبيعة التفاعل بين مستوى القدرة وعضوية المجموعة فيهما (Ackerman, 1992; Mellenbergh, 1982; Millsap & Everson, 1993; Narayanan & Swaminathan, 1996) وعرّف أكرمان (Ackerman, 1992) الأداء التفاضلي المنتظم للفقرة

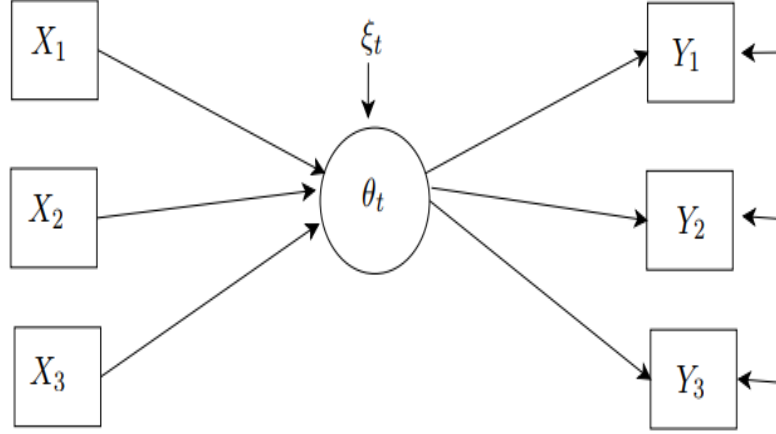
(UNDIF) على أنه اختلاف ثابت في منحنيات خصائص الفقرة (ICCs)، عبر المجموعات المختلفة عند المستويات المتساوية في القدرة؛ بمعنى عندما تكون منحنيات خصائص الفقرة (ICCs) متوازية، بينما في الأداء التفاضلي غير المنتظم (NUNDIF) تكون منحنيات خصائص الفقرة (ICCs) غير متوازية. وبين نارايانان وسواميناثان (Narayanan & Swaminathan, 1996) أن الأداء التفاضلي المنتظم (UNDIF) لا يوجد فيه تفاعل بين عضوية المجموعة ومستوى القدرة؛ أي يظهر أداء مجموعة أفضل أو أقل من مجموعة أخرى وثابت عند جميع مستويات القدرة، والاختلاف بين منحنيات الاستجابة للفقرة يكون في معلمة الصعوبة، بينما في الأداء التفاضلي غير المنتظم (NUNDIF) هناك تفاعل بين عضوية المجموعة ومستوى القدرة؛ بحيث يكون الفرق بين أداء المجموعتين غير ثابت عند جميع مستويات القدرة. والاختلاف بين منحنيات الاستجابة للفقرة يكون في معلمة التمييز. ويبين الشكل (١) أربعة أمثلة لأداء الفقرة التفاضلي. ففي اللوحة A تُظهر أداءً تفاضلياً منتظماً، بينما اللوحات B و C و D تُظهر أداءً تفاضلياً غير منتظم (Montoya & Jeon, 2019).



الشكل (١): منحنيات خصائص الفقرة التي تُظهر أداءً تفاضلياً منتظماً وغير منتظم

وتم تطوير العديد من الطرق للكشف عن DIF، ومن الأمثلة على هذه الطرق: نسبة الأرجحية-لما نزل (Mantel-Haenszel –Log Odds Ratio, MH-LOR)، وطرق المساحة، والانحدار اللوجستي، ونموذج راش، واختبارات تحيز الفقرة المتزامن (Simultaneous Item Bias, SIBTETS)، والمجموعات المتعددة (Multiple group IRT)، والنماذج متعددة المؤشرات متعددة الأسباب (Models, MIMIC) التي هي محور الاهتمام في الدراسة الحالية.

ونماذج MIMIC هي نماذج ذات معالم كتلك النماذج في نظرية الاستجابة للفقرة IRT، وتُعرف على أنها تحليل عاملي توكيدي ذو تباين مشترك (Confirmatory Factor analysis With Covariance) وتوصف نماذج MIMIC باعتبارها تطبيقات خاصة من نمذجة المعادلات البنائية (SEM) (Jöreskog Goldberger, 1975) ولهذه النماذج أهمية بالغة في التحقق من عدالة وصلاحيّة الاختبار لأنها تسمح بفحص الاختلافات في البنية الكامنة عبر مجموعات متعددة. كما يمكن استخدامها للكشف عن DIF المحتمل في المؤشرات الملاحظة (فقرات الاختبار) للمتغيرات الكامنة. ويعرض الشكل (٢) مثلاً لنموذج MIMIC يتضمن ثلاثة مؤشرات ملاحظة وثلاثة أسباب ملاحظة (Montoya & Jeon, 2019).



الشكل (٢): مثال لنموذج MIMIC

حيث تمثل (Y_1, Y_2, Y_3) ثلاثة مؤشرات ملاحظة، و (X_1, X_2, X_3) ثلاثة أسباب ملاحظة للمتغير الكامن θ_t ، كما يمثل ξ_t متغير البواقي. وتوفر نمذجة المعادلة البنائية (SEM) مع المتغيرات الكامنة طريقة مرنة لاختبار تكافؤ القياس (اللاتغير، Invariance)، لأنها تسمح باستخدام المتغيرات المشتركة المتصلة والمنفصلة، والمتغيرات الملاحظة والكامنة (Barendse et al., 2012).

وتطبق نماذج MIMIC لدراسة DIF باستخدام أساليب تقدير مناسبة للبيانات التصنيفية، ويوضح الشكل ٣ نموذج MIMIC البسيط وهو نموذج أحادي البعد ذو سمة واحدة مقاسة (θ) والمنحدرة على متغير المجموعة الملاحظة (Cause). ونستدل على وجود DIF من خلال ظهور فرق في الأداء للمتغير الكامن (θ) على متغير المجموعة، ولكن بعد التأكد من أن هذه الفروق في الأداء بين المجموعات لا تُعزى لاختلاف القدرة، فإذا ظهر اختلاف في صعوبة الفقرة بين المجموعات مع تساوي معاملات التمييز فهذا يشير إلى وجود أداء تفاضلي منتظم للفقرة (Wood, 2009).

وقد قام موثن وأسباروهوف (Muthén & Asparouhov, 2002) بتطوير معادلة لمتغير الاستجابة الكامنة ويمكن استخدامها لاختبار DIF المنتظم في الفقرات ثنائية التدرج على النحو:

$$y_i^* = \alpha_i \theta + \beta_i Z + \varepsilon_i \dots \dots (1)$$

حيث:

y_i^* : مؤشر ثنائي الاستجابة للمتغير الكامن.

α_i : معلمة التمييز.

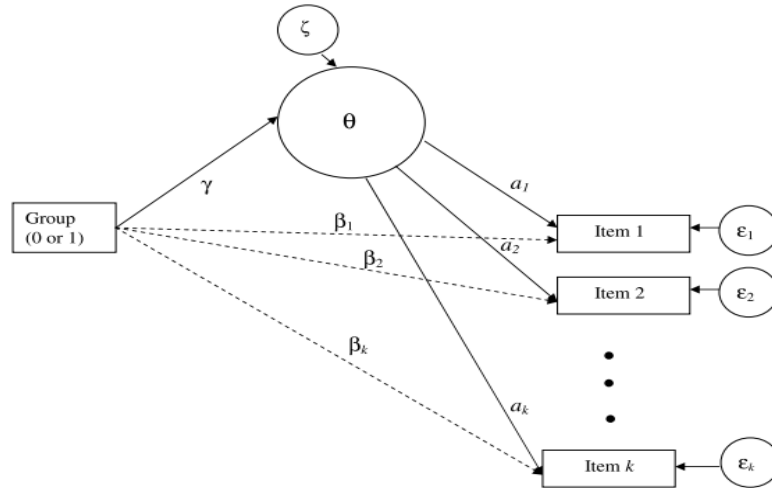
θ : قدرة المفحوص الكامنة.

