

**taqdeer ma'alm almafrdaat la'imthaan maada addaraasaat alaaajtmaa'ya
(hazhaa watnee) latlba dabloum atta'leem al'aam: daraasa maqaarna
bayn annadhrya alkalaaseekya wandhrya al'istjaaba lalmfrda.**

تقدير معالم المفردات لإمتحان مادة الدراسات الاجتماعية (هذا وطني) لطلبة دبلوم التعليم العام: دراسة
مقارنة بين النظرية الكلاسيكية ونظرية الإستجابة للمفردة

**Estimating Item Parameters for the Social Studies Exam for General
Diploma Students: A Comparative Study Between Classical Test
Theory and Item Response Theory**

الغالية الهنائية¹، و أ.د. حسين الخروصي² Alghalia Alhanaia Hussein Alkhrossi

¹ جامعة السلطان قابوس، طالبة دكتوراة قياس وتقويم، s114070@student.squ.edu.om

² قسم علم النفس، كلية التربية، جامعة السلطان قابوس، سلطنة عُمان، hussein5@squ.edu.om

تاريخ الاستلام: 2024/12/31 تاريخ القبول: 2025/10/14 تاريخ النشر: 2025/12/30

Abstract: This study aimed to evaluate the Social Studies test by comparing two measurement approaches: Classical Test Theory (CTT) and Item Response Theory (IRT). The objective was to identify which method offers more accurate and objective estimations of test item parameters, specifically item difficulty and discrimination. The study sampled 50,693 General Diploma students from the first semester of the 2022-2023 academic year. Results showed strong alignment between CTT and IRT in item difficulty estimates, but IRT proved more accurate in identifying low-discrimination items. The study recommends using IRT for more precise and objective test analysis and highlights the need for balanced item difficulties to measure diverse student abilities effectively.

Keywords: Classical Test Theory, Item Response Theory, Social Studies Exam.

ملخص: هدفت الدراسة إلى تحليل إمتحان مادة الدراسات الاجتماعية باستخدام كل من نظرية القياس الكلاسيكية CTT ونظرية الإستجابة للمفردة IRT لتحديد النظرية الأكثر دقة وموضوعية في تقدير معالم مفردات الاختبار من حيث الصعوبة والتمييز. شملت عينة الدراسة من 50693 طالباً وطالبة، وهم جميع طلبة دبلوم التعليم العام المتقدمين لإختبار الدراسات الاجتماعية للفصل الدراسي الأول من العام الدراسي 2022-2023، واستخدمت الدراسة المنهج الوصفي، ومن أهم نتائج الدراسة توافق تقديرات صعوبة المفردات بين النظرية الكلاسيكية ونظرية الإستجابة للمفردة، وبالنسبة للتمييز اتفقت النظريتين في تصنيف بعض المفردات، إلا أن نظرية الإستجابة للمفردة أظهرت دقة أكبر، خاصة في تحديد المفردات ذات التمييز المنخفض، لذلك توصي الدراسة باستخدام نظرية الإستجابة للمفردة لتحليل الإختبارات لكونها أكثر دقة وموضوعية، وضرورة الإنتباه عند تصميم الإختبارات بأن تشمل توزيعاً متناسباً للمفردات عبر مستويات الصعوبة لضمان تحقيق تقييم شامل ودقيق لقدرات الطلبة المختلفة.

الكلمات المفتاحية: النظرية الكلاسيكية، نظرية الإستجابة للمفردة، إمتحان مادة الدراسات الاجتماعية.

المؤلف المرسل: الغالية الهنائية

المقدمة

يُعدّ التقويم التربوي أحد الركائز الأساسية في تطوير العملية التعليمية وتحسين مخرجاتها. فهو يمثل الوسيلة التي يتم من خلالها قياس مدى تحقيق الأهداف التعليمية وتحديد نقاط القوة والضعف في الأداء التعليمي، مما يُساعد على اتخاذ قرارات مبنية على بيانات موثوقة تساهم في تحسين جودة التعليم (علام، 2007). من بين أدوات التقويم المتعددة: الاختبارات التحصيلية حيث تُستخدم على نطاق واسع لقياس التحصيل العلمي للمتعلمين في مختلف المراحل التعليمية. فقد عرّفها الزغول (2001، ص327) بأنها "مجموعة من الأسئلة وضعت لقياس مدى تحقيق الأهداف التعليمية المحددة مسبقاً لدى المتعلمين"، وتعتمد عليها الأنظمة التربوية لاتخاذ قرارات مؤثرة في مسارات الطلبة ومستقبلهم التعليمي والمهني.

تأتي أهمية تصميم الاختبارات التحصيلية من دورها المحوري في تقديم تشخيص دقيق لمستويات الطلبة، مما يُساعد في اتخاذ القرارات المناسبة، سواء فيما يتعلق بتطوير المناهج أو بتقديم برامج علاجية للمتعلمين الذين يحتاجون إلى دعم إضافي (علام، 2002). ومع ذلك يواجه مصممو الاختبارات تحديات متعددة تتعلق بالتحقق من دقتها وموضوعيتها، وهو ما دفع الباحثين إلى تطوير نماذج قياس أكثر كفاءة للتغلب على مشكلات النظرية الكلاسيكية (Classical Test Theory - CTT) في القياس (Hambleton et al., 1991). تعتمد النظرية الكلاسيكية على تحليل الخصائص السيكومترية للاختبار بشكل كامل، مما يجعل دقة نتائجها مرتبطة بخصائص العينة المستخدمة في الاختبار، وبالتالي يتأثر الثبات والصدق في هذه النظرية بتركيبية العينة وبطبيعة المفردات، حيث يمكن أن يؤدي الاختبار السهل إلى زيادة وتضخيم مستويات أداء الطلبة، بينما يحدث العكس مع الإختبارات الصعبة، بالإضافة إلى ذلك يتم حساب الخطأ المعياري على مستوى الاختبار ككل، مما قد يؤدي إلى تحيز في النتائج ويقلل من دقتها (Hambleton et al., 1991). ولمواجهة هذه التحديات، ظهرت النظرية الحديثة أو نظرية الاستجابة للمفردة (Item Response Theory - IRT) كنموذج أكثر دقة ومرونة، حيث تقوم بتحليل المفردات على مستوى أكثر تفصيلاً وتُقدم تقديرات مستقلة عن عينة الأفراد، مما يُعزز من موضوعية ودقة القرارات التربوية (Hambleton & Jones, 1993) (Hambleton et al., 1991).

تمثل نظرية الاستجابة للمفردة ثورة في مجال القياس والتقويم للمفردة؛ إذ تُوفّر أدوات متقدمة لتحليل المفردات وتحديد خصائصها السيكمترية بدقة، فهي تقوم على نماذج رياضية لتقدير قدرات الأفراد بدقة بناءً على استجاباتهم لفقرات الاختبار، وتهتم بدراسة المفردات الاختبارية ومعرفة خصائص كل منها، كما تعتمد على العلاقة بين قدرة الفرد واحتمال استجابته استجابة صحيحة على المفردة، حيث تقوم النظرية على مبدأ أن لكل شخص سمة كامنة (قدرة غير واضحة) يمكن التعرف عليها من خلال أدائه في الاختبار المخصص لتقدير تلك القدرة، أي أن الهدف الرئيسي من تطبيق الاختبار وفقاً لهذه النظرية هو تحديد مكان الشخص على مقياس القدرة (Baker & Kim, 2004; Baker, 2001). ومن بين الفوائد الرئيسية لهذه النظرية قدرتها على تقديم معلومات دقيقة عن أداء الطلبة على كل مفردة اختبارية من خلال دالة معلومات المفردة (IIF)، وكذلك معلومات على مستوى الاختبار ككل من خلال دالة معلومات الاختبار (TIF)، والتي بدورها ساهمت في تطور الاختبارات وأدوات القياس كما قدمت أيضاً حلول كثيرة للاختبارات والتي عجزت النظرية الكلاسيكية في تقديمها مثل تعميم الاختبار والتعرف على المفردات المتحيزة وإنشاء بنوك للأسئلة كما ساهمت في ظهور الاختبارات التكيفية المحوسبة (Hambleton et al., 1991).

في ضوء هذه التطورات، يُعد تطبيق امتحان مادة الدراسات الاجتماعية (هذا وطني) فرصة لتحليل الأداء باستخدام كل من النظرية الكلاسيكية ونظرية الاستجابة للمفردة. حيث تمثل مادة الدراسات الاجتماعية جزءاً من المقررات الدراسية في دبلوم التعليم العام في سلطنة عُمان، وتهدف إلى تنمية روح المواطنة وتعزيز الانتماء للوطن، من خلال زيادة وعي الطلبة بالتاريخ الحضاري لعمان، وتعزيز الوعي بطبيعة عمان الجغرافية وتأثيرها في صياغة الشخصية العمانية. وبالتالي فإن إجراء تحليل دقيق لمفردات الإمتحان يُسهم في تحسين جودته وتطوير أساليب التقويم المتبعة في النظام التعليمي لسلطنة عُمان. لذلك تسعى هذه الدراسة إلى تحليل ومقارنة نتائج الاختبار (الإمتحان) في ضوء النظرية الكلاسيكية (CTT) للاختبارات ونظرية الإستجابة للمفردة (IRT) لتحديد النظرية الأكثر دقة وموضوعية في تقدير معالم الاختبار.

