المجلة الأردنية في العلوم التربوية، مجلد 17، عدد 2، 2021، 285-299

دقّة قياس الاختبارات التّكيفيّة متعدّدة المراحل في ظل ظروف اختبارية مختلفة

لؤي شواشرة و محمود القرعان*

Doi: //10.47015/17.2.8

تاريخ تسلم البحث: 2020/3/30 تاريخ قبوله: 2020/6/4

ملخص: هدفت الدّراسة الحالية إلى مقارنة دقّة قياس تصاميم مختلفة للاختبارات التّكيفيّة متعدّدة المراحل تحت مجموعة من الشّروط، وهي ثلاثة مستويات لطول الاختبار (12، 24، 36) فقرة، ومستويان لتصميم اللوحة (3-3-1، 2-2-1)، ومستويان لطول وحدة التوجيه (طويل، قصير)، ومستويان لاستراتيجية التّوجيه (DPI، AMI). ولتحقيق هدف الدراسة، تم توليد بيانات القدرة الحقيقية لعينة تتكون من (5000) مفحوص من توزيع طبيعي بمتوسط حسابي (1) وانحراف معياري (0)، وتمّ تقييم نتائج الدّراسة باستخدام متوسّط التّحيّز ووسط مربعات الخطأ للقدرة الحقيقيّة والقدرة المقدّرة. وقد أظهرت نتائج الدراسة أنّ دقّة القياس تزداد بزيادة طول الاختبار، بغض النظر عن طول وحدة التّوجيه وتصميم اللوحة واستراتيجية التّوجيه المستخدمة. كما أظهرت النتائج تأثيرات من صغيرة إلى ضئيلة لمعظم متغيرات التصاميم (تصميم اللوحة، استراتيجية التّوجيه، طول وحدة التّوجيه).

(**الكلمات المفتاحية**: الاختبارات التّكيفيّة متعدّدة المراحل، دقّة القياس، تصميم اللوحة، استراتيجية التّوجيه، طول وحدة التّوجيه)

مقدمة: يتزايد الاهتمام ببناء الاختبارات والمقاييس النفسية والتربوية وتطويرها لتقدير قدرة الأفراد وتقييمها لمساعدة أصحاب القرار على التّنظيم والتّطوير والتّحسين. وبالتالي يجب أن تتمتّع هذه الاختبارات والمقاييس بالصّدق والثّبات والموضوعيّة، بحيث ينعكس ذلك على دقة القرار المتّخذ.

ولعقود طويلة، كانت اختبارات الورقة والقلم أكثر الطرق شيوعًا لقياس مهارات وقدرات الأفراد المتقدّمين للاختبار، وتمّ التّحول بشكل تدريجي إلى الاختبارات الخطيّة الحاسوبيّة مع التقدّم التّكنولوجي وتطور البرامج الحاسوبية. وتعتبر هذه الاختبارات مشابهة لاختبارات الورقة والقلم، إلا أنها توفّر الوقت والجهد من حيث إدارة الاختبار، وتجميعه، وإعطاء الدرجات، بحيث يمكن الحصول على تقرير الاختبار بصورة فورية (Yan, Von Davier & Lewis, 2014).

ويتقدّم جميع المفحوصين في اختبارات الورقة والقلم والاختبار الخطي المحوسب للفقرات نفسها بغض النظر عن مستوى قدرتهم، وبغض النظر عما إذا كانت الفقرات سهلة أو صعبة بالنسبة لهم، حيث الفقرات السّهلة جدًا والصّعبة جدًا لا تقدم معلومات عن مستوى أداء المفحوص. وفي ضوء التفاوت بين قدرة المفحوص وصعوبة الفقرة، سيؤدي ذلك إلى انخفاض في كمية المعلومات وخصوصًا عند طرفي متصل القدرة، وبالتالي إلى انخفاض في دقّة القياس (Zheng & Chang, 2015).

Assessing Measurement Accuracy of Multistage Adaptive Testing under Different Testing Modes

Loiy Shawashreh and Mahmoud Alquraan, *Yarmouk University, Jordan.*

Abstract: This study aims at comparing the measurement accuracy of multistage adaptive testing under different conditions: test length (12, 24 and 36 items), panel design (1-3-3 and 1-2-2 panel designs), routing module length (long and short) and routing strategy (AMI and DPI). To achieve the aims of this study, (5000) true abilities were generated with a normal distribution of a mean of (1) and a standard deviation of (0). Measurement accuracy was assessed using mean of bias and mean squared error (MSE). The results show that the measurement accuracy increases as the test length increases, despite of panel design (1-3-3 and 1-2-2 panel designs), routing module length (long and short) or routing strategy. Moreover, the results show that panel design, routing strategy and routing module length marginally affect the measurement accuracy of multistage adaptive testing.

(**Keywords**: Multistage Adaptive Ttesting, Measurement Accuracy, Panel Design, Routing Module Length, Routing Strategy)

وقد زاد الاهتمام بالاختبارات التّكيفيّة (Computerized Adaptive Testing: CAT) والاختبارات التّكيفيّة متعدّدة المراحل (Multistage Adaptive Testing: MST) مقابل اختبار الورقة والقلم والاختبار الخطي المحوسب؛ نتيجةً للتّطور التكنولوجي، واستخدام الحاسوب على نطاق واسع، وظهور نظريّة استجابة الفقرة وتطور البرمجيّات الإحصائية المتعلّقة بها، حيث توفر الاختبارات التّكيفيّة فقرات مصمّمة لتناسب قدرة المفحوص، وبالتالي توفر هذه الاختبارات تقديرات أكثر دقة وكفاءة، ويكون طول الاختبار التكيفي وزمنه أقل مما هي في اختبار الورقة والقلم (Wainer, 2000).

* جامعة اليرموك، الأردن.

كذلك بدأ الاهتمام بالاختبارات التّكيفيّة متعدّدة المراحل (Multistage Adaptive Testing: MST) في خمسينيات القرن الماضي. وينظر إلى الاختبارات التكيفية متعددة المراحل على أنها مزيج بين الاختبارات التّكيفيّة والاختبارات الخطيّة؛ لذلك تعتبر الاختبارات التّكيفيّة متعدّدة المراحل حلاً وسطًا بين الاختبارات الخطيّة والاختبارات التّكيفيّة من حيث المرونة والتّعقيد، حيث يمكن الحصول على اختبار متعدّد المراحل أقصر من الاختبار الخطي، ويمتلك دقّة قياس عالية مشابهة لدقّة الاختبارات التّكيفيّة في تقدير القدرة أو تصنيف المفحوصين (Zheng et al., 2012).

وتقدم لكل مفحوص في الاختبار التّكيفي المحوسب فقرات تتناسب مع مستوى قدرته. وبعد كل فقرة تقدم للمفحوص، يتم تقدير قدرته، وتقدم له فقرة جديدة تتناسب مع القدرة المقدّرة الجديدة؛ إي أنّ تقدير القدرة يتمّ بصورة مستمرّة للوصول إلى تقدير دقيق للقدرة (Magis, Yan & Von Davier, 2017).