β_i : معامل انحدار يشير إلى الفرق في صعوبة الفقرة i .

Z : المتغير الوهمي (Dummy Variable) الذي يحدد متغير المجموعة.

ε_i : خطأ القياس.

ويوضح الشكل رقم ٣ النموذج المستخدم للكشف عن الأداء التفاضلي المنتظم (Wood, 2009).



الشكل (٣): مثال لنموذج MIMIC للكشف عن الأداء التفاضلي المنتظم

حيث:

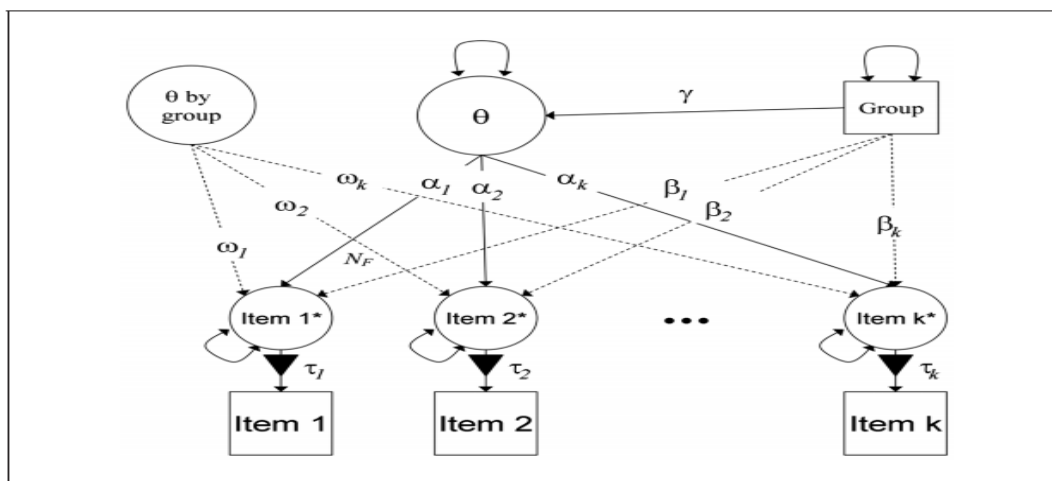
γ : معامل انحدار θ على متغير المجموعة وتشير إلى الفرق المتوسط بين متغير المجموعة على المتغير الكامن المقاس. a_i : معامل التمييز. ζ : متغير البواقي للمتغير الكامن θ

ولاختبار الأداء التفاضلي غير المنتظم NUNDIF؛ قام وودز وجريم (Woods & Grimm, 2011) ببناء نموذج بنيائي يحتوي على متغير جديد وهو متغير التفاعل؛ أي تفاعل المتغير الكامن θ مع متغير المجموعة G Causes، ونعبر عنه بالمتغير $(\theta \times G)$ ، ويتم تنفيذ ذلك باستخدام برمجية MPLUS. حيث تنحدر الفقرات على متغير التفاعل ونبحث في تأثيره على تلك الفقرات للكشف عن كلا النوعين من DIF (منتظم، وغير منتظم) في الفقرات ثنائية الاستجابة في وقت واحد، فإذا كانت النتائج تُظهر فرقاً في معاملات التمييز فهذا مؤشر على وجود الأداء التفاضلي غير المنتظم من خلال الصيغة التالية بعد إضافة متغير التفاعل:

$$y_i^* = a_i\theta + \beta_i Z_k + \omega_i \theta Z_k + \varepsilon_i \dots \dots (2)$$

حيث: ω_i : تأثير الأداء التفاضلي غير المنتظم للفقرة k .

ويوضح الشكل (٤) استخدام نموذج MIMIC للكشف عن الأداء التفاضلي المنتظم وغير المنتظم (Woods & Grimm, 2011).



الشكل (٤): نموذج MIMIC لاختبار الأداء التفاضلي غير المنتظم

حيث:

γ : علاقة تأثير تشير إلى الفروق في المتوسطات بين أداء المجموعات على المتغير الكامن θ
 α_k : تشير إلى معلمة التمييز للفقرة k . β_k : الفرق في صعوبة الفقرة k تبعاً لمتغير المجموعة.
 τ_k : العتبة الفارقة.

ويفيد أسلوب MIMIC في معالجة بعض جوانب القصور في بعض الطرق التقليدية والحديثة في الكشف عن الأداء التفاضلي للفقرة مثل: ضعف قدرتها على كشف الأداء التفاضلي غير المنتظم للفقرة، وحساسيتها للاختلافات في السمة الكامنة بين المجموعتين المستهدفة والمرجعية، وقصورها في التعامل مع بعض الحالات التي تنتهك افتراض التوزيع الطبيعي على الرغم من كونها إجراءات غير معلمية، وصغر حجم العينة الناتج عن تقسيم العينة إلى مجموعات فرعية متعددة. وهذا أدى إلى وجود قيود لاستخدام تلك الطرق (Wood & Grimm, 2011).

وقد تم تناول طرق الكشف عن الأداء التفاضلي للفقرات في الكثير من الدراسات. وبالرغم من كثرة تلك الدراسات فإن القليل منها تناول طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC خاصة على الصعيد العربي.

وبينت دراسة لتشنغ وآخرون (Cheng, et al., 2016) مرونة طريقة المؤشرات المتعددة والأسباب المتعددة (MIMIC) وفعاليتها في الكشف عن الأداء التفاضلي للفقرات (DIF) وقدرتها على تفسيره. وتوصلت دراسة لي وآخرون (Lee, Bulut, & Suh, 2016) إلى ازدياد فعالية طريق MIMIC بزيادة طول الاختبار وحجم العينة. وعند مقارنة طريقة MIMIC مع طرق SIBTEST، والانحدار اللوجستي LR، ومانتل هانزل MH في ضوء اختلاف حجم العينة، تبين أن الطرق تكون متسقة في نتائجها عندما يكون حجم العينة أكبر أو يساوي ٢٠٠٠ (Arikan, et al., 2016). وفحص جمالي وآخرون (Jamali et al., 2017) بالاعتماد على بيانات مولدة فاعلية نموذج الأسباب المتعددة والمؤشرات المتعددة MIMIC في فحص DIF تحت ظروف وشروط معينة مثل توزيع القدرة وحجم العينة وطول المقياس وفتات الاستجابة وحجم DIF. وأشارت النتائج إلى أن الالتواء المعتدل والعالي في توزيع السمات الكامنة أدى إلى انخفاض قوة نموذج MIMIC في كشف DIF المنتظم، في حين أدى زيادة طول المقياس وعدد فتات الاستجابة وحجم DIF إلى تحسين قوة النموذج MIMIC. وتوصل أوقورلو وعطار (Ugurlu & Atar, 2020) إلى أن زيادة حجم العينة وتقليل عدد الفقرات ذات الأداء التفاضلي يقلل من معدلات الخطأ من النوع الأول عند استخدام النماذج متعددة المؤشرات ومتعددة الأسباب (MIMIC) للكشف عن الأداء التفاضلي للفقرات.