مشكلة الدراسة

تتمثل مشكلة هذه الدراسة في الحاجة إلى تحديد الفروقات الجوهرية بين النظرية الكلاسيكية للقياس (CTT) ونظرية الاستجابة للمفردة ((IRT، وتقييم مدى تكافؤ استخدامهما في تحليل اختبار "هذا وطني" الخاص بطلبة دبلوم التعليم العام في سلطنة عُمان. على الرغم من أن نظرية CTT تتميز بسهولة تطبيقها وملاءمتها للتقييمات السريعة، إلا أنها تعاني من قيود تتمثل في اعتماد نتائجها على العينة، مما يؤدي إلى انخفاض دقة التقديرات في قياس القدرات المتنوعة (Hambleton & Jones, 1993; Manjunath & Banerjee, 2022). في المقابل تقدم نظرية IRT مزايا تتمثل في تقدير معالم المفردات بشكل مستقل عن العينة، مما يعزز من ثبات التقديرات وموضوعيتها عبر عينات متعددة (Manjunath & Banerjee, 2022). فعلى الرغم من المزايا النظرية التي توفرها نظرية IRT إلا أن دراسة بيتشي وآخرون (Bichi et al., 2015) لم تجد دليلاً يثبت تفوق نظرية IRT على نظرية CTT، رغم تفوقها النظري في بعض الجوانب. ولهذا السبب أجريت العديد من الدراسات والتي اختلفت في نتائجها حول قدرة كل من النظريتين على توفير معالم مفردات دقيقة لتقييم الطلبة. فتوصلت بعض الدراسات إلى وجود توافق بين النظريتين في تقدير معالم الصعوبة، لا سيما في النماذج أحادية وثنائية المعلم PL1 و PL2؛ بينما ظهر تباين أكبر في النموذج ثلاثي المعالم (Bichi et al., 2015; Manjunath & Banerjee, 2022; Subali et al., 2021؛ الصيخان والمومني، 2021؛ الدوسري، 2019). في حين أثبتت دراسة (الصيخان والمومني، 2021؛ الدوسري، 2019؛ Subali et al., 2021). أن نظرية IRT تقدم معالم تمييز أدق مقارنة بنظرية CTT، مما يعزز من دقة النتائج عند وجود تنوع في مستويات الصعوبة، ومع ذلك يتطلب هذا توافقاً جيداً بين النموذج والبيانات كما يتطلب عينات كبيرة ومعالجة رياضية معقدة لتحقيق تقديرات موثوقة (Bichi et al., 2015; Ukofia & Essen, 2021).

بناءً على ماسبق تسعى هذه الدراسة إلى تحليل اختبار "هذا وطني" باستخدام كل من النظرية الكلاسيكية CTT ونظرية الاستجابة للمفردة IRT لتحديد النظرية الأكثر دقة وموضوعية في تقدير معالم الاختبار. كما تسعى إلى تقديم توصيات عملية لتطوير

أدوات القياس في النظام التعليمي بسلطنة عُمان، بما يسهم في تعزيز جودة التعليم ودعم القرارات التربوية.

أهمية الدراسة

تتمثل أهمية الدراسة في تناولها لجانب مهم وهي توفير اختبارات على مستوى من الدقة والموضوعية في القياس لمرحلة مهمة من مراحل التعليم وهي مرحلة دبلوم التعليم العام حيث سيتم عرض الأهمية التطبيقية والنظرية للدراسة، كما يلي:

أولاً: الأهمية التطبيقية

1-1 توفير تغذية راجعة لمعدي الاختبارات وأخصائيي التقويم: حيث يساهم

التحليل في تحديد مدى صعوبة الفقرات وقدرتها على التمييز، وبالتالي يساعد في تحديد الفقرات التي تحتاج إلى تعديل أو استبعاد. وبالتالي يتم بناء اختبارات موثوقة وأكثر عدالة وتكافؤاً بين الطلبة، ويقلل من الانحياز ويعزز من مصداقية التقويم.

2-1 دعم المعلمين في تحسين الأداء التعليمي: من خلال توفير تغذية راجعة

مباشرة للمعلمين حول نقاط القوة والضعف لدى الطلبة، مما يمكنهم من تطوير أساليب التدريس ووضع خطط علاجية وإثرائية تلائم احتياجات الطلبة المختلفة حسب قدراتهم ومستوياتهم.

3-1 مساعدة معدي المناهج في تطوير المناهج الدراسية: يقدم التحليل رؤى

مهمة حول فعالية المناهج الحالية، مما يدعم تعديل وتطوير المناهج بما يتماشى مع احتياجات الطلبة ومعايير الجودة.

4-1 المعهد التخصصي للتدريب المهني للمعلمين: يساهم التحليل في مساعدة

المعهد التخصصي على وضع خطط تدريب متخصصة للجهات المعنية ببناء الاختبارات مستندة إلى نتائج التحليل.

5-1 توفير تغذية راجعة للمشرفين التربويين: يمكن للمشرفين الاستفادة من التحليل

في توجيه المعلمين وتقديم الدعم الفني المناسب، بالإضافة إلى وضع استراتيجيات إشراف فعّالة.

ثانياً: الأهمية النظرية

2-1 مقارنة النظريتين في القياس

تقدم الدراسة مقارنة علمية بين النظرية الكلاسيكية للقياس ونظرية الإستجابة للمفردة، مما يمكن الباحثين وصناع القرار من اختيار النهج الأنسب لتطوير اختبارات تتمتع بالدقة والصدق والموضوعية والثبات. كما تفيد هذه المقارنة في فهم مدى دقة وكفاءة كل نظرية في تقدير معالم المفردات، خصوصاً في الاختبارات النهائية.

2-2 إثراء البحث العلمي في مجال القياس والتقييم

تسهم الدراسة في سد الفجوة في الأدبيات المتعلقة بتطبيق النظريات الحديثة في القياس (IRT) على اختبارات المواد الدراسية في سياقات تعليمية عربية، مثل سلطنة عُمان. حيث من المؤمل أن تكون هذه الدراسة مرجعاً لمزيد من الدراسات المستقبلية التي تهدف إلى تحسين أدوات التقييم في التعليم العام.

2-3 دعم صانعي القرار في تطوير السياسات التعليمية

توفر الدراسة بيانات دقيقة تساعد صناع القرار في تطوير سياسات تقييم فعالة ومبنية على معايير علمية، مما يعزز من جودة التعليم.

أهداف الدراسة

تتمثل أهداف الدراسة الحالية في ما يلي:

- التأكد من مطابقة مفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) لنظرية القياس الكلاسيكية.
- استخراج مؤشرات الصعوبة والتمييز لمفردات الاختبار وفقاً لنظرية القياس الكلاسيكية.
- التأكد من مطابقة مفردات الاختبار لافتراضات نظرية الإستجابة للمفردة.
- استخراج مؤشرات الصعوبة والتمييز لمفردات الاختبار وفقاً لنظرية الإستجابة للمفردة.
- مقارنة معاملات الصعوبة والتمييز المستخرجة بالنظريتين.

أسئلة الدراسة

تسعى الدراسة الحالية للإجابة عن الأسئلة التالية:

السؤال الأول : ما مدى مطابقة مفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) لنظرية القياس الكلاسيكية (CCT)؟

السؤال الثاني : ما مؤشرات الصعوبة والتمييز لمفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) وفقاً لنظرية القياس الكلاسيكية (CCT)؟

السؤال الثالث: ما مدى مطابقة مفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) لافتراضات نظرية الإستجابة للمفردة (IRT)؟

السؤال الرابع : ما مؤشرات الصعوبة والتمييز لمفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) وفقاً لنظرية الإستجابة للمفردة (IRT)؟

السؤال الخامس: هل يوجد فرق في استخراج معاملات الصعوبة والتمييز المستخرجة بالنظريتين (CCT) و (IRT)؟

حدود الدراسة

الحدود الموضوعية والزمانية

ركزت الدراسة على تحليل استجابات طلبة دبلوم التعليم العام لاختبار مادة الدراسات الاجتماعية (هذا وطني) فقط، والمتعلق بنهاية الفصل الدراسي الأول للعام الدراسي 2022-2023م.

الحدود المكانية

اقتصرت التحليل على طلبة دبلوم التعليم العام، وهم طلبة الصف الثاني عشر في المدارس الحكومية فقط داخل سلطنة عمان، دون التطرق إلى مدارس القطاع الخاص أو المؤسسات التعليمية الأخرى.

الإطار النظري والدراسات السابقة

أولاً: النظرية الكلاسيكية للقياس (CTT)

1-1 النظرية الكلاسيكية هي إطار نظري يهدف إلى تفسير نتائج الاختبارات عن طريق ثلاثة مفاهيم رئيسية (Rust et al., 2021):

- الدرجة الظاهرة (Observed Score): وهي الدرجة التي يحصل عليها المفحوص في الاختبار.

- الدرجة الحقيقية (True Score): وهي التقدير الدقيق لأداء المفحوص لو كان الاختبار خالياً من الأخطاء.
- درجة الخطأ (Error Score): وهي الفروقات التي تنشأ عن عوامل عشوائية تؤثر على الأداء في الاختبار.

يتم التعبير عن هذه العلاقة الأساسية في CTT بالصيغة الرياضية: $X=T+E$
X: الدرجة الظاهرة، T: الدرجة الحقيقية، و E: الخطأ.

تعتمد هذه النظرية على افتراضات معينة منها:

- الخطأ في القياس يكون موزعاً بشكل طبيعي وله متوسط يساوي صفر.
- لا توجد علاقة ارتباط بين الأخطاء والدرجات الحقيقية.
- الأخطاء في اختبارين متوازيين غير مرتبطة.

1-2 خصائص المفردات (الصعوبة، التمييز)

الصعوبة (Item Difficulty): يتم التعبير عنها بنسبة الإجابات الصحيحة على المفردة أي نسبة الأفراد الذين أجابوا على المفردة بشكل صحيح، ويُشار إليها بـ p-value، بحيث إذا قلت النسبة دل على صعوبة المفردة.
وذكر رست وآخرون (Rust et al., 2021) أنه إذا كانت قيمة صعوبة المفردة أقل من 0,25 فتكون مفردة صعبة، وإذا كانت بين 0,25 - 0,75 فمتوسطة الصعوبة، وإذا زادت عن 0,75 فتكون مفردة سهلة.