يتشابه تصميم MST مع تصميم CAT بحيث تكون الفقرات مناسبة لمستوى قدرة المفحوص، ويتكون اختبار MST من مجموعة اختبارات قصيرة مجمعة مسبقًا تسمى الوحدات (Modules)، وتدار على مراحل وبحدّ أدنى مقداره مرحلتان، ويتم توجيه المفحوصين إلى الوحدات اللاحقة بناءً على قدرتهم المقدّرة في الخطوة السابقة، وتغطي هذه الوحدات مستويات مختلفة من الصعوبة (Yan et al., 2014).

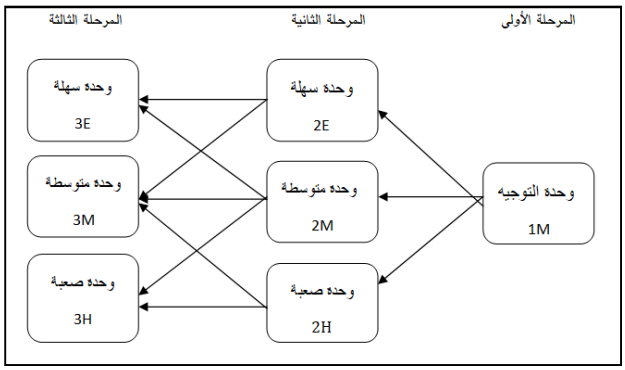
وتتشابه المكونات الأساسية للاختبارات التّكيفيّة متعدّدة المراحل والاختبارات التّكيفيّة من حيث تجمع الفقرات، واستراتيجية التوجيه التي تعين الوحدة التالية للمفحوص، وطرق تسجيل نتيجة المفحوص، بينما تمتلك MST عوامل فريدة، مثل الوحدات والألواح والمراحل. تمثل الوحدة مجموعة من الفقرات يتم إنشاؤها مسبقًا من قبل مطوّر الاختبار، بحيث تراعي الخصائص الإحصائية للاختبار (مثل كمية المعلومات، والخصائص السيكومترية للفقرات)، والخصائص غير الإحصائية (مثل مواصفات المحتوى، ومعدل التعرض). ويتم دمج الوحدات لإنشاء اللوحة (Panel) التي تعتبر بمثابة اختبار يقدّم للمفحوصين (Luecht 2000; Wang, 2017).

ويبدأ تنفيذ اختبارMST بتقديم مجموعة من الفقرات تسمى وحدة التوجيه (Routing module) لجميع المفحوصين، ويتم بعد ذلك تقدير قدرة كل مفحوص بناءً على نمط استجابته على هذه الفقرات ومقارنتها بمحك أداء محدد مسبقًا، وتقدم للمفحوص بعد ذلك -إذا كانت قدرته المقدرة أعلى من المحك- وحدة أكثر صعوبة (Difficult module)، بينما تقدم له وحدة أسهل (Easy module) إذا كانت قدرته المقدرة أقل من المحك. ويتم تحقيق الجزء التكيفي للاختبار عن طريق اختيار وحدة مناسبة للمفحوص في كلّ مرحلة وفقًا لأدائه في المراحل السابقة. وتقلل هذه العملية من طول الإختبار دون فقدان الكثير من المعلومات؛ لأن الفقرات تستهدف مستويات قدرة محدّدة بدقّة أكثر (Magis et al., 2017).

ويوضح الشكل (1) مثالاً للوحة ذات تصميم 3-3-1، تتكون من ثلاث مراحل وسبع وحدات. تتكوّن المرحلة الأولى من وحدة واحدة متوسطة الصعوبة (وحدة التوجيه)، بينما تتكوّن كل من المرحلتين الثّانية والثّالثة من ثلاث وحدات، مع وجود سبعة مسارات محتملة لكل مفحوص.

الشكل 1

لوحة ذات تصميم 3-3-1 لاختبار تكيفي متعدّد المراحل.



ويبدأ تنفيذ الاختبار باختيار لوحة عشوائيًا لكلّ مفحوص، بحيث يتقدّم جميع المفحوصين الذين تمّ تعيينهم للوحة في المرحلة الأولى لفقرات وحدة التوجيه، ويتم بعد ذلك تقدير قدرة كل مفحوص بناءً على نمط استجابته على تلك الفقرات، ومقارنتها بمحكّ محدّد مسبقًا لتوجيه المفحوصين إلى المرحلة الثّانية. وتقدم للمفحوص في المرحلة الثانية وحدة جديدة من بين ثلاث وحدات تتناسب مع مستوى قدرته المقدّرة في المرحلة الأولى، ويتمّ تكرار الأمر في المرحلة الثّالثة والأخيرة للوصول إلى التّقدير النّهائي للقدرة (Yan et al., 2014).

وتستخدم اختبارات MST التجميع الآلي للاختبار، مع التّركيز على الخصائص الإحصائية للاختبار، والخصائص غير الإحصائية (مثل: توازن المحتوى، والتحكّم بمعدّل عرض الفقرات). وتستخدم عادةً دالّة المعلومات شائعة الاستخدام في اختبار الورقة والقلم؛ لأنها تسهل بناء صور متوازية للوحدات وتتيح استخدامًا فعالاً لتجمع الفقرات (Armstrong & Roussos, 2003; Melican et al., 2009).

وثمّة طرق تستخدم لبناء الألواح المتوازية في اختبارات MST، منها: طرق البرمجة الخطية (Linear programming methods)، والطرق التوجيهية (Heuristic methods)، وتوفر كلتا الاستراتجيتين حلولاً لبناء الالواح المتوازية، بحيث تلبي جميع قيود الاختبار. فطرق البرمجة الخطية توفّر أفضل الحلول، بحيث يمكنها تلبية جميع قيود الاختبار، بينما توفّر الطرق التّوجيهية

حلولاً قريبة من الحل الأمثل بتكلفة أقل، وإضافة إلى العديد من المزايا العمليّة ومنها سهولة الاستخدام (Zheng et al., 2012).

ونحتاج إلى بناء وحدات بالاعتماد على تجمّع الفقرات، وتجميع اللوحات من الوحدات النّاتجة لبناء لوحات متوازية. وتوجد استراتيجيّتان لتحقيق التوازي بين الألواح: استراتيجية من أسفل إلى أعلى (Bottom-up) التي يتمّ فيها تجميع النّماذج المتوازية لكل وحدة، وبعد ذلك يتمّ خلطها ومطابقتها لبناء عدد كبير من الألواح المتوازية، واستراتيجية من أعلى إلى أسفل (Top- down) التي تكون فيها النماذج المجمعة لكل وحدة ليست متوازية تمامًا، ولبناء ألواح متوازية، هناك حاجة إلى إجراء تحسين إضافي تلبّي فيه المسارات المختلفة قيودًا محدّدة. والجدير بالذكر أن التجميع من أعلى إلى أسفل أكثر تعقيدًا من التّجميع من أسفل إلى أعلى، نظرًا للجولة الإضافية من التّحسين في الخطوة الثانية (Zheng et al., 2012; Luecht & Nungester, 1998).

وتستخدم العديد من الأساليب للتجميع الآلي للاختبار منها: نموذج الانحراف المطلق موزون المعيار (Normalized Weighted Absolute Deviation: NWAD Luecht, 2000). نموذج الانحراف الموزون (Weighted Deviation (Model: WDM). (Swanson & Stocking, 1993)، وطرق البرمجة الخطية (Linear Programming)، وقد تمّ استخدام طرق البرمجة الخطيّة لتجميع الوحدات في هذه الدّراسة (Diao & van der Linden, 2011).