وبتفحص الدراسات السابقة التي تناولت طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC في الكشف عن أداء الفقرة التفاضلي يتبين أن معظمها اعتمدت على بيانات مولدة وفق شروط محددة ومعروفة مسبقاً، لذلك جاءت هذه الدراسة للتحقق من النتائج السابقة من خلال استخدام بيانات حقيقية.

٢. مشكلة الدراسة وأسئلتها

يُعدّ الكشف عن الأداء التفاضلي للفقرات من القضايا المهمة للاختبارات، ونظراً لأنها قد تتأثر ببعض المتغيرات مثل حجم العينة، وطول الاختبار وخصائص الأفراد وصياغة الفقرات وغيرها، فقد أكد فينش (Finch, 2005) بالاعتماد على بيانات مولدة أن طريقة MIMIC فعالة في الكشف عن DIF عندما يكون عدد الفقرات ٥٠ فقرة وتتناقص الفعالية بنقصان عدد الفقرات إلى ما دون ٢٠ فقرة. وأشار وودز (Woods, 2009) إلى أن طريقة MIMIC أكثر فعالية من أساليب نظرية الاستجابة للفقرة في العينات صغيرة الحجم. وتوصل وودز وجريم (Woods & Grimm, 2011) إلى أن طريقة MIMIC أفضل من كل من طريقة مانتل هانزل وطريقة نسبة الأرجحية القصوى لنماذج استجابة الفقرة (LR) في الكشف عن الأداء التفاضلي DIF غير المنتظم.

وجاءت هذه الدراسة لدراسة فاعلية نموذج المؤشرات المتعددة والأسباب المتعددة MIMIC في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات لدى عينات من طلبة الصف العاشر الأساسي الذين تقدموا لاختبار بيزا الدولي عام ٢٠١٨ في ضوء متغيري حجم العينة وطول الاختبار بالاعتماد على بيانات حقيقية.

٢,١ أسئلة الدراسة

وبالتحديد فإن هذه الدراسة حاولت الإجابة عن السؤالين الآتيين:

١. هل تختلف فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC اختلافاً دالاً إحصائياً ($\alpha=0.05$) في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي باختلاف حجم العينة؟
٢. هل تختلف فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC اختلافاً دالاً إحصائياً ($\alpha=0.05$) في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي باختلاف طول الاختبار؟

٣. أهداف وأهمية الدراسة

٣,١ أهداف الدراسة

تهدف هذه الدراسة إلى ما يلي:

١. الكشف عن فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC في ضوء اختلاف حجم العينة.
٢. الكشف عن فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC في ضوء اختلاف طول الاختبار.

٣,٢ أهمية الدراسة

تأتي أهمية هذه الدراسة من عدة نواحي أهمها:

١. تقصي تأثير بعض المتغيرات مثل حجم العينة وطول الاختبار على فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي لعام ٢٠١٨ باختلاف حجم العينة وطول الاختبار؛ وهذا يفيد في تحديد الظروف المناسبة لاستخدام هذه الطريقة.
٢. الحاجة الماسة لدراسة DIF للاختبارات التي يبني على نتائجها قرارات مهمة تتعلق بالطالب أو جودة النظام التعليمي كما هو الحال في اختبار (PISA 2018) الدولي في الرياضيات.
٣. فتح المجال أمام المزيد من الاستخدام لطريقة المؤشرات المتعددة والأسباب المتعددة (MIMIC)، خاصة على الصعيد العربي، حيث إن نتائج هذه الدراسة ستثري المكتبة العربية بهذا النوع من البحوث نظراً لقلتها.

٤. التعريفات الإجرائية

الأداء التفاضلي للفقرة

مؤشر إحصائي يُعبر عن اختلاف في خصائص قياس فقرة عبر المجموعات المختلفة والمتساوية في القدرة.

الأداء التفاضلي تبعاً لمتغير الجنس

ويُقاس بدلالة الأداء التفاضلي على الفقرة لمفحوصين لهم القدرة نفسها، لكنهم يختلفون تبعاً لنوع الجنس.

فاعلية النموذج Model Effectiveness

درجة التوافق في نسبة الفقرات التي تبدي أداء تفضلياً باستخدام نموذج MIMIC، في ضوء اختلاف حجم العينة وطول الاختبار.

٥. محددات الدراسة

- اقتصرَت هذه الدراسة على الأداء التفاضلي لمتغير الجنس.
- استخدام اختبار الدولي PISA 2018 في الرياضيات للصف العاشر للطلبة الأردنيين.
- حددت الدراسة ثلاثة مستويات لحجم العينة ٤٠٠٠، ٣٠٠٠، ٢٠٠٠ مفحوص، وثلاثة مستويات لطول الاختبار ٣٠ فقرة، ٢٠ فقرة، ١٠ فقرات.

٦. الطريقة والإجراءات

٦,١ مجتمع الدراسة وعينتها

تألف مجتمع الدراسة من جميع طلبة الصف العاشر الأساسي من جميع مدارس الأردن التي شاركت في الاختبار الدولي بيزا للرياضيات الذي أجري عام ٢٠١٨، حيث بلغ عدد المدارس المشاركة (٣١٣) مدرسة، وبلغ عدد الطلبة الذين شاركوا في الاختبار ٨٩٦٣ طالبًا وطالبة، وتكون المجتمع بصورته النهائية من استجابات ٤٩٩٣ طالبًا وطالبة بعد استبعاد الاستجابات المفقودة وغير المكتملة على جميع فقرات الاختبار، ولتحقيق غرض الدراسة جرى اختيار ثلاث عينات تختلف في حجم العينة بطريقة عشوائية وهي: ٤٠٠٠، ٣٠٠٠، ٢٠٠٠ طالب وطالبة، وتم استبعاد الحجوم التي تقل عن ذلك؛ حيث كانت طريقة MIMIC قليلة الفاعلية في الكشف عن الأداء التفاضلي في تلك الحجوم. وقد يؤدي استخدام الحجوم الصغيرة الى نتائج مظللة أو متباينة بين طرق كشف الأداء التفاضلي (Arikan, et al., 2016). كما تم اعتماد ثلاثة مستويات لطول الاختبار: المستوى الأول يتكون من ٣٠ فقرة مختارة عشوائيًا لتمثل المجالات المقاسة في الرياضيات، والمستوى الثاني يتكون من ٢٠ فقرة مختارة عشوائيًا من فقرات المستوى الأول، والمستوى الثالث يتكون من ١٠ فقرات مختارة عشوائيًا من فقرات المستوى الثاني. وكانت الاستجابات لهذه المجموعات من الأفراد على هذه الفقرات هي بمثابة عينة الدراسة.

٦,٢ إجراءات الدراسة

تم الحصول على البيانات اللازمة لإجابة أسئلة الدراسة والتي تمثل استجابات الطلبة عن فقرات اختبار الرياضيات في الاختبار الدولي بيزا ٢٠١٨ والذي تكون من ٨٣ فقرة بصورته الأصلية، تم الحصول عليها من مركز تنمية الموارد البشرية المسؤول عن تنفيذ الاختبار في الأردن. وتم استبعاد ٨ فقرات من نوع الاستجابات الجزئية (partial credit) أي غير ثنائية التصحيح، وكذلك تم استبعاد ١١ فقرة موضوعية؛ وذلك لعدم وجود أي استجابات عليها من جميع الطلبة، ليصبح الاختبار الذي تم دراسته بواقع ٦٤ فقرة.