التمييز (Item Discrimination): يعني أن المفردة قادرة على التمييز بفعالية بين الممتحنين ذوي الدرجات العالية والدرجات المنخفضة. حيث يتم قياسه باستخدام معاملات الارتباط بين أداء الأفراد على المفردة وأدائهم الكلي في الاختبار. يُستخدم عادةً معامل الارتباط (corrected item-total correlation)، وكلما كان هذا المعامل أكبر، كلما كانت الفقرة قادرة بشكل أفضل على التمييز بين الأفراد (Hambleton & Jones, 1993). ويجب أن لا تقل قيمة التمييز عن 0,20 Rust et al., (2021)، ويتم تصنيف قيمة التمييز حسب ما ورد في (Bichi, 2015)، والذي تم عرضه في جدول 1:

جدول 1: معايير الحكم على قيمة التمييز حسب النظرية الكلاسيكية للقياس (CTT)

قيمة التمييز	الحكم على التمييز	جودة الفقرة
أعلى من 0,40	تمييز مرتفع جداً	الفقرة مميزة جداً
0,30 - 0,39	تمييز مرتفع	تحتاج إلى قليل من المراجعة أو لا تحتاج
0,20 - 0,29	تمييز متوسط	تحتاج إلى مراجعة
أقل من 0,20	تمييز ضعيف	ينبغي حذف الفقرة أو مراجعتها

ثانياً: النظرية الحديثة للقياس أو نظرية الاستجابة للمفردة (Item Response Theory - IRT)

1-2 نظرية الاستجابة للمفردة (Item Response Theory - IRT) تقوم على

نماذج رياضية لتقدير قدرات الأفراد بدقة بناءً على استجاباتهم ل فقرات الاختبار، وتهتم بدراسة المفردات الاختبارية ومعرفة خصائص كل منها، كما تعتمد على العلاقة بين قدرة الفرد واحتمال استجابته استجابة صحيحة على المفردة، حيث تقوم النظرية على مبدأ أن لكل شخص سمة كامنة (قدرة غير واضحة) يمكن التعرف عليها من خلال أدائه في الاختبار المخصص لتقدير تلك القدرة، (Baker & Kim, 2001; Baker, 2004). تهدف هذه النظرية إلى تحسين تقدير أداء الأفراد من خلال تحليل الاستجابات على مستوى المفردة، مما يوفر تقديرات دقيقة ومستقلة عن العينة المستخدمة في الاختبار (Rust et al., 2021).

2-2 افتراضات النظرية (أحادية البعد، الاستقلال الموضوعي، التحرر من السرعة)

أحادية البعد Unidimensionality: يمكن تعريف أحادية البعد على أن جميع مفردات الاختبار تقيس سمة واحدة فقط، أي وجود عامل واحد يؤثر على أداء الاختبار وهو القدرة أو السمة التي يسعى الاختبار لقياسها (Hambleton et al., 1991).

الاستقلال الموضوعي Local Independence:

هو أن تكون استجابات الفرد للبند المختلفة في الاختبار مستقلة، وتتحدد استجابته فقط بموقع الشخص على المتصل الكامن وهذا يعني أن لا تؤثر استجابة الفرد لإحدى المفردات على استجاباته للمفردات الأخرى، ويعد افتراض الإستقلال الموضوعي ضروري تحقيقه في بيانات الاختبار (أداة القياس) حيث أن عدم تحقيقه يؤدي إلى عدم القدرة على استخدام نماذج أحادية البعد لتقدير البارامترات وكذلك

سيؤدي إلى التأثير على تقديرها؛ ولربما ستكون القيم المقدرة أكبر من قيمها في الحقيقة وبالتالي تزداد أخطاء القياس (علام، 2005).

وذكر العديد من الباحثين (Hambleton et al., 1991; Ayla, 2009; DeMars, 2010) أن افتراض الاستقلال الموضوعي يتحقق إذا تحقق افتراض أحادية البعد، وفي بعض الأحيان يمكن اعتبار افتراض أحادية البعد وافتراض الاستقلال الموضوعي شيئاً واحداً إلا أنه في حالات أخرى لا يمكن اعتبار ذلك أي أن افتراض الاستقلال الموضوعي لا يتحقق ويتم انتهاكه؛ ولتلافي ذلك يفضل الانتباه في مرحلة تصميم المفردات سواء للمقياس أو الاختبار من خلال البحث عن فقرات الاختبار (أداة القياس) التي تكون مرتبطة جميعها بنفس النص، أو عندما تكون الفقرات مرتبة بشكل متتابع بحيث إجابة الفقرة تعتمد على إجابة الفقرة التي تسبقها، وكذلك إذا وجدت مفردات تتضمن تلميح لإجابة مفردة أخرى، ففي هذه الحالة يمكن إعادة صياغة المفردات؛ إلا أنه في حالات أخرى لا يمكن إعادة صياغتها لأن طبيعة المفردات يتطلب وجود ترابط منطقي بينها، وفي هذه الحالة يمكن أن تشكل هذه المفردات اختبار فرعي، وأيضاً لا يتحقق افتراض الاستقلال الموضوعي في الاختبارات التي تعتمد على السرعة أو الاختبارات الطويلة (Ayla, 2009).

التحرر من السرعة (Speediness): وتعني أن استجابة الفرد للمفردات تتوقف على مقدار ما يمتلكه الفرد من القدرة أو السمة التي يراد قياسها، ولا تتأثر استجابته نتيجة لتأثير السرعة؛ وهذا يدل على أن عدم إجابة الفرد للمفردات إجابة صحيحة يعود إلى انخفاض قدرته وليس لتأثير سرعة الإجابة (Hambleton & Swaminathan, 1985).

2-3 خصائص المفردات (الصعوبة، التمييز)

صعوبة المفردة (Item Difficulty): صعوبة المفردة يرمز لها بالرمز b وهي نقطة على متصل القدرة أو السمة يكون عندها احتمال الإجابة عن المفردة إجابة صحيحة يساوي $(1+c)/2$ ؛ حيث ترمز c إلى معلمة التخمين (Baker, 2001). وذكر Hambleton et al., 1991 أنه إذا تم إجراء تحويل خطي للقدرة بمتوسط حسابي = صفر، وانحراف معياري = 1؛ فإن القيم العملية للصعوبة تتراوح بين -2 و 2 بحيث تشير القيم القريبة من -2 إلى سهولة المفردة والقيم القريبة من 2 إلى صعوبة المفردة،

وتم تصنيف قيم صعوبة المفردة حسب (Baker & Kim, 2004; Baker, 2001)، والتي تم عرضها في جدول 2 كما يلي:

جدول 2: معايير الحكم على قيمة صعوبة المفردة حسب نظرية الاستجابة للمفردة

الحكم على صعوبة المفردة	قيمة الصعوبة
سهلة جدا	أقل من (2-)
سهلة	(2-) - (0,5-)
متوسطة الصعوبة	(0,5) - (0,5-)
صعبة	(0,5) - (2)
صعبة جدا	أعلى من (2)

تمييز المفردة (Item Discrimination): يرمز له بالرمز a ويعبر عنه من خلال ميل منحنى خصائص المفردة عند مستوى القدرة المقابل لصعوبة المفردة فتمييز المفردة يتناسب طرديا مع ميل المنحنى المميز لها، وبالنسبة للاختبارات ينبغي أن يكون التمييز موجب وهو يشير إلى أن احتمال حصول الفرد على إجابة صحيحة يزداد مع قدرته (Baker, 2001). والجدول 3 يوضح معايير الحكم على قيمة تمييز المفردة ما ورد في بيكر (2001) Baker .

جدول 3: معايير الحكم على قيمة تمييز المفردة حسب نظرية الاستجابة للمفردة

الحكم على تمييز المفردة	قيمة التمييز
تمييز ضعيف جدا	(0,01) - (0,34)
تمييز ضعيف	(0,35) - (0,64)
متوسطة التمييز (مقبول)	(0,65) - (1,34)
تمييز عال	(1,35) - (1,69)
تمييز عال جدا	أعلى من (1,70)

الدراسات السابقة

هدفت دراسة مانونات وبانيرجي (Manjunath & Banerjee (2022) إلى تقييم مدى التكافؤ بين النظرية الكلاسيكية (CTT) للاختبارات ونظرية الإستجابة للمفردة (IRT) في تقدير معالم الفقرات مثل: الصعوبة والتمييز والثبات، كما هدفت إلى تحديد ما إذا كانت الفوائد النظرية لنظرية IRT تتفوق على سهولة تطبيق نظرية CTT. حيث قامت الدراسة بتحليل ومراجعة 16 دراسة تجريبية أجريت بين

عامي 1991 و2014، أظهرت النتائج أن معلمات الصعوبة كانت متقاربة بين النظريتين، وخاصة في النماذج أحادية وثنائية المعلم، مع وجود اختلاف عند استخدام نماذج ثلاثية المعالم، كما توصلت الدراسة إلى أن معلمات التمييز في IRT كانت أكثر دقة من نظرية CTT عند تنوع مستوى الصعوبة. وأشارت الدراسة إلى أن معلمات نظرية IRT أظهرت ثباتاً أعلى من معلمات نظرية CTT عند اختلاف العينات الجغرافية وقدرات الطلاب، لكن هذا الثبات كان مشروطاً بتوافق النموذج مع البيانات. لذا أوصت الدراسة بإجراء مزيد من الأبحاث باستخدام عينات أكبر ومن مناطق متنوعة لاختبار استقرار المعلمات بشكل أعمق، مشيرة إلى أن استخدام نظرية CTT يمكن أن يكون مناسباً للاختبارات الصغيرة والتطبيقات السريعة، في حين أن استخدام نظرية IRT يكون أكثر فاعلية في الاختبارات واسعة النطاق التي تتطلب تقديرات دقيقة وثابتة. (Manjunath & Banerjee, 2022; Ukofia & Essen, 2021).