وتستخدم العديد من استراتيجيات التّوجيه لتحديد نقطة التوجيه (علامة القطع)، التي تحدّد الوحدة التالية التي تقدّم للمفحوص، بحيث تتناسب مع قدرته المقدرة في المرحلة السابقة. ومن هذه الاستراتيجيات: استراتيجية الحد الأقصى للمعلومات (Approximate Maximum Information: AMI): التي تعمل على تحديد نقط التّقاطع بين دالّة معلومات الوحدة التي تمّ تقديمها للمفحوص في الخطوة السابقة ودالّة المعلومات التي ستقدّم للمفحوص في المرحلة اللاحقة. ومنها أيضًا طريقة فترات المجتمع المحدّدة (Defined Population Interval: DPI) التي تعمل على تحديد النسب المئويّة للمفحوصين في المجتمع الذين من المتوقع اتباعهم لمسار معين عبر المسارات الرئيسة داخل اللوحة (Luecht et al., 2006).

وتعتبر طرق تقدير القدرة خطوة أساسية في إجراءات MST، حيث يتم تقدير قدرة المفحوص بعد إجابته عن فقرات تقدير الأرجحية العظمى (Maximum Likelihood Estimation: MLE) في كل وحدة تقدم له. والطرق التي تستخدم في تقدير القدرة هي طرق التقدير البيزية (Bayesian Estimation). ومن الطرق البيزية طريقة التقدير البعدي المتوقع (Expected A Posterior: EAP) التي تعتمد على معلومات سابقة عن توزيع القدرة للمفحوصين. وهذا التّوزيع يتم افتراضه بناءً على معلومات سابقة عن المفحوصين. وعادة يتم افتراض التوزيع الطبيعي. وبناءً على نتائج دراسة كيم وآخرين (Kim et al., 2015) التي أظهرت تفوق طريقة EAP على طريقة الأرجحية العظمى في تقدير القدرة لاختبارات MST، فقد تمّ استخدام طريقة EAP في الدراسة الحالية.

وأجرى جودون (Jodoin, 2003) دراسة هدفت إلى مقارنة دقة القياس، ودقة التصنيف، والخطأ المعياري في القياس بين الاختبار الخطّي ذي الطّول الثابت واختبار MST واختبار CAT، تحت ثلاثة مستويات من جودة تجمّع الفقرات، ومستويين من التّطابق بين مواصفات محتوى تجمع الفقرات، ومستويين لطول الاختبار، وعدة مستويات للتحكم في معدل عرض الفقرات لعدد من برامج الاختبار. وتكونت عينة الدراسة من توليد بيانات (2000) مفحوص. وأظهرت النّتائج زيادة في دقّة القياس مع زيادة جودة تجمّع الفقرات و زيادة طول الاختبار.

وأجرى جودون وآخرون (Jodoin et al., 2006) دراسة هدفت إلى مقارنة تصاميم اختبارات ذات طول ثابت (اختبار خطّي، واختبار تكيّفي ثلاثي المراحل، واختبار تكيّفي ثنائي المراحل). وتمّت عمليّة المقارنة على أربعة اختبارات خطيّة حقيقية مكونة من (60) فقرة واختبار تكيفي ثلاثي المراحل مكون من (60) فقرة وكل وحدة مكونة من (20) فقرة، واختبار تكيفي ثنائي المراحل مكون من (40) فقرة وكل وحدة مكونة من (20) فقرة. تكوّنت عينّة الدراسة من بيانات (5000) مفحوص تمّ توليدها من توزيع طبيعي N(0,1). وأظهرت النّتائج أنّ جميع تصاميم الاختبارات المكوّنة من (60) فقرة أعطت تقديرات دقيقة للقدرة. كما أظهرت النّتائج أنّ جميع تصاميم الاختبارات المكونة من (40) فقرة كانت أقل دقّة في تقدير القدرة من الاختبارات بطول (60) فقرة، ولكنّها أعطت تقديرات ضمن النطاق المعقول والمقبول.

وأجرى ساري وهغنز-مانلي (Sari & Huggins-Manley, 2017) دراسة هدفت إلى استكشاف دقة قياس CAT وMST عندما تتباين مجالات المحتوى عبر مجموعة متنوعة من أطوال الاختبار. وتمت مقارنة تصميم واحد للاختبارات التكيفية مع تصميمات الاختبارات التكيفية متعددة المراحل (3-3-1, 1-3-1) عبر العديد من الظروف الاختبارية، بما في ذلك طول الاختبار (24، 48) فقرة، وعدد مناطق المحتوى. تكونت عينة الدراسة من (4000) مفحوص تمّ توليد بيانات القدرة الحقيقيّة لهم من توزيع طبيعي N(0, 1). وأظهرت نتائج الدّراسة أنّ لطول الاختبار تأثيرًا في النتائج أكثر من عدد مناطق المحتوى.

وقام كيم وآخرون (Kim et al., 2013) بإجراء دراسة هدفت إلى المقارنة بين أربعة تصاميم للألواح، وثلاث طرق للتوجيه، ومستويين من طول الاختبار، وثلاث نسب للنجاح، استنادًا إلى نموذج التّصحيح الجزئي لماستر. وتمّ توليد بيانات لعيّنة تتكون من (1000) مفحوص، وأظهرت النّتائج أنه بغضّ النّظر عن طريقة التّوجيه المستخدمة، فإن تصميم اللوحة يؤدي إلى دقّة قرار التّصنيف نفسها مع شرط طول الاختبار نفسه، كما أظهرت النتائج أنه بزيادة طول الاختبار، تزداد دقّة القياس.

وأجرت وانغ (Wang, 2017) دراسة هدفت إلى المقارنة العادلة بين أداء الاختبارات التّكيفيّة والاختبارات التّكيفيّة متعددة المراحل باستخدام تجمع رئيسي للفقرات مكون من (8100) فقرة تحت (16) شرطًا، وهي مستويان من طول الاختبار (40،60) فقرة، ومستويان من تصميم اللوحة (3-3-2،1-2-1)، ومستويان من استراتيجية التوجيه (AMI, DPI)، ومستويان من أولويّة التّجميع (التجميع الأمامي، التجميع الخلفي)، ومقارنة كلّ اختبار ناتج من تفاعل الشّروط السّابقة بالاختبار التّكيّفي المقابل. تكوّنت عيّنة الدّراسة من (5000) مفحوص، تمّ توليد بيانات القدرة الحقيقيّة لهم من توزيع طبيعي N(0,1). وأظهرت نتائج الدّراسة ان دقة قياس MST تزداد بزيادة طول الاختبار. كما أظهرت النتائج تأثيرات ضئيلة لتصميم اللوحة وطريقة التوجيه، وأوصت بمقارنة دقة قياس تصاميم MST باستخدام اختبارات قصيرة.