وتم الكشف عن الأداء التفاضلي للفقرات باستخدام طريقة MIMIC في ضوء اختلاف حجم العينة وطول الاختبار من خلال برمجة MPLUS، على النحو الآتي:

تم التحقق من الأداء التفاضلي المنتظم للفقرة وذلك باختبار أثر المجموعة على الفقرة حسب الشكل رقم ٣، وأما الأداء التفاضلي غير المنتظم فتم انشاء متغير جديد وهو متغير التفاعل بين المجموعة والسمة الكامنة المقاسة ($\theta \times G$) حسب الشكل رقم ٤، وتم دراسة أثر كل من المجموعة Causes والتفاعل والمتغير الكامن كمتغيرات مستقلة على المؤشرات Indicators. فإذا وجد تأثير للتفاعل على المؤشرات (الفقرات) فهذا مؤشر على وجود الأداء التفاضلي غير المنتظم. وتم إجراء التحليلات عند حجوم عينات (٤٠٠٠، ٣٠٠٠، ٢٠٠٠)، وأطوال اختبار (٣٠، ٢٠، ١٠) فقرات. وتم استبعاد العينات ذات الحجم ١٠٠٠ فما دون؛ لأنه تبين من خلال تجربتها في التحليل عدم قدرة طريقة MIMIC على كشف الفقرات ذات الأداء التفاضلي.

٧. نتائج الدراسة ومناقشتها

للإجابة عن تساؤلات الدراسة تم تطبيق برنامج MPLUS على استجابات الأفراد في مستويات حجوم العينة وأطوال الاختبار، حيث تم إجراء تحليل للبيانات في تسع حالات مختلفة باستخدام نموذج MIMIC، وكانت نتائج الكشف عن الفقرات التي أظهرت أداءً تفاضلياً على الشكل الآتي:

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٣٠ فقرة.

يبين الجدول (١) أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٣٠ فقرة.

جدول (١): أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٣٠ فقرة.

رقم الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
	الوزن المعياري β	SE	قيمة t	p-value	الوزن المعياري ω	SE	قيمة t	p-value	
٢	٠,٠٣٨	٠,٠١١	٣,٤٤٨	٠,٠٠١	٠,٠١٤	٠,٠١٢	١,٢١١	٠,٢٢٦	منتظم
٥	٠,٠٢٣	٠,٠٠٩	٢,٥٢٩	٠,٠١١	-٠,٠٢٢	٠,٠٠١	-٢,٢٥٦	٠,٠٢٤	غير منتظم
٦	٠,٠٤٤	٠,٠٠١	٤,٣٨٢	٠,٠٠٠	-٠,٠٣٨	٠,٠١١	-٣,٥٣٩	٠,٠٠٠	غير منتظم
١٤	-٠,٠٣١	٠,٠١٤	-٢,٢٣١	٠,٠٢٦	٠,٠٢٤	٠,٠١٥	١,٥٤٩	٠,١٢١	منتظم
١٧	-٠,٠٣٤	٠,٠١٢	-٢,٧٢١	٠,٠٠٧	٠,٠٠٩	٠,٠١٤	٠,٦٥١	٠,٥١٥	منتظم
٢٠	٠,٠٣٢	٠,٠١٤	٢,٣٢٣	٠,٠٢٠	٠,٠٠٨	٠,٠١٠	٠,٧٥٥	٠,٤٥٠	منتظم
٢٢	٠,٠٦٦	٠,٠١٧	٣,٩١١	٠,٠٠٠	٠,٠٣٨	٠,٠١٠	٣,٧٢٣	٠,٠٠٠	غير منتظم
٢٣	٠,٠٣١	٠,٠١٤	٢,٣٠٣	٠,٠٢١	٠,٠٠٥	٠,٠١٠	٠,٤٨٠	٠,٦٣١	منتظم
٢٤	٠,٠٢٨	٠,٠١١	٢,٥٧٩	٠,٠١٠	-٠,٠٠١	٠,٠١٠	-٠,٠٦٩	٠,٩٤٥	منتظم
٢٧	-٠,٠٥٧	٠,٠١٣	-٤,٣٦٤	٠,٠٠٠	٠,٠٠٢	٠,٠١٤	٠,١٦٩	٠,٨٦٦	منتظم

تشير النتائج في الجدول ١ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ١٠ فقرات من أصل ٣٠ فقرة، أي ما نسبته ٣٣,٣٪ من فقرات الاختبار عندما يكون حجم العينة ٤٠٠٠، وكان الأداء التفاضلي منتظماً في ٧ فقرات وغير منتظم في ٣ فقرات، ولصالح الذكور في الفقرات (٢، ٢٠، ٢٣، ٢٤)، ولصالح الإناث في الفقرات (١٤، ١٧، ٢٧).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٢٠ فقرة.

يبين الجدول (٢) أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٢٠ فقرة.

جدول (٢): أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ٢٠ فقرة.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	p-value	الوزن المعياري ω	SE	قيمة t	p-value	
٢	١٢	٠,٠٣٨	٠,٠١١	٣,٤٧٩	٠,٠٠١	٠,٠١٠	٠,٠١٢	٠,٧٨٢	٠,٤٣٤	منتظم
٩	١١٤	-٠,٠٣٠	٠,٠١٤	-٢,٢٠٦	٠,٠٢٧	٠,٠٢٢	٠,٠١٥	١,٤٥٨	٠,١٤٥	منتظم
١٢	١١٧	-٠,٠٠٣	٠,٠١٣	-٢,٢٩٩	٠,٠٢٢	٠,٠٠٩	٠,٠١٤	٠,٦٢١	٠,٥٣٤	منتظم
١٥	١٢٢	-٠,٠٠٧	٠,٠٢٥	-٠,٢٧٧	٠,٧٨٢	٠,٠٤٣	٠,٠١٠	٤,٢١٢	٠,٠٠٠	غير منتظم
١٩	١٢٧	-٠,٠٤٨	٠,٠١٢	-٣,٨٣٧	٠,٠٠٠	٠,٠٠٣	٠,٠١٤	٠,١٨٢	٠,٨٥٦	منتظم

تشير النتائج في الجدول ٢ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٥ فقرات من أصل ٢٠ فقرة، أي ما نسبته ٢٥٪ من فقرات الاختبار عندما يكون حجم العينة ٤٠٠٠، وكان الأداء التفاضلي منتظماً في ٤ فقرات وغير منتظم في فقرة واحدة، ولصالح الذكور في الفقرة ٢، ولصالح الإناث في الفقرات (١٤، ١٧، ٢٧).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ١٠ فقرات.

يبين الجدول (٣) أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ١٠ فقرات.