وبالمثل هدفت دراسة أكوفيا وإيسن (Ukofia & Essen, 2021) إلى مقارنة صعوبة وتمييز المفردات باستخدام النظرية الكلاسيكية (CTT) والنظرية الحديثة للاستجابة للمفردة (IRT)، حيث شملت العينة 5192 استجابة، اختيرت عشوائياً من بين استجابات المتقدمين لاختبار الرياضيات (UTME) لعام 2020 في نيجيريا، والذي يتضمن 50 مفردة، وتم استخدام برنامج BILOG-MG3 لاستخراج المؤشرات، بالاعتماد على نموذج ثنائي المعالم، كشفت النتائج ظهور اختلافات في تصنيف صعوبة المفردات؛ إذ صنفت النظرية الكلاسيكية 36 مفردة بأنها سهلة، 6 مفردات متوسطة الصعوبة، ولم تصنف أي مفردة على أنها صعبة؛ بينما باستخدام نظرية الاستجابة للمفردة تم تصنيف 4 مفردات فقط بأنها سهلة وهي نفس المفردات السهلة في النظرية الكلاسيكية- ولم يتم تحديد أي مفردة ضمن المفردات متوسطة الصعوبة، بالإضافة إلى ذلك، تم استبعاد 3 مفردات غير ملائمة. وبالنسبة لتمييز المفردات أظهرت النظريتان توافقاً في تصنيف المفردات ذات التمييز المنخفض، إلا أنهما اختلفتا في تصنيف المفردات عبر المستويات الأخرى من التمييز، مما يشكل تحدياً في المقارنة بين النظريتين. وتوصلت الدراسة على إمكانية استخدام كلا النظريتين لتصنيف المفردات؛ إلا أن نظرية الاستجابة للمفردة قدمت معلومات أكثر دقة وموثوقة

لتقييم المفردات وتحديد المفردات غير الملائمة مقارنة بالنظرية الكلاسيكية (CTT)، مما يجعل نظرية IRT أكثر فاعلية في تطوير الاختبارات.

كذلك هدفت دراسة سوبالي وآخرون (Subali et al., 2021) إلى مقارنة خصائص المفردات الاختبارية من منظور النظرية التقليدية CTT والنظرية الحديثة IRT للقياس فيما يتعلق بتحديد صعوبة الفقرات وخصائصها، واختبار مدى التباين بين المفردات بناءً على نوع المفردة (اختيار من متعدد (3خيارات أو خيارين)، مفردات صح أم خطأ)، شملت العينة 234 فصلاً دراسياً من الصف الأول إلى السادس، وتم تحليل استجابات الطلاب باستخدام برنامج Quest ، باستخدام نموذج راش، أظهرت النتائج تشابهاً كبيراً بين النظريتين في تصنيف صعوبة الفقرات؛ حيث صُنِّت الفقرات السهلة والصعبة في نظرية CTT بنفس التصنيف في نظرية IRT . كما بينت النتائج أن الفقرات ذات الثلاثة خيارات كانت أصعب من ذات الخيارين وفقاً لنظرية CTT؛ بينما لم يظهر هذا الفرق عند تحليلها باستخدام نظرية IRT ، وخلصت الدراسة إلى إمكانية استخدام أي من الطريقتين في تطوير الاختبارات، وتوصي الدراسة باستخدام نظرية CTT في التقييمات الصفية لسهولة تطبيقها، في حين يُفضل استخدام نظرية IRT في التقييمات واسعة النطاق نظراً لدقتها.

في السياق نفسه هدفت دراسة الصيخان والمومني (2021)، إلى مقارنة النظرية الكلاسيكية (CTT) للاختبارات مع النظرية الحديثة للاستجابة للمفردة (IRT)، لاختبار الفقرات المناسبة في اختبار تحصيلي في مادة اللغة الإنجليزية، شملت العينة 1076 طالبة في المرحلة الثانوية تم اختيارهم بالطريقة العنقودية من 11 مدرسة للطالبات، وتم بناء اختبار مكون من 42 مفردة من نوع اختيار من متعدد بأربع بدائل، وتم استخدام SPSS وبرنامج BILOG-MG3 لاستخراج المؤشرات، بالاعتماد على نموذج ثلاثي المعالم. كشفت النتائج أن كلا النظريتين توفران نتائج متقاربة في اختيار الفقرات، حيث بلغت نسبة الاتفاق 76%. كما أظهرت النتائج تشابهاً في تقدير صعوبة وتمييز الفقرات بين النظريتين CTT وIRT، إلا أن نظرية IRT أظهرت دقة أكبر في تمييز الفقرات ذات التمييز المنخفض، مما يجعلها أكثر فاعلية في التعامل مع الفقرات التي تتطلب تمييزاً دقيقاً في مستويات الصعوبة، كما أظهرت نتائج الدراسة

ميزة نظرية IRT في تقديم تقديرات مستقلة عن العينة مقارنة بنظرية CTT، مما يعزز من دقة نتائج القياس عبر مختلف العينات. أوصت الدراسة بضرورة تطبيق نظرية IRT في تحليل الاختبارات التحصيلية لتوفير تقديرات دقيقة ومستقلة، كما أشارت إلى إمكانية استخدام نظرية CTT في التطبيقات البسيطة التي لا تتطلب تحليلاً عميقاً للبيانات. وبالنسبة لدراسة الدوسري (2019)، فقد هدفت إلى مقارنة بين النظرية الكلاسيكية للقياس ونظرية الاستجابة للمفردة باستخدام نموذج راش في تقدير الخصائص السيكومترية لاختبار القدرات العامة، كما هدفت إلى التحقق من ثبات التقديرات في كلا النظريتين، والكشف عن المفردات غير الملائمة. تضمنت عينة الدراسة من 36910 طالب وطالبة ممن تقدموا لإختبار القدرات العامة، الذي شمل 96 مفردة. أظهرت النتائج توافقاً كبيراً بين تقديرات الصعوبة والتميز في كلا النظريتين، مع ثبات أعلى في التقديرات في نظرية الإستجابة للمفردة، كما أن النتائج في النظرية الكلاسيكية حددت عدداً أكبر من المفردات غير الملائمة مقارنة بنظرية الإستجابة للمفردة، وأوصت الدراسة باستخدام نظرية الإستجابة للمفردة في تطوير الاختبارات لزيادة الموضوعية والدقة وتبني نموذج راش لضمان استقرار التقديرات وفعالية التحليل. إضافة إلى ذلك هدفت دراسة بيشي وآخرون (2015) Bichi et al. إلى مقارنة النظرية الكلاسيكية (CTT) للاختبارات ونظرية الإستجابة للمفردة (IRT) من خلال مراجعة دراسات تجريبية أجريت بين عامي 1998 و2013، بهدف تحليل الفروقات والتشابهات بين النظريتين في تقدير معلمات وخصائص المفردات والقدرات وتطوير الاختبارات، أظهرت النتائج توافقاً كبيراً بين النظريتين في تقدير صعوبة المفردات وتمييزها، إلا أن الدراسة لم تجد دليلاً يثبت تفوق نظرية IRT على نظرية CTT، رغم تفوقها النظري في بعض الجوانب، كما أشارت الدراسة إلى أن معظم الأبحاث السابقة تعاني من قيود متعلقة بحجم العينات وقدم البيانات المستخدمة، مما أثر على موثوقية النتائج. أوصت الدراسة بإجراء أبحاث إضافية باستخدام بيانات حديثة وعينات أكبر لفهم أفضل للفروق بين النظريتين وتطوير أدوات التقييم بشكل أكثر دقة.

وفي الختام هدفت دراسة بيشي وآخرون (2019) Bichi et al. إلى مقارنة النظرية الكلاسيكية (CTT) للاختبارات ونظرية الإستجابة للمفردة (IRT) من خلال تحليل

اختبارات تحصيل الكيمياء، وتقديم رؤى حول مدى ملائمة العناصر المستخدمة في الاختبارات التعليمية في نيجيريا، استخدمت الدراسة المنهج الوصفي لعينة بلغت 530 طالبًا، حيث تم تحليل اختبار الكيمياء الذي يحتوي على 40 مفردة باستخدام برامج مثل X-calibre 4 و ITEMAN 4. كشفت النتائج ملائمة 27 مفردة وتشكل نسبة (67,5%)؛ بينما أن 13 مفردة (32,5%) تحتاج إلى تعديل، كما أظهرت الدراسة ارتباطاً قوياً بين نتائج النموذجين؛ حيث بلغ معامل الارتباط لصعوبة المفردات (-0,985) وللتمييز (0,801)، ويشير هذا إلى أن النظريتين قدمتا نتائج مقاربة وموثوقة؛ إلا أن الدراسة أشارت إلى أن اختبار الكيمياء لم يخضع لعمليات المراجعة والتدقيق. توصي الدراسة بضرورة إخضاع الاختبارات التي يصممها المعلمون لجميع مراحل المراجعة والتدقيق لتحسين جودتها، وتشجع على دمج النظريتين CTT و IRT في تطوير وتحليل العناصر لضمان تقليل أخطاء القياس وتعزيز الموثوقية.