وأجرى زنزكي (Zenisky, 2004) دراسة هدفت إلى البحث في الآثار المترتبة على تفاعل العديد من متغيرات تصميم MST. وتمت دراسة المتغيرات الآتية: تصميم اللوحة وله أربعة مستويات (1-3-2-3;1-2-3-3;1-3-3-2;1-2-2-2)، واستراتيجية التّوجيه بين المراحل ولها أربعة مستويات، والنسب المئويّة للتوجيه ولها ثلاثة مستويات. وتكونت عينة الدّراسة من بيانات (9000) مفحوص تمّ توليدها من توزيع طبيعي N(0,1). وأظهرت نتائج الدّراسة تأثيرات من صغيرة إلى ضئيلة لمعظم متغيّرات التصاميم (تصميم اللوحة، استراتيجية التوجيه).

وأجرى زنزكي وهامبلتون (Zenisky & Hambleton, 2004) دراسة هدفت إلى استكشاف طريقة تفاعل متغيّرات تصاميم الاختبارات التّكيفيّة متعدّدة المراحل مع بعضها البعض، وكيفيّة تأثيرها على دقّة التّصنيف وتقدير القدرة. وكانت المتغيّرات للتصاميم التي تمت دراستها: كميّة معلومات الاختبار الإجماليّة، ولها أربعة مستويات، وكذلك أربعة مستويات لتصميم اللوحة (1-3-2-3;1-2-3-3;1-3-3-2;2-2-1-3)، وتوزيع المعلومات على المراحل وله مستويان، واستراتيجية التوجيه ولها أربعة مستويات. وتكونت عينة الدراسة من بيانات (9000) مفحوص تمّ توليدها من توزيع طبيعي N(0,1). وأظهرت نتائج الدّراسة تأثيرات من صغيرة إلى ضئيلة لمعظم متغيرات التصاميم (تصميم اللوحة، استراتيجية التوجيه). كما أظهرت نتائج الدّراسة أن هناك تأثيرًا ضئيلاً على تقدير القدرة عندما تزيد كمية المعلومات على (10) بشكل كبير.

وقد سعت دراسة باتسولا (Patsula, 1999) إلى مقارنة دقة تقدير القدرة بين الاختبارات التكيفية والاختبارات التكيفية متعددة المراحل من خلال عدة تصميمات هي: عدد المراحل، وعدد الوحدات في المرحلة الواحدة، وعدد الفقرات المتضمنة في الوحدة الواحدة مع الاحتفاظ بطول ثابت للاختبار، وطريقة تحديد علامة القطع بين المراحل. وقد تكونت عينة الدراسة من (500) مفحوص. وأظهرت نتائج الدراسة أنّ زيادة عدد المراحل من مرحلتين إلى ثلاث مراحل بشكل عام تقلل من حجم الخطأ في تقدير القدرة، كما أنّ زيادة عدد الوحدات من ثلاث إلى خمس أدت إلى زيادة في دقة تقدير القدرة، وكان لتغير عدد الفقرات في كل مرحلة تأثير ضئيل على دقة تقدير القدرة.

كما أجرى كيم وآخرون (Kim et al., 2015) دراسة هدفت إلى التّحقّق من فاعليّة طرق تقدير القدرة في نظريّة استجابة الفقرة من حيث الخطأ وتحيز التقدير في الاختبارات التكيفية متعددة المراحل. واعتمدت هذه الدّراسة تصميم اللوحة المكون من مرحلتين (3-1)، ومستويين لصعوبة الفقرات في المرحلة الثّانية (متداخلة؛ منفصلة)، وأربعة مستويات لطول وحدة التّوجيه (10-15 ,25-15 ,20-20 ,15-25 ,30)، وسبع طرق لتقدير القدرة؛ لمعرفة ما إذا كان هناك تفاعل بين طرق تقدير القدرة وتصميمات MST. تكونت عيّنة الدراسة من بيانات القدرة الحقيقية لـ (8200) مفحوص، تمّ توليدها من توزيع منتظم، بحيث تنحصر قيم القدرة بين (-3) و (3). وأظهرت نتائج الدّراسة أنّ هناك تأثيرًا ضئيلاً لطول وحدة التّوجيه على دقة التّقدير للتصميمات المستخدمة في الدراسة، كما أظهرت النتائج تفوق الطرق البيزية على طرق الأرجحية.

وأجرى أوزتورك (Öztürk, 2019) دراسة هدفت إلى فحص تأثير طول وحدة التّوجيه وخصائصها المختلفة على دقة القياس. واستخدمت ستة أطوال مختلفة لوحدة التّوجيه (5 ,10 ,15 ,20 ,25 ,30) فقرة مع طول متغير للاختبار (25 ,30,35 ,40 ,45 ,50) فقرة، وتسع خصائص سيكومترية لفقرات وحدة التّوجيه، وتصميمان مختلفان للوحة. وتكونت عيّنة الدراسة من محاكاة القدرة الحقيقية لـ (5000) مفحوص، وتمّ توليدها من توزيع طبيعي. وأظهرت نتائج الدراسة أنه كلمّا زاد طول وحدة التّوجيه تنخفض قيمة الجذر التّربيعي لمربعات متوسّط الخطأ، وتزداد قيمة معامل الارتباط. وأظهرت النّتائج أنّ تصميم اللوحة ثلاثية المراحل يعطي دقّة قياس أعلى من التّصميم الذي يتكوّن من مرحلتين بشكل عام.

وبعد مراجعة الدراسات السابقة، تتضح قلة الدراسات التي تناولت تصاميم الاختبارات متعدّدة المراحل باستخدام اختبارات قصيرة، وأن أغلب الدراسات تناولتها باستخدام الاختبارات الطويلة والمتوسطة. لذلك جاءت الدّراسة الحالية لتقييم دقّة قياس الاختبارات التّكيفيّة متعدّدة المراحل باستخدام اختبارات قصيرة ودراسة كيفيّة تفاعل طول الاختبار مع طول وحدة التوجيه وتصميم اللوحة واستراتيجية التّوجيه. هذا مع العلم بأنه لا توجد دراسة عربية تناولت فاعلية الاختبارات التّكيفيّة متعددة المراحل بحسب علم الباحثيْن.

مشكلة الدّراسة وسؤالها

يهدف التقييم واسع النطاق (Large-scale Assessment) في التعليم أو المجالات الأخرى إلى توفير معلومات أساسية، مثل متوسط الكفاءة ودرجة الإتقان، باستخدام تقنيات تقييم تقلل من التكلفة ومقدار الوقت الذي يمضيه المفحوص في التقييم، بحيث

تساعد هذه المعلومات أصحاب القرار على تحسين النظم والسياسات (Al-Gamdi, 2018). لذلك يسعى مستخدمو الاختبارات التي تستخدم على نطاق واسع للحصول على دقّة قياس عالية باستخدام أقل عدد من الفقرات؛ إذ إنّ تطوير فقرة واحدة يكلف ما بين (1500) و(2500) دولار (Rudner, 2009)، كما أن إدارة الاختبارات الطويلة وتنفيذها بحاجةٍ إلى أوقات طويلة.

ويلجأ مطوّرو الاختبارات في بعض الأحيان إلى استخدام اختبارات قصيرة أو متوسطة مع الحصول على دقة قياس مقبولة؛ لتوفير الوقت والجهد والمال. وفي مجال الاختبارات التّكيفيّة متعدّدة المراحل، هناك بعض الاختبارات القصيرة التي تستخدم تصاميم MST (Wang, 2017). مثل اختبارات التّقييم الوطني للتقدم التعليمي (National Assessment of Educational Progress: NAEP).