جدول (٣): أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٤٠٠٠ وطول الاختبار ١٠ فقرات.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	١٢	٠,٠٤١	٠,٠١١	٣,٦٤٧	٠,٠٠٠	-٠,٠٠٣	٠,٠١٣	-٠,٢٠٧	٠,٨٣٦	منتظم
٧	١٢٢	٠,٠٣٨	٠,٠١١	٣,٣١٣	٠,٠٠١	-٠,٠١٢	٠,٠١٣	-٠,٠٩٢	٠,٣٥٨	منتظم
١٠	١٣٠	٠,٠١١	٠,٠١١	٠,٩٧٩	٠,٣٢٨	٠,٠٢٤	٠,٠١١	٢,١٥٧	٠,٠٣١	غير منتظم

تشير النتائج في الجدول ٣ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٣ فقرات من أصل ١٠ فقرات، أي ما نسبته ٣٠٪ من فقرات الاختبار عندما يكون حجم العينة ٤٠٠٠، وكان الأداء التفاضلي منتظماً في فقرتين وغير منتظم في فقرة واحدة، ولصالح الذكور في الفقرتين (٢، ٢٢).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٣٠ فقرة.

يبين الجدول (٤) أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٣٠ فقرة.

جدول (٤): أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٣٠ فقرة.

رقم الفقرة	الوزن المعياري β	SE	قيمة t	P-value	مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
					الوزن المعياري ω	SE	قيمة t	P-value	
٢	٠,٠٤٥	٠,٠١٣	٣,٤٨٦	٠,٠٠٠	٠,٠١٥	٠,٠١٤	١,٠٤٣	٠,٢٩٧	منتظم
٥	٠,٠٣٦	٠,٠١١	٣,٢٧٥	٠,٠٠١	-٠,٠٠٣	٠,٠١١	-٢,٦٢٥	٠,٠٠٩	غير منتظم
٦	٠,٠٦٠	٠,٠١٣	٤,٧٢٩	٠,٠٠٠	-٠,٠٤٨	٠,٠١٣	-٣,٦٢٦	٠,٠٠٠	غير منتظم
٧	٠,٠٣١	٠,٠١٥	٢,١١٨	٠,٠٣٤	-٠,٠٢٦	٠,٠١٥	-١,٧١١	٠,٠٨٧	منتظم
١٧	-٠,٠٣٥	٠,٠١٥	-٢,٢٨٩	٠,٠٢٢	٠,٠٠٩	٠,٠١٧	٠,٥٥٠	٠,٥٨٢	منتظم
١٩	٠,٠٥١	٠,٠٢١	٢,٤٦٩	٠,٠١٤	-٠,٠١٥	٠,٠١٣	-١,١٨٣	٠,٢٣٧	منتظم
٢٠	٠,٠٥٠	٠,٠١٦	٣,٠٥٣	٠,٠٠٢	٠,٠٠٨	٠,٠١٢	٠,٦٥٩	٠,٥١٠	منتظم
٢٢	٠,٠٩٠	٠,٠٢٠	٤,٥٣٢	٠,٠٠٠	٠,٠٣٨	٠,٠١٢	٣,١٣٣	٠,٠٠٢	غير منتظم
٢٣	٠,٠٥٥	٠,٠١٦	٣,٣٢٦	٠,٠٠١	٠,٠١٢	٠,٠١٢	١,٠٣١	٠,٣٠٣	منتظم
٢٧	-٠,٠٦٧	٠,٠١٦	-٤,٠٨٥	٠,٠٠٠	٠,٠٠٣	٠,٠١٦	٠,١٧٨	٠,٨٥٩	منتظم

تشير النتائج في الجدول ٤ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ١٠ فقرات من أصل ٣٠ فقرة، أي ما نسبته ٣٣,٣٪ من فقرات الاختبار عندما يكون حجم العينة ٣٠٠٠، وكان الأداء التفاضلي منتظماً في ٧ فقرات وغير منتظم في ٣ فقرات، ولصالح الذكور في الفقرات (٢، ٧، ١٩، ٢٠، ٢٣)، ولصالح الإناث في الفقرتين (١٧، ٢٧).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٢٠ فقرة.

يبين الجدول (٥) أرقام الفقرات التي أظهرت أداءً تفضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٢٠ فقرة.

جدول (٥): أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ٢٠ فقرة.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	١٢	٠,٠٤٥	٠,٠١٣	٣,٥٠٩	٠,٠٠٠	٠,٠١٠	٠,٠١٤	٠,٧٢٣	٠,٤٦٩	منتظم
٤	١٧	٠,٠٢٩	٠,٠١٥	١,٩٨	٠,٠٤٨	-٠,٠١٦	٠,٠١٦	-١,٠٢٣	٠,٣٠٦	منتظم
١٢	١١٧	-٠,٠٣٢	٠,٠١٦	-٢,٠٢٩	٠,٠٤٢	٠,٠٠٧	٠,٠١٧	٠,٣٩٧	٠,٦٩١	منتظم
١٥	١٢٢	٠,٠١٥	٠,٠٢٨	٠,٥٣٧	٠,٥٩١	٠,٠٤٥	٠,٠١٢	٣,٦٢٤	٠,٠٠٠	غير منتظم
١٩	١٢٧	-٠,٠٥١	٠,٠١٥	-٣,٣٤٤	٠,٠٠١	٠,٠٠١	٠,٠١٧	٠,٠٦٦	٠,٩٤٨	منتظم

تشير النتائج في الجدول ٥ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٥ فقرات من أصل ٢٠ فقرة، أي ما نسبته ٢٥٪ من فقرات الاختبار عندما يكون حجم العينة ٣٠٠٠، وكان الأداء التفاضلي منتظماً في ٤ فقرات وغير منتظم في فقرة واحدة، وكانت لصالح الذكور في الفقرتين (٢، ٧)، ولصالح الإناث في الفقرتين (١٧، ٢٧).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ١٠ فقرات.

يبين الجدول (٦) أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ١٠ فقرات.

جدول (٦): أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٣٠٠٠ وطول الاختبار ١٠ فقرات.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	١٢	٠,٠٤٩	٠,٠١٣	٣,٦٩٦	٠,٠٠٠	-٠,٠٠٧	٠,٠١٥	-٠,٤٤٥	٠,٦٥٦	منتظم
٧	١٢٢	٠,٠٤٥	٠,٠١٤	٣,١٨	٠,٠٠١	-٠,٠١٤	٠,٠١٦	-٠,٨٩٥	٠,٣٧١	منتظم
١٠	١٣٠	٠,٠١١	٠,٠١٣	٠,٨٦١	٠,٣٨٩	٠,٠٣٢	٠,٠١٣	٢,٤٤٦	٠,٠١٤	غير منتظم

تشير النتائج في الجدول ٦ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٣ فقرات من أصل ١٠ فقرات، أي ما نسبته ٣٠٪ من فقرات الاختبار عندما يكون حجم العينة ٣٠٠٠، وكان الأداء التفاضلي منتظماً في فقرتين وغير منتظم في فقرة واحدة، ولصالح الذكور في الفقرتين (٢، ٢٢).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٣٠ فقرة.

يبين الجدول (٧) أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٣٠ فقرة.

جدول (٧): أرقام الفقرات التي أظهرت أداءً تفاضلياً للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٣٠ فقرة.