التعليق على الدراسات السابقة

أظهرت الدراسات السابقة التي تناولت المقارنة بين نظريتي القياس فوائد واضحة وحدوداً لكل من النظريتين. على الرغم من أن النظرية الكلاسيكية تمتاز بسهولة التطبيق وملاءمتها للتقييمات البسيطة؛ إلا أنها تعاني من اعتماد تقديراتها على العينة مما قد يؤدي إلى تحيز النتائج عند تنوع قدرات المفحوصين (Manjunath & Banerjee, 2022؛ الدوسري، 2019). من ناحية أخرى، تقدم نظرية IRT تقديرات أكثر دقة واستقلالية عن العينة، مما يعزز من دقة التحليل، لاسيما في الاختبارات واسعة النطاق (Ukofia & Essen, 2021؛ Bichi et al., 2015) ومع ذلك تتطلب نظرية IRT عينات أكبر وحسابات رياضية معقدة، وهو ما قد يحد من سهولة تطبيقها (Subali et al., 2021).

كما أظهرت بعض الدراسات وجود توافق بين تقديرات الصعوبة والتمييز في النموذجين، إلا أن نظرية CTT ميّزت عدداً أكبر من المفردات الغير ملائمة مقارنة بنظرية IRT، مما يعكس حاجة أكبر لمراجعة المفردات في إطار نظرية CTT (الصيخان والمومني، 2021؛ Bichi et al., 2015) في المقابل، أوصت معظم الدراسات بدمج النموذجين لتعزيز جودة التقييمات وتحقيق صدق وثبات أعلى في النتائج، لاسيما مع

تطبيق نظرية IRT في التقييمات التي تتطلب دقة أكبر (Manjunath & Banerjee, 2022; الدوسري، 2019).

منهجية الدراسة

منهج الدراسة

اعتمدت الدراسة الحالية على المنهج الوصفي؛ نظرًا لملاءمته لطبيعتها وأهدافها، حيث يعد المنهج الوصفي أحد أشكال التحليل والتفسير العلمي المنظم لوصف ظاهرة أو مشكلة محددة وتصويرها كمياً عن طريق جمع بيانات ومعلومات عن الظاهرة أو المشكلة وتصنيفها وتحليلها وإخضاعها للدراسة الدقيقة (ملحم، 2005).

مجتمع الدراسة وعينتها

تألفت عينة الدراسة من 50693 طالب وطالبة وهي تمثل نسبة 100% من مجتمع الدراسة، وهم جميع طلبة دبلوم التعليم العام المتقدمين لإختبار الدراسات الاجتماعية (هذا وطني) للعام الدراسي 2022-2023 الفصل الأول - الدور الأول.

أداة الدراسة

ملف بيانات يحتوي على استجابات طلبة دبلوم التعليم العام لامتحان نهاية الفصل الدراسي الأول للعام الدراسي 2022-2023م في مادة الدراسات الاجتماعية (هذا وطني) (مركز القياس والتقويم التربوي، وزارة التربية والتعليم).

حيث احتوى امتحان مادة الدراسات الاجتماعية (هذا وطني) على (45) مفردة، مقسمة على ثلاثة مجالات للمعرفة هي: (المعرفة) و(الفهم والتطبيق) و(القدرات العليا)، وعلى موضوعين وهما: (عمان تاريخ وحضارة) و(بيئتنا تنوع وثراء واستدامة). واحتوى ملف البيانات على عمود به رمز للنوع الاجتماعي ثم يليه أعمدة درجات الطلبة في مفردات الاختبار. يتكون الاختبار من 45 مفردة مقسمة إلى مفردات اختيار من متعدد وعددها 14 سؤال (المفردات من 1- 14)، ومفردات مقالية وعددها 31 سؤال (المفردات من 15- 45).

الدرجة الكلية للاختبار 70 درجة، حيث صحت مفردات اختيار من متعدد بطريقة (0، 1) بحيث يكون للإجابة الصحيحة درجة واحدة والإجابة الخاطئة صفراً. أما

المفردات المقالية فاختلفت في طريقة تصحيحها، والجدول 4 يوضح تدرج تصحيح المفردات المقالية.

جدول 4: تدرج تصحيح المفردات المقالية لإختبار الدراسات الإجتماعية

المفردات	تدرج تصحيح المفردة	درجة المفردة
SC15	0/1/2	2
SC16	0/1/2	2
SC17	0/1	1
SC18	0/1	1
SC19	0/1	1
SC20	0/1	1
SC21	0/1	1
SC22	0/1	1
SC23	0/1	1
SC24	0/1	1
SC25	0/1/2/3/4	4
SC26	0/1/2/3/4	4
SC27	0/1/2/3/4	4
SC28	0/1/2/3/4	4
SC29	0/1	1
SC30	0/1	1
SC31	0/1	1
SC32	0/1	1
SC33	0/2	2
SC34	0/2	2
SC35	0/2	2
SC36	0/2	2
SC37	0/2/4	4
SC38	0/1/2/3/4	4
SC39	0/1	1
SC40	0/1	1
SC41	0/1	1
SC42	0/1	1
SC43	0/1	1
SC44	0/1	1
SC45	0/1/2	2

صدق الأداة

قدم كل من هاردستي وبيردن (2004) Hardesty & Bearden توضيحاً لمفهوم صدق المحتوى الذي يشير إلى مدى تمثيل مفردات الاختبار لعينة مناسبة وشاملة من المحتوى النظري المرتبط بهدف الاختبار، كما أوضحا مفهوم الصدق الظاهري الذي يرتبط بصدق المحكمين (الخبراء)، ويقصد به الدرجة التي يقيم فيها المحكمون ملائمة مفردات الاختبار للمفهوم المستهدف وأهداف التقييم. وعادة يتراوح عدد المحكمين الخبراء بين 3 إلى 10 من الخبراء.

في هذه الدراسة تم التحقق من الصدق الظاهري لإختبار مادة الدراسات الاجتماعية من خلال تشكيل لجنة مكونة من 9 أعضاء، تكونت اللجنة من أحد إخصائي تقويم مادة الدراسات الاجتماعية، عضو من المناهج، عضو من إشراف عام الوزارة، وعضو من إشراف المحافظات التعليمية، وأحد الأكاديميين المختصين في المادة العلمية، وعدد 3 من واضعي الاختبار ومدقق الاختبار، وقامت اللجنة بمراجعة ورقة الاختبار ونموذج الإجابة، بالاستعانة بجدول المواصفات الفنية للاختبار، والمخرجات التعليمية التي وردت في وثيقة التقييم، واشتملت المراجعة على مدى وضوح مفردات الاختبار، وسلامتها العلمية واللغوية، وارتباطها بالمجال المعرفي المحدد في جدول المواصفات، وملائمة درجة كل مفردة مع نموذج الإجابة، بالإضافة إلى أي ملاحظات أخرى تراها اللجنة ضرورية. ويتم أخذ الملاحظات بعين الاعتبار، وتعديل الاختبار بناءً على توصيات اللجنة.

الأساليب الإحصائية

لتحليل البيانات، واستخراج النتائج تم استخدام ثلاثة برامج إحصائية وهي : برنامج X-calibre، وبرنامج SPSS، وبرنامج Iteman.

باستخدام برنامج X-calibre تم تقدير معالم المفردات (الصعوبة، التمييز) لأسئلة الاختيار من متعدد والأسئلة المقالية ثنائية التدرج وهي المفردات رقم (15، 16، 25، 26، 27، 28، 37، 38، 45) باستخدام نموذج ثنائي المعالم؛ وهو الأنسب في هذه الحالة لأننا نهدف إلى مقارنة الصعوبة والتمييز في النظرية الكلاسيكية ونظرية الإستجابة للمفردة، كما أن لدينا يوجد لدينا عينة كبيرة من البيانات حيث يشترط

لاستخدام النموذج الثنائي عينة لا تقل عن 1000، مما ينتج عن ذلك دقة تقديرات المعالم (علام، 2005). وأشار (DeMars, 2010) أن دقة التقدير في النموذج ثلاثي المعالم ممكن أن تلاحظ في العينات التي تصل إلى 2000 مفحوص، ولكنها لا تتحسن بشكل كبير مع عينات أكبر من 4000، وفي هذه الدراسة وصلت العينة لأكثر من 50000، فلذلك دقة التقدير بين النموذج الثنائي والثلاثي قد تكون متقاربة. كما تم استخدام نموذج الاستجابة المتدرجة لتقدير معالم الأسئلة المقالية المتبقية. لمناسبتها للبيانات حيث اختلفت المفردات في تدرجها وذكر (علام، 2005) أن هذا النموذج لا يتطلب أن تشتمل جميع مفرداته على نفس التدرج.

كما تم تقدير معالم المفردات (الصعوبة والتمييز) وكذلك القدرة (θ) بطريقة الاحتمال الأقصى (Maximum Likelihood Estimation - MLE)، وكذلك تم استخدام برنامج Iteman، لإستخراج مؤشرات الصعوبة والتمييز لمفردات الإختيار من متعدد للنظرية الكلاسيكية.

نتائج الدراسة ومناقشتها

السؤال الأول : ما مدى مطابقة مفردات اختبار مادة الدراسات الاجتماعية(هذا وطني) لنظرية القياس الكلاسيكية(CCT)؟

للإجابة على السؤال الأول تم حساب ثبات الاختبار وبلغ معامل الاتساق الداخلي للفقرات (0,928) حسب معامل ألفا كرونباخ، وهو يشير إلى تمتع الامتحان بثبات مرتفع، كما تم استخراج الإحصاءات الوصفية على مستوى الدرجة الكلية للإختبار، والجدول 5 يوضح ذلك.