ونتيجة لشيوع الاختبارات التّكيفيّة متعدّدة المراحل واستخدامها بشكل واسع في القياس التربوي والنّفسي والتّقييمات الواسعة، وندرة الدراسات التي اهتمت بقياس دقة قياس الاختبارات التّكيفيّة متعدّدة المراحل باستخدام اختبارات قصيرة، واستجابة لتوصيات دراسة وانغ (Wang, 2017) بضرورة إجراء المزيد من الدراسات حول دقّة قياس الاختبارات القصيرة التي تستخدم تصاميم MST وطريقة تفاعلها مع متغيّرات أخرى مثل تصميم اللوحة وطول وحدة التوجيه. فإنّ الدراسة الحالية سعت للمقارنة بين تصاميم الاختبارات التّكيفيّة متعدّدة المراحل من حيث دقّة القياس في ظلّ ظروف مختلفة من طول الاختبار(36،24،12) فقرة، وتصميم اللوحة (3-3-1،2-2-1)، واستخدام مستويين لطول وحدة التوجيه (طويل، قصير) واستراتيجيتين للتوجيه (AMI; DPI).

وبالتحديد، فإنّ هذه الدّراسة تحاول الإجابة عن السؤال الآتي: هل تختلف دقّة قياس تقدير القدرة للاختبارات التّكيفيّة متعدّدة المراحل في ظلّ شروط مختلفة من طول الاختبار ومستويين من تصميم الألواح، ومستويين من طول وحدة التوجيه، وإستراتيجيتين للتوجيه (AMI و DPI)؟

هدف الدراسة

هدفت الدّراسة الحالية إلى مقارنة دقّة قياس تصاميم مختلفة للاختبارات التّكيفيّة متعددة المراحل تحت مجموعة من الشروطّ وهي: ثلاثة مستويات لطول الاختبار (12، 24، 36) فقرة، ومستويان لتصميم اللوحة (3-3-1، 2-2-1)، ومستويان لطول وحدة التوجيه (طويل، قصير)، واستراتيجيتان للتوجيه (AMI, DPI).

أهميّة الدّراسة

تبرز أهمية الدراسة الحالية من خلال سعيها لسدّ الفجوة في الأدب التربوي، ومقارنة فاعليّة تصاميم مختلفة لـ MST باستخدام اختبارات قصيرة نسبيًا، والتّحقّق من مختلف تصميمات الاختبارات متعددة المراحل؛ (مثل: تصميم اللوحة، وطول وحدة التوجيه، واستراتيجية التّوجيه) وطريقة تفاعلها مع اختبارات قصيرة، وذلك على العكس من الدراسات السابقة التي سعت إلى عملية المقارنة باستخدام اختبارات طويلة.

محدّدات الدّراسة

اقتصرت الدراسة الحالية على طرق تجميع محدّدة للاختبارات التّكيفيّة متعددة المراحل، واستخدام الألواح المتوازية. كما اقتصرت على استخدام فقرات من نوع الاختيار من متعدّد (ثنائية الاستجابة) لإنشاء تجمّع الفقرات، واختبارات المحاكاة، وعلى استخدام بيانات مولدة لبناء تجمع الفقرات، وبيانات القدرة الحقيقية لعينة الدراسة.

الطريقة

هدفت الدراسة إلى مقارنة دقّة قياس تصاميم MST تحت ظروف اختبارية مختلفة، وتمّ استخدام بيئة R من خلال الحزم lpsolve و lpsolveAP لأغراض بناء الوحدات والألواح في MST، كما تمّ استخدام الحزمة mstR لأغراض التحليل وتسجيل النّتائج، علمًا بأنه تمت برمجة بعض المعادلات غير المتوفرة في الحزم السابقة.

توليد بيانات MST

تمّ توليد بيانات القدرة الحقيقية لـ (5000) مفحوص من توزيع طبيعي بمتوسط حسابي (1) وانحراف معياري (0) بواسطة حزمة (mstR) التي تعمل ضمن بيئة R. وتمّ العمل على مقارنة العديد من شروط تصاميم اختبار(MST) ، وهي: طول الاختبار وله ثلاثة مستويات (12 ;24 ;36)، وعدد الفقرات داخل وحدة التّوجيه وله مستويان (طويلة؛ قصيرة)، وتصميم اللوحة وله مستويان (3-3-1 ;2-2-1) واستخدمت استراتيجيّتان للتوجيه (AMI; DPI). وبالتالي تمّ تكوين (24) شرطًا منفصلاً لتصميم (MST) عبر خمس لوحات متوازية.

عملت الدراسة الحالية على المقارنة بين مستويين لطول وحدة التوجيه (قصير؛ طويل). فمثلاً عندما كان طول الاختبار(12) فقرة، تمّت دراسة مستويين لطول وحدة التوجيه (3 فقرات؛ 6 فقرات). والجدول (1) يوضح طول كل وحدة لجميع أطوال الاختبار التي تمّت دراستها.

الجدول 1

طول الوحدة لتصاميم MST.

تصميم اللوحة 2-2-1		تصميم اللوحة 3-3-1	
عدد الفقرات في الوحدة	طول الاختبار	عدد الفقرات في الوحدة	طول الاختبار
18-9-9	36	18-9-9	36
9-9-18	36	9-9-18	36
12-6-6	24	12-6-6	24
6-6-12	24	6-6-12	24
6-3-3	12	6-3-3	12
3-3-6	12	3-3-6	12

وتعتمد عمليّة تجميع الوحدات في MST على خصائص الفقرات المتوفرة في تجمع الفقرات. وللحصول على تجمّع فقرات يمكنه تحقيق متطلّبات بناء الوحدات والألواح، يمكن استخدام مجموعة من الاختبارات الثابتة المثلى التي تتوافق مع مستوى قدرة المفحوصين لبناء هذا التجمّع (Van der Linden & Veldkamp, 2006).

تمّ بناء تجمع الفقرات باستخدام مجموعة من الاختبارات الثابتة المثلى التي تتوافق مع مستوى قدرة المفحوصين. وتم توليد تجمع الفقرات باستخدام النموذج ثلاثي المعلمة بحيث تكون الفقرات موزعة على ثلاث مناطق محتوى بنفس العدد من الفقرات، وتمّ التّحكم بمعدل عرض الفقرات بحيث يكون بمعدل (0.2). وأشار لورد (Lord, 1980) إلى أن الصيغة الرياضيّة للنموذج اللوجيستي ثلاثي المعلمة تعطى بالعلاقة الآتية:

$$P_j(\theta) = c_j + (1 + c_j)\frac{1}{1+e^{-Da_j(\theta-b_j)}} \quad \ldots\ldots\ldots (1)$$

تمثل b_j مستوى صعوبة الفقرة، وa_j معلمة التمييز للفقرة، وc_j تمثل معلمة التخمين (الخط التقاربي الأدنى)، وهي تمثل احتمال إجابة المفحوص إجابة صحيحة في حال غياب القدرة، و$D=1.7$ ثابت التّدريج. وقد تم توليد معاملات التمييز ومعلمات التخمين من التّوزيعات $a \sim$ lognormal (0, 0.3)، $c \sim$ Uniform (0.1, 0.2) على التوالي لمحاكاة تجمع فقرات اختبار رياضيات. كما تم توليد بيانات القدرة الحقيقية للمفحوصين من توزيع طبيعي$N(0, 1)$، بينما تم توليد معلمة الصعوبة لتجمع الفقرات من توزيع طبيعي للحصول على أفضل تطابق بين مستوى قدرة المفحوصين ومعلمة الصعوبة (Leung et al., 2005). والجدول (2) يوضح توزيع معلمة الصعوبة لكلّ وحدة.