رقم الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
	الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	٠,٠٤١	٠,٠١٧	٢,٤٥٤	٠,٠١٤	٠,٠٢١	٠,٠١٧	١,٢١٨	٠,٢٢٣	منتظم
٥	٠,٠٣٩	٠,٠١٤	٢,٧٩١	٠,٠٠٥	-٠,٠٣١	٠,٠١٤	-٢,١٨٢	٠,٠٢٩	غير منتظم
٦	٠,٠٦٣	٠,٠١٧	٣,٨٣٦	٠,٠٠٠	-٠,٠٤٨	٠,٠١٧	-٢,٨٣٤	٠,٠٠٥	غير منتظم

منتظم	٠,١٣٩	-١,٤٧٩	٠,٠١٩	-٠,٠٢٨	٠,٠٥٠	١,٩٦٤	٠,٠١٨	٠,٠٣٦	٧
منتظم	٠,٩٣٨	٠,٠٧٧	٠,٠١٨	٠,٠٠١	٠,٠٤٢	٢,٠٣٥	٠,٠١٧	٠,٠٣٥	١٨
منتظم	٠,٨١٢	٠,٢٣٨	٠,٠١٥	٠,٠٠٤	٠,٠٢٢	٢,٢٩٧	٠,٠٢٠	٠,٠٤٦	٢٠
غير منتظم	٠,٠٠٤	٢,٨٦٣	٠,٠١٥	٠,٠٤٣	٠,٠٠٠	٣,٥٢٦	٠,٠٢٤	٠,٠٨٥	٢٢
منتظم	٠,٩٨٥	٠,٠١٩	٠,٠١٥	٠,٠٠٠	٠,٠٣٢	٢,١٤٩	٠,٠٢٠	٠,٠٤٣	٢٣
منتظم	٠,٧٩٩	٠,٢٥٥	٠,٠٢١	٠,٠٠٥	٠,٠٠٠	-٣,٥٤	٠,٠٢١	-٠,٠٧٣	٢٧
منتظم	٠,٧٨٨	٠,٢٦٩	٠,٠١٨	٠,٠٠٥	٠,٠١٢	-٢,٥٠٥	٠,٠١٧	-٠,٠٤٣	٢٩

تشير النتائج في الجدول ٧ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ١٠ فقرات من أصل ٣٠ فقرة، أي ما نسبته ٣٣,٣٪ من فقرات الاختبار عندما يكون حجم العينة ٢٠٠٠، وكان الأداء التفاضلي منتظمًا في ٧ فقرات وغير منتظم في ٣ فقرات، وكانت لصالح الذكور في الفقرات (٢، ٧، ١٨، ٢٠، ٢٣)، ولصالح الإناث في الفقرتين (٢٧، ٢٩).

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٢٠ فقرة.

يبين الجدول (٨) أرقام الفقرات التي أظهرت أداءً تفاضليًا للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٢٠ فقرة.

جدول (٨): أرقام الفقرات التي أظهرت أداءً تفاضليًا للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ٢٠ فقرة.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	١٢	٠,٠٤١	٠,٠١٧	٢,٤٧١	٠,٠١٣	٠,٠١٦	٠,٠١٨	٠,٨٨٢	٠,٣٧٨	منتظم
١٥	١٢٢	٠,٠٢٥	٠,٠٣٤	٠,٧٢٢	٠,٤٧٠	٠,٠٤٩	٠,٠١٥	٣,٢٥٩	٠,٠٠١	غير منتظم
١٩	١٢٧	-٠,٠٥٩	٠,٠٢٠	-٢,٩٨٢	٠,٠٠٣	٠,٠٠٤	٠,٠٢١	٠,١٨٣	٠,٨٥٥	منتظم

تشير النتائج في الجدول ٨ إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٣ فقرات من أصل ٢٠ فقرة، أي ما نسبته ١٥٪ من فقرات الاختبار عندما يكون حجم العينة ٢٠٠٠، وكان الأداء التفاضلي منتظمًا في فقرتين وغير منتظم في فقرة واحدة، ولصالح الذكور في الفقرة ٢، ولصالح الإناث في الفقرة ٢٧.

نتائج فحص الأداء التفاضلي عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ١٠ فقرات.

يبين الجدول (٩) أرقام الفقرات التي أظهرت أداءً تفاضليًا للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ١٠ فقرات.

جدول (٩): أرقام الفقرات التي أظهرت أداءً تفاضليًا للجنس وفق طريقة MIMIC عندما يكون حجم العينة ٢٠٠٠ وطول الاختبار ١٠ فقرات.

رقم الفقرة	رمز الفقرة	مؤشرات الفقرات على متغير الجنس				مؤشرات الفقرات على التفاعل (الجنس × θ)				نوع DIF
		الوزن المعياري β	SE	قيمة t	P-value	الوزن المعياري ω	SE	قيمة t	P-value	
٢	١٢	٠,٠٤٥	٠,٠١٧	٢,٦٩١	٠,٠٠٧	-٠,٠١٠	٠,٠١٩	-٠,٥٢٣	٠,٦٠١	منتظم
٧	١٢٢	٠,٠٥١	٠,٠١٨	٢,٧٧٧	٠,٠٠٥	-٠,٠١٧	٠,٠٢٠	-٠,٨٢٨	٠,٤٠٨	منتظم
٨	١٢٥	-٠,٠٤٠	٠,٠١٩	-٢,٠٤٩	٠,٠٤٠	-٠,٠٠٩	٠,٠١٦	-٠,٥٨١	٠,٥٦١	منتظم
١٠	١٣٠	٠,٠٣١	٠,٠١٦	١,٩٤٦	٠,٠٥٢	٠,٠٣٢	٠,٠١٦	١,٩٧٩	٠,٠٤٨	غير منتظم

تشير النتائج في الجدول (٩) إلى أن طريقة MIMIC كشفت عن الأداء التفاضلي في ٤ فقرات من أصل ١٠ فقرات، أي ما نسبته ٤٠٪ من فقرات الاختبار عندما يكون حجم العينة ٢٠٠٠، وكان الأداء التفاضلي منتظمًا في ٣ فقرات وغير منتظم في فقرة واحدة، ولصالح الذكور في الفقرتين (٢، ٢٢)، وفي الفقرة ٢٥ كانت لصالح الإناث.

النتائج المتعلقة بالإجابة عن السؤال الأول من أسئلة الدراسة

للإجابة عن السؤال الأول من أسئلة الدراسة "هل تختلف فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC اختلافًا دالًا احصائيًا ($\alpha=0.05$) في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي باختلاف حجم العينة؟

تم تلخيص نتائج التحليل في الجدول (١٠) الذي يبين عدد ونسب الأداء التفاضلي للفقرات وفق نموذج MIMIC في ضوء اختلاف حجم العينة في الأطوال المختلفة للاختبار.

جدول (١٠): عدد ونسبة الفقرات التي تبدي أداءً تفاضليًا لاختبار الرياضيات الدولي PISA للصف العاشر وفق نموذج MIMIC، ودلالة اختبار مربع كاي للاستقلال χ^2 في ضوء اختلاف حجم العينة في الأطوال المختلفة للاختبار.

طول الاختبار	حجم العينة	العدد	النسبة	χ^2	Sig
٣٠ فقرة	٤٠٠٠	١٠	٣٣,٣%	٠	١
	٣٠٠٠	١٠	٣٣,٣%		
	٢٠٠٠	١٠	٣٣,٣%		
	المجموع	٣٠	٣٣,٣%		
٢٠ فقرة	٤٠٠٠	٥	٢٥%	٠,٧٨٦	٠,٦٧٥
	٣٠٠٠	٥	٢٥%		
	٢٠٠٠	٣	١٥%		
	المجموع	١٣	٢١,٧%		
١٠ فقرات	٤٠٠٠	٣	٣٠%	٠,٣	٠,٨٦١
	٣٠٠٠	٣	٣٠%		
	٢٠٠٠	٤	٤٠%		
	المجموع	١٠	٣٣,٣%		

يُلاحظ في الجدول ١٠ أن جميع قيم (sig) أكبر من ($\alpha = 0.05$)، أي أن قيم مربع كاي للاستقلال غير دالة إحصائيًا، ما يدل على قبول الفرضية الصفرية التي تنص على عدم وجود اختلاف في نسبة الفقرات التي تبدي أداءً تفاضليًا للجنس باختلاف حجم العينة بغض النظر عن طول الاختبار. وتعني هذه النتيجة أن فاعلية نموذج المؤشرات المتعددة والأسباب المتعددة في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي لا تختلف باختلاف حجم العينة.