جدول 5: الإحصاءات الوصفية على مستوى الدرجة الكلية لاختبار مادة الدراسات

الإجتماعية (ن=50693)

الدرجة الكلية	أقل درجة	أعلى درجة	المتوسط الحسابي	الانحراف المعياري	الإلتواء	التفرطح
الدرجة الكلية 100	1,43	100	72,89	20,99	0,70-	0,38-
الدرجة الكلية 70	1	70	51,03	14,694	0,70-	0,38-

* تم تحويل الدرجات إلى التدرج المنوي (الدرجة الكلية = (100/70) * (100))

من جدول 5 يتضح لنا أن البيانات تتوزع توزيع طبيعي وذلك من خلال معاملات الإلتواء والتفرطح فهي في المدى المقبول والذي يتراوح بين -3 و 3 (George& Mallery,2010).

السؤال الثاني: ما مؤشرات الصعوبة والتمييز لمفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) وفقا لنظرية القياس الكلاسيكية (CCT)؟
تم حساب معاملات الصعوبة والتمييز لمفردات اختبار الدراسات الاجتماعية، والجدول 6 يوضح نتائج ذلك.

جدول 6: معاملات الصعوبة والتمييز لمفردات اختبار الدراسات الاجتماعية وفقا لنظرية القياس

الكلاسيكية (CCT)

المفردة	الصعوبة	التمييز	المفردة	الصعوبة	التمييز
SC01	*0.82	0.45	SC24	0.69	0.51
SC02	*0.94	**0.16	SC25	*0.77	0.55
SC03	*0.87	0.32	SC26	*0.76	0.67
SC04	*0.75	0.49	SC27	0.43	0.66
SC05	*0.85	0.39	SC28	0.58	0.50
SC06	*0.83	0.43	SC29	0.57	0.56
SC07	*0.92	0.22	SC30	0.45	0.36
SC08	0.62	0.63	SC31	0.68	0.58
SC09	*0.85	0.35	SC32	0.63	0.61
SC10	0.67	0.51	SC33	0.60	0.65
SC11	*0.88	0.33	SC34	0.64	0.65
SC12	0.58	0.67	SC35	0.58	0.52
SC13	*0.99	**0.02	SC36	*0.87	0.47
SC14	*0.85	0.35	SC37	*0.87	0.61
SC15	0.73	0.67	SC38	*0.93	0.36
SC16	*0.76	0.71	SC39	0.64	0.60
SC17	0.62	0.68	SC40	0.67	0.45
SC18	*0.75	0.67	SC41	*0.78	0.51
SC19	0.63	0.67	SC42	0.68	0.58
SC20	0.41	0.48	SC43	*0.95	0.37
SC21	*0.83	0.56	SC44	*0.85	0.39
SC22	0.62	0.54	SC45	*0.86	0.55
SC23	0.71	0.52			

** معامل التمييز أقل من (0.20).

* مفردات سهلة (أعلى من 0.75)

من جدول 6 نلاحظ ما يلي:

- تراوحت معاملات الصعوبة بين (0,41 - 0,99) ، ولم يتم تصنيف أي مفردات على أنها صعبة، وهذه النتيجة تتفق مع ما توصلت إليه دراسة (Ukofia & Essen (2021)، التي لم تكتشف وجود صعوبة كذلك، حيث صنفت جميع المفردات ضمن المستويين السهلة والمتوسط من حيث الصعوبة.
- تراوحت معاملات التمييز بين (0,02 - 0,71) ، حيث كان معامل تمييز المفردتين (2 ، 13) أقل من (0,20) مما يشير إلى ضعف قدرتهما على التمييز، كما أن المفردتين تصنفان كمفردات سهلة، الأمر الذي أدى إلى تمكن جميع الطلبة تقريبا من الإجابة عليها بشكل صحيح، وبالتالي فقدت المفردتان قدرتهما على التمييز.

السؤال الثالث: ما مدى مطابقة مفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) لافتراضات نظرية الإستجابة للمفردة (IRT)؟

للإجابة على السؤال الثالث تم حساب معامل الثبات للاختبار حسب نظرية الإستجابة للمفردة، حيث بلغ معامل الثبات (0,927)، وهو مؤشر عال لثبات الاختبار، وهذا يدل على موثوقية الاختبار في قدرته على تقدير قدرات الطلبة. حيث يعني أن 92,7% من التباين في الدرجات يعكس القدرة الحقيقية للطلبة؛ بينما 7,3% ترجع لأخطاء القياس (DeMars, 2010).

كما تم التحقق من افتراض أحادية البعد Unidimensionality من خلال إجراء تحليل عاملي استكشافي بطريقة تحليل العوامل الرئيسية Principle Analysis Components للتأكد أن الاختبار يقيس سمة واحدة فقط ، إذ أفرز التحليل وجود 4 عوامل كانت قيم جذرها الكامن EigenValues أكبر من الواحد صحيح، وفسرت جميع العوامل 39,108 % من التباين الكلي.

وتبين من الجدول 7 أن قيمة الجذر الكامن للعامل الأول بلغت 13,007 وهو يفسر ما نسبته 28,904% من التباين الكلي، وهي قيمة مرتفعة إذا ما قورنت مع قيم الجذور الكامنة لبقية العوامل. أما العامل الثاني فقد بلغت قيمة جذره الكامن 2,022 وفسرت ما نسبته 4,49% من التباين، بمعنى أن العامل الأول فسر نسبة عالية من

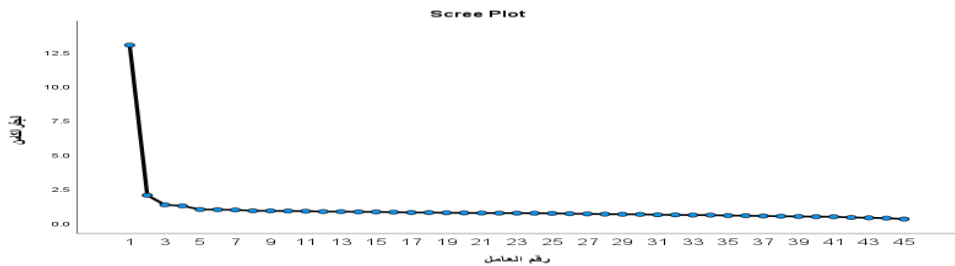
التباين أكبر عن ما تفسره بقية العوامل، وهذا مؤشر على تحقق افتراض أحادية البعد للاختبار، أي أن الاختبار يقيس سمة واحدة فقط

(Hambleton & Swaminathan, 1985). والجدول 7 يوضح نتائج التحليل العاملي.

جدول 7 : نتائج التحليل العاملي لاختبار الدراسات الإجتماعية (ن=50693)

العامل	قيمة الجذر الكامن	نسبة التباين المفسر %	نسبة التباين المفسر التراكمية %
1	13,007	28,904	28,904
2	2,022	4,494	33,397
3	1,325	2,944	36,342
4	1,245	2,767	39,108

كما أمكن التحقق من افتراض أحادية البعد من خلال استخدام ما يعرف باختبار الهضبة Scree Plot ، ويوضح شكل 1 التمثيل البياني لقيم الجذور الكامنة للعوامل المكونة للاختبار على البيانات الكلية، ونلاحظ من الشكل 1 أن هناك مسافة كبيرة بين العامل الواحد والعامل الآخر، أي أن هذا الاختبار له بنية عاملية لأحادية السمة لأن المؤثر هو عامل واحد فقط. (DeMars, 2010)



شكل 1. التمثيل البياني لقيم الجذور الكامنة للعوامل المكونة للاختبار على البيانات الكلية

وبالنسبة لافتراض الإستقلال الموضعي Local Independence يُعتبر هذا الافتراض وثيق الصلة بأحادية البُعد، ويعني هذا الافتراض أنه إذا تحقق افتراض أحادية البعد، فإن استجابات المفحوصين على فقرة معينة لن تعتمد على استجاباتهم على الفقرات الأخرى. وبالتالي، يُعد انتهاك افتراض أحادية البُعد مؤشراً على وجود أكثر من بُعد يؤثر على الأداء، مما يخل بالاستقلالية المحلية (DeMars, 2010)، وبما أن بيانات إختبار مادة الدراسات الاجتماعية تحقق افتراض أحادية البعد، فإنها ستحقق افتراض الإستقلال الموضعي أيضاً Local Independence (Hambleton & Swaminathan, 1985).

أما بالنسبة لافتراض التحرر من السرعة (Speediness)، فإن آلية تطبيق اختبارات دبلوم التعليم العام تتم بتخصيص مدة زمنية ثابتة قدرها 3 ساعات للإختبار مادة الدراسات الاجتماعية، حيث يحصل جميع الطلبة على نفس الفترة الزمنية للإجابة على الأسئلة. ولا يُؤخذ زمن الانتهاء من الإجابة كمعيار في التقييم، مما يعني أن التقييم يعتمد فقط على جودة ودقة الإجابات وليس على سرعة إنجازها (Hambleton et al., 1991).

وبالنسبة لمطابقة النماذج المستخدمة للبيانات فقد تم تقدير معالم الفقرات (الصعوبة والتمييز) وتقدير القدرة (θ) باستخدام النموذج ثنائي المعالم 2PL ونموذج الاستجابة المترتبة Graded Response Model (GRM) بطريقة الاحتمال الأقصى Maximum Likelihood Estimation (MLE) من خلال برنامج X-calibre 4.2. تم قراءة مؤشر مربع كاي (Chi-square) المتاح في البرنامج، لكن هذا المؤشر أظهر معدلاً مرتفعاً، وهذا يظهر في العادة مع العينات الكبيرة، بالإضافة إلى مؤشر مربع كاي يتيح البرنامج نوعين من مؤشرات المطابقة لكل فقرة في الاختبار:

1. مؤشر مربع كاي (Chi-square): يقيس مدى توافق كل فقرة على حدة مع النموذج.