الجدول 2

توزيع معلمة الصعوبة لجميع الوحدات.

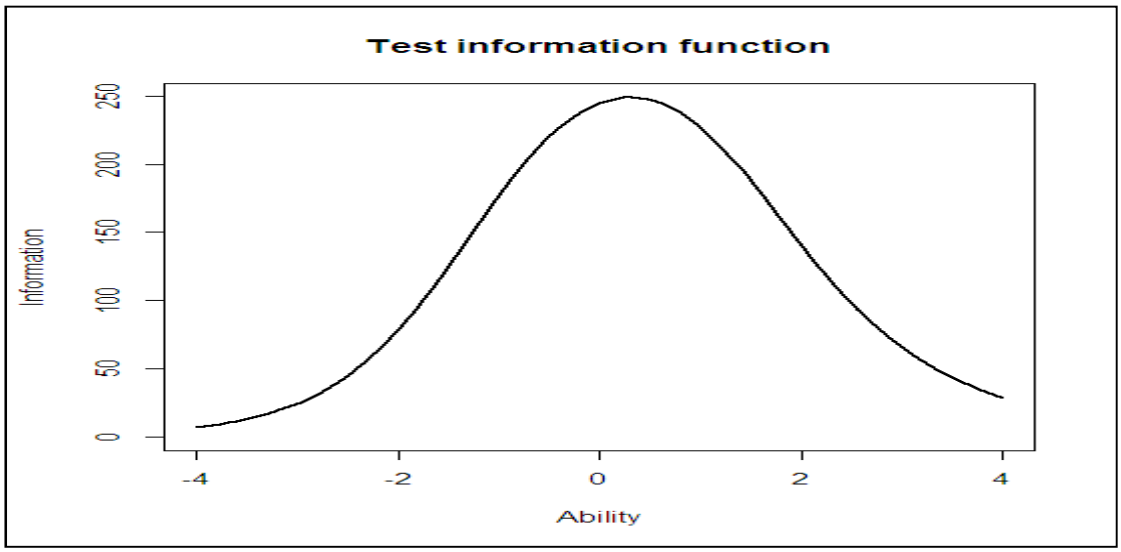
التصميم	1-3-3	1-2-2
المرحلة الأولى	N (0,1)	N (0,0.3)
المرحلة الثانية (2E)	N (-1, 0.6)	N (-0.7, 0.6)
المرحلة الثانية (2M)	N (0, 0.6)	----
المرحلة الثانية (2H)	N (1, 0.6)	N (0.7, 0.6)
المرحلة الثالثة (3E)	N (-1, 0.3)	N (-0.7, 0.3)
المرحلة الثالثة (3M)	N (0, 0.3)	----
المرحلة الثالثة (3H)	N (1, 0.3)	N (0.7, 0.3)

وتمّ توليد معلمة الصّعوبة لوحدة التّوجيه من توزيع طبيعي (N(0, 0.3 لتصنيف المفحوصين بدقّة إلى الوحدتين التاليتين في المرحلة الثّانية لتصميم 2-2-1، ونظرًا لأنّ توزيع القدرة يتبع التوزيع الطبيعي، فإن من المتوقّع توجيه المفحوصين بعدد متساوٍ إلى الوحدة الصّعبة والوحدة السّهلة في المرحلة الثّانية. وبسبب العدد المتساوي من المفحوصين لكل من الوحدتين السهلة والصعبة في المرحلة الثّانية تمّ تركيز معلمة الصعوبة للفقرات عند (0.7-) و(0.7) على التوالي، وتمّ تحديد القيم من خلال حساب وسيط مستوى القدرة لنصفين من المفحوصين، وتم تحديد متوسّط الصعوبة للوحدتين للمرحلة الثالثة بنفس الطريقة عند (0.7-) و(0.7). ولأنه قد يتم توجيه بعض المفحوصين إلى الوحدة غير الصحيحة في المرحلة اللاحقة، تمّ تخفيض قيمة التباين لمعلمة الصعوبة في المرحلة الثالثة ليكون هناك تداخل أعلى لتوزيع الصعوبة على مقياس القدرة (Wang, 2017).

وقد تم توليد معلمة الصعوبة لوحدة التّوجيه للتصميم 1-3-3 من توزيع طبيعي $N(0, 1)$ بحيث تتطابق تمامًا مع مستوى قدرة المفحوص، والاختلاف هنا عن توزيع التّصميم السابق أنّ المفحوصين تم توجيههم إلى ثلاث وحدات في المرحلة الثّانية. وتم توجيه المفحوصين في المرحلة الثّانية والمرحلة الثّالثة إلى ثلاث وحدات مختلفة بحيث تم تحديد متوسط صعوبة (1-)، (0)، (1) للوحدات الثلاث على التوالي. وتمّ بناء تجمع فقرات مكون من (1485) فقرة ثنائية الاستجابة (Wang et al., 2012). والشكل (2) يظهر دالّة معلومات تجمع الفقرات.

الشكل 2

دالة معلومات تجمع الفقرات.



تمّ استخدام استراتيجية من أسفل إلى أعلى (Bottom-up) لبناء الوحدات، حيث يمكن لهذه الاستراتيجية تجميع وبناء أكثر من وحدة في وقت واحد. وأشار لويشت (Luecht et al., 2006) إلى أن هذه الاستراتيجية تتطلب دوال معلومات وقيود محتوى بشكل منفصل لكل وحدة، ولذلك تمّ استخدام طريقة الحدّ الأقصى التقريبي للمعلومات (AMI) لتحديد دالّة المعلومات المستهدفة لكل وحدة بشكل منفصل عن الوحدات الأخرى (Luecht, 2000). وتم ذلك وفق الخطوات الآتية:

أ- لكل فقرة في تجمع الفقرات، تم حساب دالة معلومات الفقرة عند مستوى قدرة معين (مثلاً تم حساب دالة معلومات الفقرة عند مستوى قدرة مقداره صفر لوحدة التوجيه).

ب- تم ترتيب الفقرات ترتيبًا تنازليًا حسب كمية المعلومات للفقرة ولجميع الفقرات المتضمّنة في تجمّع الفقرات.

ج- اختيار أعلى (n*f) فقرة تزوّد بكميّة معلومات، حيث (n) طول الوحدة، و(f) عدد الألواح المراد تكوينها.

د- حساب دالة معلومات لـ (n*f) فقرة عند كل مستوى قدرة بين (3 , 3-) بفاصل (0.1) مثلاً.