النتائج المتعلقة بالإجابة عن السؤال الثاني من أسئلة الدراسة

للإجابة السؤال الرئيسي الثاني من هذه الدراسة "هل تختلف فاعلية طريقة المؤشرات المتعددة والأسباب المتعددة MIMIC اختلافًا دالًا احصائيًا ($\alpha=0.05$) في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار بيزا الدولي باختلاف طول الاختبار؟" تم تلخيص ما توصلت إليه الدراسة من نتائج في الجدول (١١) على الشكل الآتي.

جدول (١١): عدد ونسبة الفقرات التي تبدي أداءً تفاضليًا لاختبار الرياضيات الدولي PISA للصف العاشر وفق نموذج MIMIC، ودلالة اختبار مربع كاي للاستقلال χ^2 في ضوء اختلاف طول الاختبار في حجومات العينات المختلفة.

حجم العينة	طول الاختبار	العدد	النسبة	χ^2	Sig
٤٠٠٠	٣٠ فقرة	١٠	٣٣,٣%	٠,٣٩٧	٠,٨٢٠
	٢٠ فقرة	٥	٢٥%		
	١٠ فقرات	٣	٣٠%		

		المجموع	١٨	٣٠.%	
٣٠٠٠	٠,٣٩٧	٣٠ فقرة	١٠	٣٣,٣%	٠,٨٢٠
		٢٠ فقرة	٥	٢٥%	
		١٠ فقرات	٣	٣٠.%	
		المجموع	١٨	٣٠.%	
٢٠٠٠	٢,٧٩١	٣٠ فقرة	١٠	٣٣,٣%	٠,٢٤٨
		٢٠ فقرة	٣	١٥%	
		١٠ فقرات	٤	٤٠.%	
		المجموع	١٧	٢٨,٣%	

يُلاحظ في الجدول (١١) أن جميع قيم (*sig*) أكبر من ($\alpha=0,05$) أي أن قيم مربع كاي للاستقلال غير دالة إحصائياً، ما يدل على قبول الفرضية الصفرية التي تنص على عدم وجود اختلاف في نسبة الفقرات التي تبدي أداءً تفاضلياً للجنس باختلاف طول الاختبار بغض النظر عن حجم العينة. وتعني هذه النتيجة أن فاعلية نموذج المؤشرات المتعددة والأسباب المتعددة في الكشف عن الأداء التفاضلي للجنس في فقرات الرياضيات في اختبار ييزا الدولي لا تختلف باختلاف طول الاختبار.

٧,١ مناقشة النتائج

إذا أمعنا النظر في النتائج المعروضة في الجداول المرقمة من ١ - ٩ والمتعلقة بالكشف عن الأداء التفاضلي للفقرات لمتغير الجنس تبعاً لاختلاف حجم العينة، وطول الاختبار، وإذا قمنا بتثبيت طول الاختبار عند طول معين والتغيير في حجم العينة وملاحظة أعداد الفقرات ذات الأداء التفاضلي التي تكتشفها الطريقة في كل مرة تغير حجم العينة، يُلاحظ مثلاً عند طول اختبار ٣٠ فقرة بغض النظر عن حجم العينة كشفت طريقة MIMIC عن ١٠ فقرات، نصف تلك الفقرات تقريباً كانت لصالح الذكور، وفقرتين لصالح الإناث، و٣ فقرات أبدت DIF غير منتظم، وعند طول اختبار ٢٠ كشفت طريقة MIMIC عن ٥ فقرات في أغلب مرات تغيير حجم العينة، فقرة لصالح الذكور، و٣ لصالح الإناث، وفقرة أبدت DIF غير منتظم، وعند طول اختبار ١٠ كشفت طريقة MIMIC عن ٣ فقرات في أغلب مرات تغيير حجم العينة، وكانت لصالح الذكور في فقرتين، وقد يُعزى الأداء التفاضلي للفقرات بين الذكور والإناث إلى تأثيره بمحتوى الفقرة، فقد تبين أن الفقرات التي أبدت أداءً تفاضلياً لصالح الذكور تناسب اهتماماتهم؛ كالرياضة وتأجير الدراجات والسرعة وغيرها، أما الفقرات التي أبدت أداءً تفاضلياً لصالح الإناث فقد كانت تناسب اهتمامتهن؛ كترتيب المقاعد، والبيانات الوظيفية، وغيرها.

كما نلاحظ بالرجوع إلى نسبة الفقرات ذات الأداء التفاضلي، فإننا نجد عدم اختلافها بشكل جوهري، وهذا ما أكدته نتيجة اختبار مربع كاي للاستقلال χ^2 المبينة في الجدول رقم ١٠، والتي تشير إلى عدم وجود دلالة إحصائية عند مستوى الدلالة الإحصائية ($\alpha=0,05$) وبهذه النتيجة نتوصل إلى عدم اختلاف فعالية الطريقة في قدرتها على الكشف على الأداء التفاضلي للفقرات في العينات التي تزيد عن ٢٠٠٠ بمعنى أن أصغر حجم عينة استخدمته هذه الدراسة كان كافياً لإظهار فاعلية الطريقة وقدرتها على الكشف عن الأداء التفاضلي للفقرات بشكل دقيق، وزيادة حجم العينة لأكثر من ذلك لن يزيد من فعاليتها بشكل جوهري، فهو بحد ذاته حجم مرتفع، وقد يكون سبباً وراء ظهور ذلك، إلا أن تقليص حجم العينة وخصوصاً لأقل من ١٠٠٠ قد لا يظهر فعاليتها بشكل جيد، وقد يقود إلى نتائج مظلمة أحياناً، وهذا ما تم ملاحظته عند إجراء تحليل لعينات أقل من ١٠٠٠ في هذه الدراسة، حيث تم استبعادها لعدم قدرة الطريقة على كشف الفقرات ذات الأداء التفاضلي.

من هنا يمكن القول إن فاعلية الطريقة في الكشف عن الأداء التفاضلي للفقرات تظهر عند استخدام حجوم عينات ٢٠٠٠ أو أكثر وتكون النتائج أكثر ثباتاً واستقراراً. وتتفق هذه النتيجة مع نتائج العديد من الدراسات التي أفادت أنه بزيادة حجم العينة تكون النتائج أكثر دقة وأكثر استقراراً (Jamali, Ayatollahi, & Jafari, 2017; Arian & Ugural & Atar, 2016; Lee, Bulut, & Suh, 2016; Ugurlu & Atar, 2020).