2. البواقي المعيارية (Standardized Residuals - Z Resid): يُحسب من الفرق

بين النسبة الملاحظة للإجابات الصحيحة والنسبة المتوقعة وفق النموذج.

وأظهرت بعض المفردات دلالة إحصائية للمؤشرين، وهو أمر شائع ومقبول بالنسبة للدراسات التي تعتمد على عينات كبيرة فقد بلغت العينة في هذه الدراسة (50693) طالب وطالبة؛ وبالتالي مؤشر مربع كاي والبواقي المعياري يكونان ذو حساسية عالية، مما يعني أن أي انحراف طفيف عن القيم المتوقعة يمكن أن ينتج عنه نتيجة ذات دلالة إحصائية ($p < 0.05$) (Min et al., 2022). لذا في هذه الدراسة تم تقييم ملائمة النموذج من خلال منحنى خصائص المفردة (ICC) الذي تم عرضه في الملحق 1، حيث يوفر المنحنى التوقعات التي يوفرها النموذج لكل مفردة، وتمثل كل نقطة على هذا المنحنى مجموعة كبيرة من المفحوصين الذين يمتلكون تقديرات متقاربة لقدرتهم. فيكون النموذج متوافقاً مع البيانات إذا كانت النسبة الملاحظة للإجابات الصحيحة لكل مجموعة من المفحوصين قريبة من القيم المتوقعة وفقاً للنموذج (DeMars, 2010)، حيث تم التوصل إلى مطابقة جيدة للبيانات والنماذج المستخدمة.

السؤال الرابع : ما مؤشرات الصعوبة والتمييز لمفردات اختبار مادة الدراسات الاجتماعية (هذا وطني) وفقا لنظرية الإستجابة للمفردة (IRT)؟

تم استخراج معاملات الصعوبة والتمييز باستخدام برنامج X-calibre ، ولكي يمكن إستخدام برنامج X-calibre لتحليل IRT تم إعادة ترميز المفردات (SC33, SC34, SC35, SC36) من (0/2) لتصبح (0/1) ، وكذلك المفردة (SC37) من (0/2/4) إلى (0/1/2)، وعند معايرة المعاملات لإستخراجها حذف البرنامج المفردة 13 وذلك لسهولةها العالية وعدم تمييزها، والجدول 8 يوضح معاملات الصعوبة والتمييز لمفردات إختبار الدراسات الاجتماعية وفقا لنظرية القياس الحديثة (IRT).

جدول 8: معاملات الصعوبة والتمييز لمفردات إختبار الدراسات الاجتماعية وفقا

لنظرية القياس الحديثة (IRT)

رقم المفردة	التمييز	الصعوبة	رقم المفردة	التمييز	الصعوبة
SC01	1.02	**1.27-	SC24	0.95	**0.72-
SC02	1.12	**2.05-	SC25	*0.47	**1.11-
SC03	0.65	**2.01-	SC26	0.69	**0.96-
SC04	0.66	**1.21-	SC27	0.69	0.43
SC05	0.81	**1.61-	SC28	*0.30	0.40-
SC06	0.87	**1.39-	SC29	1.12	0.25-
SC07	0.90	**2.05-	SC30	*0.52	0.28
SC08	0.65	**0.56-	SC31	1.13	**0.62-
SC09	0.65	**1.91-	SC32	1.37	0.44-
SC10	*0.44	**1.04-	SC33	1.45	0.33-
SC11	0.87	**1.74-	SC34	1.53	0.43-
SC12	0.67	0.33-	SC35	0.89	0.30-
SC13	-	-	SC36	1.14	**1.48-
SC14	0.68	**1.80-	SC37	1.19	**1.36-
SC15	0.95	**0.77-	SC38	*0.33	**3.44-
SC16	1.28	**0.81-	SC39	1.27	0.47-
SC17	1.95	0.37-	SC40	0.71	**0.74-
SC18	2.11	**0.76-	SC41	1.06	**1.06-
SC19	1.71	0.41-	SC42	1.16	**0.63-
SC20	0.86	0.36	SC43	1.33	**1.98-
SC21	1.51	**1.12-	SC44	0.87	**1.58-
SC22	1.00	0.45-	SC45	0.96	**1.46-
SC23	1.05	**0.76-			

*تمييزضعيف $\geq 0,64$ ، **مفردات سهلة

من جدول 8 نلاحظ ما يلي:

- تراوحت معاملات الصعوبة بين (-3,44 - 0,43) ، ولا توجد مفردات صعبة، وقد اتفقت مع نتيجة التحليل بالنظرية الكلاسيكية، وهي أن أكثر من 50% من المفردات صُنفت على أنها سهلة، مما يعني أن معظم الطلبة قد يجدونها مناسبة وملائمة لمستوياتهم. اللافت أن النظريتين قد اتفقتا على هذا التصنيف، مما يعزز الثقة في دقة تصنيف المفردات، ويشير إلى أن الاختبار يتمتع بتوازن معقول في درجة الصعوبة لصالح فئة كبيرة من الطلبة.
- تراوحت معاملات التمييز بين (0,30 - 2,11)، حيث جاء معامل تمييز المفردات (10، 25، 28، 30، 38) بين (0,01 - 0,64) مما يدل على ضعف قدرتها على التمييز، وقد أكد (DeMars 2010) على أهمية دقة التمييز لكل مفردة، حيث أن المفردة التي لا تميز بشكل كافٍ بين الطلبة ذوي القدرات المختلفة قد لا تكون فعالة في تحقيق هدف التقييم، وقد يعزى ضعف التمييز لهذه المفردات إلى عدم وضوح الصياغة أو إلى سهولتها العالية.

السؤال الخامس: هل يوجد فرق في استخراج معاملات الصعوبة والتمييز المستخرجة بالنظريتين (CCT) و(IRT)؟

تم استخراج معاملات الصعوبة والتمييز لمفردات الاختبار وفقاً لنظرية القياس الكلاسيكية (CTT) ونظرية الاستجابة للمفردة (IRT). يوضح الجدول 9 الإحصاءات الوصفية لمعاملات الصعوبة والتمييز لكل من النظريتين.

جدول 9: الإحصاءات الوصفية لمعاملات الصعوبة والتمييز لمفردات الاختبار وفقاً

لنظرية القياس الكلاسيكية ونظرية الإستجابة للمفردة

النظرية	معاملات	أقل قيمة	أعلى قيمة	المتوسط	الانحراف المعياري
CTT	الصعوبة	0.41	0.99	0.73	0.14
	التمييز	0.02	0.71	0.50	0.15
IRT	الصعوبة	-3.44	0.43	-0.95	0.75
	التمييز	0.00	2.11	0.97	0.42

يتضح من جدول 9 أن تقديرات معاملات الصعوبة حسب نظرية CTT تتراوح بين (0,41) و(0,99) بمتوسط (0,73) وانحراف معياري (0,14)، مما يشير إلى أن غالبية المفردات تقع في مستوى متوسطة الصعوبة إلى سهلة. فقد بلغ عدد المفردات

السهلة 23 مفردة وهي تشكل نسبة (51,1%)، و 22 مفردة متوسطة الصعوبة (48,9%)، ولا توجد مفردات صعبة، وبالنسبة لنظرية IRT تتراوح الصعوبة بين (-3,44) و (0,43) بمتوسط (-0,95) وانحراف معياري (0,75)، بلغ عدد المفردات السهلة 30 مفردة (66,7%)، وعدد المفردات متوسطة الصعوبة 14 مفردة (31,1%)، ولا توجد مفردات صعبة أيضاً، تجدر الإشارة هنا إلى أنه تم استبعاد المفردة 13 من التقدير وذلك لسهولةها وعدم تمييزها، من نتائج جدول 9 نلاحظ وجود تقارب في استخراج معاملات الصعوبة لمفردات الاختبار وفقاً لنظرية القياس الكلاسيكية ونظرية الإستجابة للمفردة، ولتأكيد ذلك تم حساب معامل ارتباط بيرسون لمعاملات الصعوبة بين النظريتين وبلغ معامل الارتباط (-0,927)، وهو دال إحصائياً عند مستوى الدلالة (0,01). ويشير هذا إلى وجود ارتباط قوي بين معاملات الصعوبة في كلا النظريتين، بينما تعكس الإشارة السالبة اختلاف طريقة حساب الصعوبة بينهما، وهذا يتفق مع ما توصلت عليه العديد من الدراسات، (Subali et al., 2021؛ الصيخان والمومني، 2021؛ الدوسري، 2019). عموماً نستطيع القول بأن كلتا النظريتين قادرتان على تقديم تقديرات متقاربة لمعاملات الصعوبة، وإن اختلفت منهجيتهما في التقدير. وبالنسبة للتمييز تم استخراج معاملات التمييز لمفردات الاختبار وفقاً لنظرية القياس الكلاسيكية ونظرية الإستجابة للمفردة وتم عرضها في الجدول 10:

جدول 10: معاملات التمييز لمفردات الاختبار وفقاً لنظرية القياس الكلاسيكية ونظرية

الإستجابة للمفردة

نظرية الإستجابة للمفردة (IRT)		النظرية الكلاسيكية (CTT)		
رقم المفردة	قيمة التمييز	رقم المفردة	قيمة التمييز	الحكم
38،30،28،25،10	0,64 - 0,01	2،13	$0,19 \geq$	تمييز ضعيف
بقية المفردات عددها 32 مفردة (71%)	1,34 - ,65	7	-0,20 0,29	تمييز مقبول
34،33،32،21	1,69-1,35	3،5،9،11،14 38،43،44	-0,30 0,39	تمييز عالٍ
19،18،17	2 - 1,70	بقية المفردات عددها مفردة 34 (75,5%)	$0,40 <$	تمييز عالٍ جدا

كما تم حساب معامل ارتباط بيرسون لمعاملات التمييز بين النظريتين وبلغ معامل الارتباط (0,416)، وهو أيضًا دال إحصائيًا عند مستوى الدلالة (0,01). ورغم أن الارتباط إيجابي، إلا أنه ضعيف، مما يشير إلى أن هناك بعض الفروق في تقدير معاملات التمييز، فنظرية الاستجابة للمفردة (IRT) توفر دقة أكبر في تقدير معاملات التمييز. كما يظهر ذلك في جدول (5)، فقد تم تحديد 32 مفردة ذات تمييز مقبول وفقًا لنظرية IRT، أي بنسبة 71% من المفردات. أما وفق نظرية القياس الكلاسيكية (CCT) قد وُجدت 34 مفردة ذات تمييز عالٍ جدًا، ما يمثل 75,5% من المفردات. وفي المقابل حددت نظرية الاستجابة للمفردة ثلاث مفردات فقط بأنها ذات تمييز ضعيف أو مقبول، وهو ما أكدته الدراسات مثل (Manjunath & Banerjee (2022 و Ukofia & Essen (2021).