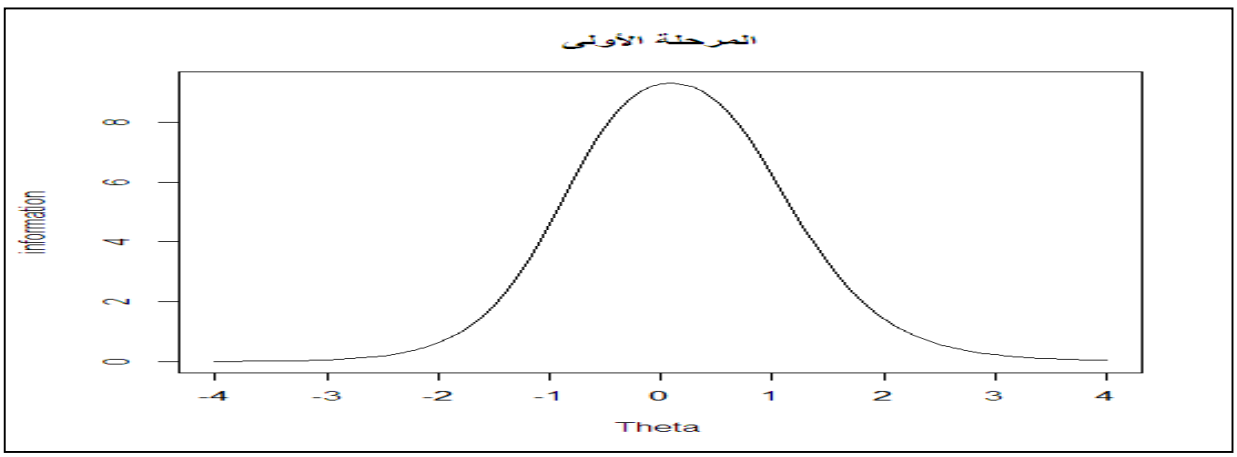
ودالة معلومات الهدف تعرف كما يأتي:

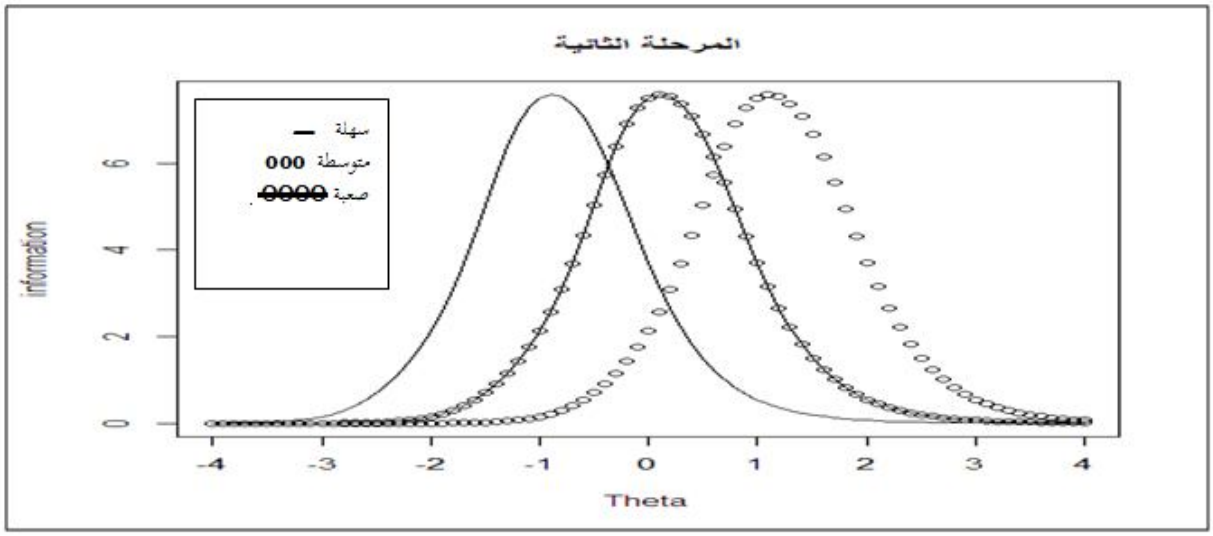
$$TIF(\theta_j) = \frac{\sum_{i=1}^{n*m} I_I(\theta_j)}{m} \quad \ldots\ldots\ldots\ldots\ldots(2)$$

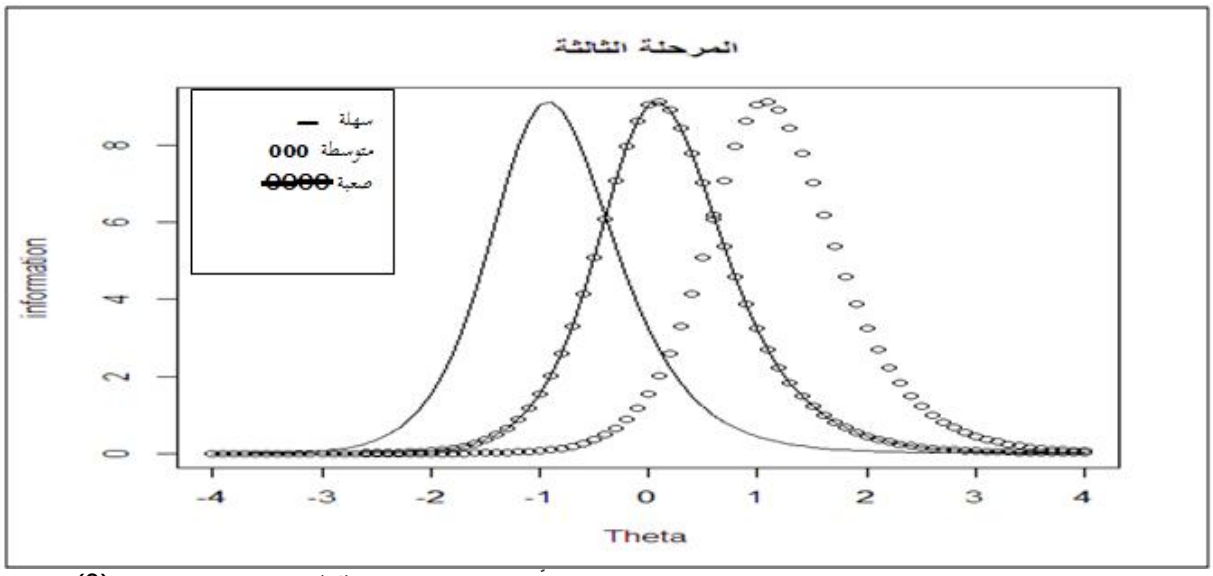
تم بناء خمس وحدات متوازية بفقرات غير متداخلة ولكل مرحلة عبر اللوحات المختلفة باستخدام الخطوات السابقة المذكورة أعلاه. والشكل (3) يعرض دوال المعلومات للوحدات في لوحة من اللوحات ذات التصميم 3-3-1.

الشكل 3

دالة المعلومات للوحدات لواحدة من اللوحات ذات التصميم 3-3-1؛ طول الاختبار (36) فقرة؛ طول وحدة التوجيه (18) فقرة.