وبشكل عام عند تثبيت حجم العينة وتغيير طول الاختبار لوحظ أن عدد الفقرات ذات الأداء التفاضلي والذي تكتشفه الطريقة يقل عند تقليص طول الاختبار، فمثلاً عند حجم عينة ٤٠٠٠ كشفت الطريقة عن ١٠، ٥، ٣ فقرات عند أطوال الاختبارات ٣٠، ٢٠، ١٠ على الترتيب، وعند حجم عينة ٣٠٠٠ كشفت عن ١٠، ٥، ٣ فقرات عند أطوال الاختبار ٣٠، ٢٠، ١٠ على الترتيب، وعند مستوى عينة ٢٠٠٠ كشفت عن ١٠، ٣، ٤ فقرات، وإذا تم النظر إلى مجموع الفقرات ذات الأداء التفاضلي والمكتشفة عبر مستويات حجم العينة عند كل طول اختبار نجد أيضاً أن مجموع ما تكتشفه كل طريقة يقل تبعاً لتقليص طول الاختبار المستخدم، وإذا تم النظر إلى نسبة الفقرات المكتشفة وتحديداً عند المجموع نجد أنها في الغالب لا

تختلف جوهرياً؛ وهذا يعني عدم اختلاف فاعلية الطريقة باختلاف طول الاختبار، وهذا ما تأكدته القيم غير الدالة إحصائياً لمربع كاي χ^2 والمعروضة في الجدول رقم ١١، إلا أن انخفاض عدد الفقرات عند تقليص طول الاختبار من ٣٠ إلى ٢٠ ومن ٢٠ إلى ١٠؛ فهذه نتيجة منطقية وخصوصاً أن تقليص طول الاختبار من ٣٠ فقرة إلى ٢٠ فقرة، كان من خلال اختبار ٢٠ فقرة من نفس طول الاختبار ٣٠، وتقليص طول الاختبار من ٢٠ إلى ١٠ فقرات كان من نفس طول الاختبار ٢٠. ولا يختلف الحال في نتائج الدراسات التي استخدمت متغير طول المقياس أو الاختبار في تأثيره على اكتشاف الأداء التفاضلي للفقرات (Lee et al., 2016; Jamali et al., 2017; Finch, 2005).

وإذا أخذنا بعين الاعتبار نوع الأداء التفاضلي الذي تكتشفه الطريقة نلاحظ أن عدد ونسبة الأداء التفاضلي غير المنتظم قليلة مقارنة بالمنتظم، وعلى الرغم من هذا تبقى هذه ميزة لطريقة MIMIC في قدرتها على كشف ذلك مقارنة بالطرق الأخرى، حيث تعاني معظم الطرق من صعوبة ومشاكل في الكشف عن الأداء التفاضلي غير المنتظم، وتبقى فاعليتها أعلى في الكشف عن الأداء التفاضلي المنتظم (Lee & Bulut & Suh, 2016; Walker, 2001). إلا أن ذلك يتأثر بتقلبات طول الاختبار، بحيث يقل عدد الفقرات التي تبدي أداءً تفضلياً غير منتظم والتي تكتشفها الطريقة في أطوال الاختبارات الصغيرة منها، وهذه النتيجة تتفق مع دراسة سواميناثان وروجرز (Swaminathan & Rogers, 1990).

وكما هو الحال في حجوم العينات الصغيرة قد تؤدي أيضاً أطوال الاختبارات القصيرة إلى نتائج مظلمة أحياناً، فمثلاً الفقرة التي رمزها في الملف الأصلي ١٢٢ أظهرت أداءً تفضلياً غير منتظم عند طول اختبار ٢٠، ٣٠ عند كل مستويات حجم العينة، في حين أن الفقرة نفسها لم تظهر كذلك عند طول الاختبار الصغير ١٠ وإنما أظهرت أداءً تفضلياً منتظماً، والفقرة ١٣٠ عند طول اختبار ١٠ أبدت أداءً تفضلياً غير منتظم على الرغم أنها لم تبدي أداءً تفضلياً في أطوال الاختبارات ٢٠، ٣٠ عند كل مستويات حجم العينة، وربما يكون آلية وإجراءات الطريقة وحساسيتها في تقدير الأوزان وعدد مرات التدوير (Iterations) للوصول إلى تقديرات دقيقة في حالة طول الاختبار الصغير سبباً وراء ذلك. إجمالاً إن فاعلية طريقة MIMIC لا تتأثر بشكل جوهري بتقلبات حجم العينة وطول الاختبار عند استخدام حجم عينة يفوق ٢٠٠٠، وهذا يقدم صورة واضحة أمام الباحث عندما يستخدم هذه الطريقة في اختيار حجم العينة وطول الاختبار المناسب.

٨. التوصيات

بناء على نتائج الدراسة فإنه يمكن التوصية بما يأتي:

١. ضرورة مراجعة الفقرات التي أبدت أداءً تفضلياً تبعاً لمتغير الجنس في اختبار بيزا الدولي، وإعادة صياغتها، والنظر في التحليل الإحصائي والخصائص السيكمترية لها لتلافي أخطاء القياس وتأثيرها على نتائج المفحوصين.
٢. استخدام حجوم عينات ٢٠٠٠ وأكثر يجعل فاعلية طريقة MIMIC أكثر ثباتاً واستقراراً، لذلك يُنصح بزيادة حجوم العينات لأكثر من ١٠٠٠ إذا كان هدف الباحث الوصول إلى أكبر عدد ممكن من الفقرات التي تبدي أداءً تفضلياً في مثل هذا النوع من الاختبارات.
٣. استخدام طريقة MIMIC في حالة تقصي الأداء التفاضلي غير المنتظم، لكن مع ضرورة استخدام حجوم عينات وأطوال اختبار مناسبة.

بيان تضارب المصالح

يقر جميع المؤلفين أنه ليس لديهم أي تضارب في المصالح.

المراجع

المركز الوطني لتنمية الموارد البشرية. (٢٠٢٠). التقرير الوطني الأردني عن الدراسة الدولية بيزا لعام ٢٠١٨. وحدة المتابعة والتخطيط، عمان، الأردن.

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.

- 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.
- Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(4), 561-579.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. London: SAGE Publications.
- Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, 76(1), 43-63.
- 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.
- Hamilton, L. (1999). Detecting gender-based differential item functioning on constructed response science tests. *Applied Measurement in Education*, 12(3), 30-56.
- Jamali, J., Ayatollahi, M., & Jafari, P. (2017). The effect of small sample size on measurement equivalence of psychometric questionnaires in MIMIC model: A simulation study. *Shiraz University of Medical Sciences, Iran*. Article ID 7596101, 12 pages.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351), 631-639.
- Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational Psychological Measurement*.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7(2), 105-118.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297-334.
- Montoya, A. K., & Jeon, M. (2019). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, 44(2), 118-136.
- Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Retrieved from <http://statmodel2.com/download/webnotes/CatMGLong.pdf>
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.
- National Center for Human Resources Development. (2020). *The Jordanian national report on the international study PISA for the year 2018*. Follow-up and Planning Unit, Amman.
- Osterlind, S. J. (1983). *Test item bias*. Beverly Hills: SAGE Publications.
- Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.

-
- Uğurlu, 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.
- Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, 38(2), 147-163.
- Wall, D., & Horak, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe – Phase 2. Coping with change*. Princeton, NJ: Educational Testing Service.
- Williams, S. (1997). The unbiased anchor bridging the gap between DIF and item bias. *Applied Measurement in Education*, 10(3), 253-267.
- Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1), 1-27.
- Woods, C. M., & Grimm, K. J. (2011). Testing for non-uniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35(5), 339–361.