أثبتت نظرية الاستجابة للمفردة (IRT) فعاليتها في تقديم تقديرات دقيقة للمفردات ذات التمييز المنخفض (الضعيف)؛ إذ تمكنت هذه النظرية من تحديد خمس مفردات منخفضة (ضعيفة) التمييز، مقارنة بالنظرية الكلاسيكية (CCT) التي كشفت عن وجود مفردتين فقط بتمييز ضعيف (الصيخان والمومني، 2021). يعكس هذا الفرق قدرة نظرية IRT على وصف خصائص المفردات بشكل أكثر تفصيلًا ودقة.

يُعد وجود مفردات ذات تمييز منخفض تحديدًا، حيث يمكن أن تؤثر سلبًا على دقة الاختبار وأداء الطلبة. فقد تعكس هذه المفردات بُعدًا غير مرتبط بالسمعة المستهدفة، أو تكون صيغتها غامضة، أو تعاني من مشكلات في خيارات الاستجابة فوجود مثل هذه المفردات تؤدي إلى إهدار وقت الاختبار، كما قد تُربك الطلبة وتؤثر سلبًا على أدائهم في باقي الفقرات، مما ينتج عنه تقديرات غير دقيقة لأدائهم، لذا ينصح بمراجعة هذه المفردات من قبل معدي الإختبارات وتجنب استخدامها لاحقًا أو إعادة صياغتها (DeMars, 2010).

بشكل عام تشير النتائج إلى أن نظرية الاستجابة للمفردة (IRT) تقدم تقديرات أكثر دقة لمعاملات التمييز مقارنة بنظرية القياس الكلاسيكية (CTT)، خصوصًا في المفردات ذات التمييز المنخفض. ومع ذلك يمكن الاعتماد على كلتا النظريتين في تحليل الصعوبة والتمييز، رغم وجود بعض الفروق في النتائج.

التوصيات

بناءً على نتائج الدراسة وأهدافها، توصي الدراسة بالآتي:

- 1- إعادة النظر في تصميم إختبار مادة الدراسات الاجتماعية ليشمل توزيعاً متناسباً للمفردات عبر مستويات الصعوبة المختلفة (سهلة، متوسطة، صعبة)، وينصح بتقليل نسبة المفردات السهلة وزيادة المفردات المتوسطة والصعبة لضمان تحقيق تنوع أكبر يمكّن من تقييم قدرات الطلبة المختلفة بشكل دقيق وشامل.
- 2- مراجعة المفردات ذات التمييز المنخفض، وتجنب تضمينها في إمتحانات مادة الدراسات الاجتماعية المستقبلية؛ لضمان دقة وفعالية الإختبار في قياس مستويات الطلبة. وبناءً على نتائج التحليل توصي الدراسة باستبعاد المفردتين 2 و 13 وفقاً للنظرية الكلاسيكية، بالإضافة إلى المفردات (10، 13، 25، 28، 30، 38) وفقاً لنظرية الإستجابة للمفردة، يلاحظ اتفاق النظريتين على استبعاد المفردة 13 لسهولةها العالية، مما يعزز ضرورة حذفها.
- 3- استخدام نظرية الإستجابة للمفردة لتحليل إمتحانات دبلوم التعليم العام، نظراً لدقتها العالية في تقدير المفردات بشكل عام، وبخاصة المفردات ذات التمييز المنخفض.
- 4- الاستفادة من نتائج الدراسة لتحديد نقاط الضعف والقوة لدى الطلبة، وتعميمها على المعلمين، والمشرفين التربويين، مما يعزز تطوير خططهم التعليمية والإشرافية.
- 5- تدريب المعلمين، والمشرفين التربويين المعنيين بوضع الإختبارات الفترية، بالإستناد إلى نتائج التحليل، لتعزيز قدراتهم في تصميم الإختبارات.
- 6- الإستمرار في إجراء دراسات مقارنة بين النظريتين الكلاسيكية ونظرية الإستجابة للمفردة، بهدف تحسين أدوات القياس وضمان جودة التقييم.

المقترحات

بناءً على نتائج الدراسة وأهدافها، تقترح الدراسة بالآتي:

- 1- إنشاء بنك أسئلة لمادة الدراسات الاجتماعية (هذا وطني)؛ بحيث يتم الاستفادة من تحليل الإختبارات ومؤشرات الصعوبة والتمييز للمفردات، في بناء وإعداد

الإختبارات النهائية، كما يمكن أن يستفيد منه المعلمين والمشرفين التربويين عند إعداد الإختبارات القصيرة.

2- **تنفيذ ورش عمل لمعلمي الدراسات الاجتماعية** لتعريفهم بنتائج الدراسة وتدريبهم على تصميم اختبارات أكثر دقة وثباتاً. يمكن لورش العمل أن تركز على كيفية بناء وصياغة المفردات الإختبارية، ونشر الوعي في الميدان التربوي بأهمية الإستفادة من نتائج تحليل استجابات الطلبة سواء بالنظرية الكلاسيكية أو نظرية الإستجابة للمفردة.

3- **إعداد دليل إرشادي لمعلمي الدراسات الاجتماعية** حول كيفية إعداد اختبارات دقيقة وموثوقة، استناداً إلى نتائج التحليل ونظريات القياس الحديثة، ما يسهم في توحيد معايير بناء الاختبارات بين المعلمين.

المراجع العربية

الدوسري، سعيد (2019). مقارنة بين النظرية الكلاسيكية للقياس ونظرية الاستجابة للمفردة في تقدير الخصائص السيكمترية لاختبار القدرات العامة مجلة جامعة تبوك للعلوم الإنسانية والاجتماعية، 6(1)، 47-64.

الزغول، عماد (2001). مبادئ علم النفس التربوي، دار الكتاب الجامعي.

الصيخان، رحاب بنت صالح؛ والمومني، رنا ثاني (2021). مقارنة بين النظرية الكلاسيكية والنموذج ثلاثي المعلم في انتقاء فقرات اختبار تحصيلي للغة الإنجليزية. المجلة الدولية للدراسات التربوية والنفسية، 10(1) 135-156.

علام، صلاح الدين محمود (2007). القياس والتقويم التربوي في العملية التدريسية. ط1. عمان، الأردن: دار المسيرة للنشر والتوزيع.

علام، صلاح الدين محمود (2002). القياس والتقويم التربوي والنفسي أساسياته وتطبيقاته وتوجهاته المعاصرة، دار الفكر العربي.

علام، صلاح الدين محمود (2005). نماذج الإستجابة للمفردة الإختبارية أحادية البعد

ومتعددة الأبعاد وتطبيقاتها في القياس النفسي والتربوي، دار الفكر العربي.

ملحم، سامي محمد (2005). مناهج البحث في التربية وعلم النفس، دار المسيرة.

المراجع الأجنبية

Ayla, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse of Assessment and Evaluation.

- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.).
- Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Ibrahim, A. B. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(5C), 1260–1266.
- Bichi, A. A. (2015). Item analysis using a derived science achievement test data. *International Journal of Science and Research (IJSR)*, 4(5), 1655-1662.
- Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Australian Journal of Basic and Applied Sciences*, 9(7), 549-556.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs: Prentice-Hall.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park: The International Professional Publisher.
- Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57(1), 98–107.
- Manjunath, M., & Banerjee, S. (2022). Assessing the equivalence between classical test theory and item response theory. *Assessment resources: Azim Premji University*.
- Min, S., Bishop, K., & Cook, G. H. (2022). Reading is a multidimensional construct at child-L2-English-literacy onset, but comprises fewer dimensions over time: Evidence from multidimensional IRT analysis. *Language Testing*, 39(2), 265–288.
- Rust, J., Kosinski, M., & Stillwell, D. (2021). *Modern psychometrics: The science of psychological assessment* (4th ed.). Routledge.
- Subali, B., Kumaidi, & Aminah, N. S. (2021). The comparison of item test characteristics viewed from classic and modern test theory. *International Journal of Instruction*, 14(1), 647-660.
- Ukofia, I. F., & Essen, C. B. (2021). A comparative analysis of item difficulty and discrimination parameters estimation in classical test and item response theories. *The Colloquium*, 9(1), 1-8.

ملحق(1) منحنيات ملائمة المفردات للنماذج المستخدمة

يكون النموذج متوافقاً مع البيانات إذا كانت النسبة الملاحظة للإجابات الصحيحة لكل مجموعة من الطلبة والمثلة بالمنحنى الأحمر قريبة من القيم المتوقعة وفقاً للنموذج والمثلة بالمنحنى الأسود، حيث تشير الانحرافات الكبيرة للمنحنين إلى عدم ملائمة المفردة.