في الشكل (3)، كان طول الوحدة في المرحلة الأولى (وحدة التّوجيه) (18) فقرة، وزوّدت بأكبر كميّة معلومات عند مستوى القدرة (θ=0)، بينما كان طول الوحدة في كل من المرحلة الثّانية والمرحلة الثّالثة (9) فقرات وزودت الوحدات السهلة بأكبر كمية معلومات عند مستوى القدرة (θ = -1) والوحدات المتوسطة زودت بأكبر كمية معلومات عند مستوى قدرة (θ = 0)، بينما زودت الوحدات الصعبة بأكبر كمية معلومات عند مستوى قدرة (θ = 1).

وتمّ استخدام البرمجة مختلطة العدد الصحيح (Mixed-Integer Programming: MIP) (Diao & van der Linden, 2011) لحل المشكلة رياضيًا وتكوين الألواح من تجمع الفقرات. ويمكن لهذه الطريقة التعامل مع عدد كبير من قيود المحتوى والقيود الأخرى، وتمّ ترميز الفقرات في تجمع الفقرات i =1, 2, ……I، (V_c) رمز المجموعة الفرعية من الفقرات التي تنتمي لمنطقة المحتوى (c)، وتمّ إعطاء الرمز(n_c) للحد الأدنى لعدد الفقرات في هذه المجموعة، واستخدمت $I_i(\theta)$ للدلالة على دالّة المعلومات للفقرة (i) عند مستوى القدرة (θ). وتم إعطاء طول الاختبار الرمز (N). ويعرف متغير القرار كالآتي:

……(3)

$$X_{if} = \begin{cases} f & \text{إذا كانت الفقرة متضمنة في النموذج} \\ 0 & \text{: خلاف ذلك} \end{cases}$$

$Min\ Z$ ……(4)

تابع إلى:

(5) ……. لكل قيم θ

$\sum_{i=1}^{I} I_i(\theta) X_{if} \leq T_\theta + Z$ ولجميع النماذج،

(6)..... لكل قيم θ

$\sum_{i=1}^{I} I_i(\theta) X_{if} \geq T_\theta + Z$ ولجميع النماذج،

(7) ..... لكل قيم i، $\sum_{f=1}^{5} X_{if} \leq 1$

(8) …..لكل قيم c ولجميع النماذج، $\sum_{i \in V_c} X_{if} \geq n_c$

$\sum_{i=1}^{I} X_{if} = N$ لكل النماذج، .....(9)

$X_i \in \{0, 1\}$ ..... (10)

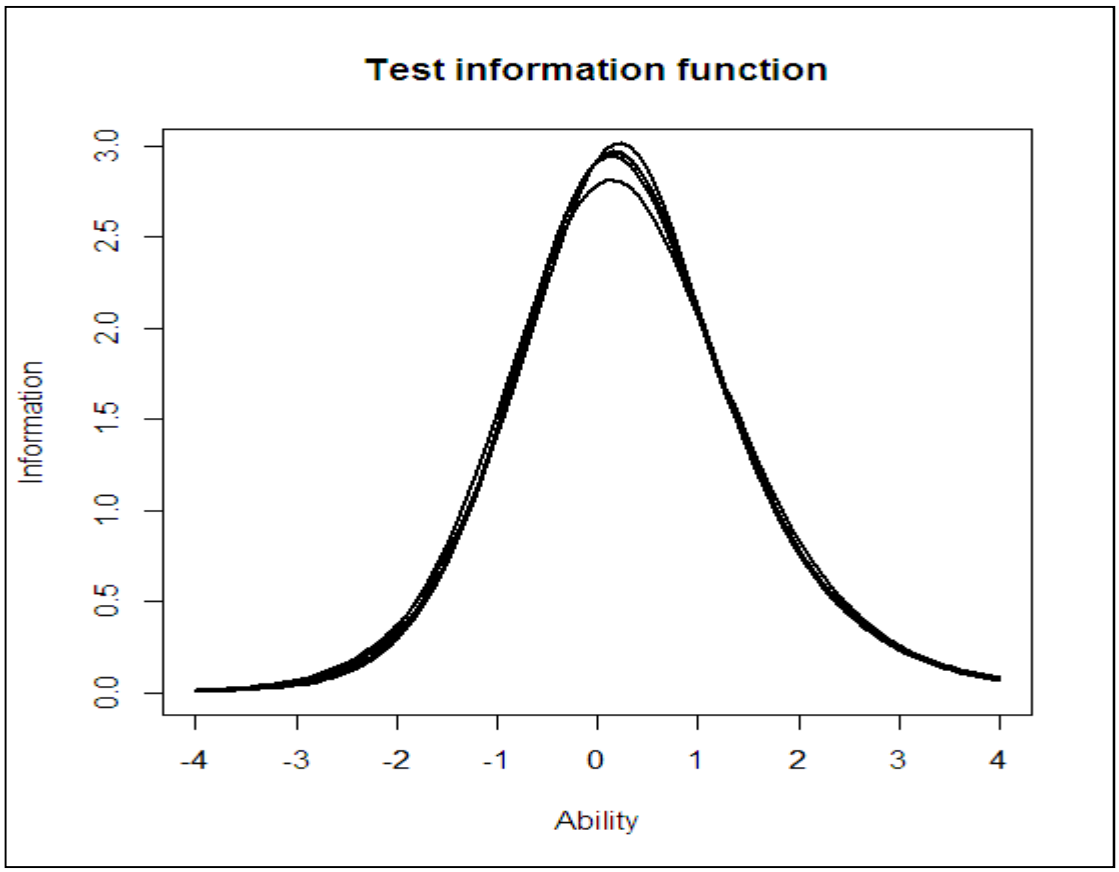
$Z \geq 0$ ..... (11)

ويتعلق القيد في المعادلة (7) بعدم تداخل الفقرات بين الوحدات المختلفة، بينما يتعلق القيد في المعادلة (8) بتوازن المحتوى وعدد الفقرات من كل منطقة من مناطق المحتوى المتضمّنة في كل وحدة، ويتعلق القيد في المعادلة (9) بطول الوحدة، وتتعلق القيود في المعادلتين (5) و(6) بالمسافة بين دالّة المعلومات المستهدفة ودالّة المعلومات للاختبار المجمع. ويجب أن لا تزيد المسافة بين دالة معلومات الوحدة المجمعة ودالة معلومات الهدف لكل قيم (θ) ولجميع النماذج على قيمة (Z).

تمّ بناء الوحدات باستخدام استراتيجية من (أسفل إلى أعلى) وطرق البرمجة المختلطة، ومنها تمّ بناء اللوحات بنجاح. والشكل (4) يعرض دوال المعلومات لوحدة التّوجيه لخمسة ألواح.

الشكل 4

دوال المعلومات لوحدات التوجيه لخمسة ألواح لتصميم 3-3-1؛ طول الوحدة (18) فقرة.



يوضح الشكل (4) دوال المعلومات لوحدات التوجيه لجميع الألواح وطول كل منها (18) فقرة وكل منها جاءت من مناطق المحتوى بعدد متساو من الفقرات بواقع (6) فقرات لكلّ منطقة. ونلاحظ من الشكل أنّ جميع الوحدات زودت بأكبر كمية معلومات عند مستوى قدرة (0). ويظهر الشكل تطابقًا كبيرًا في دوال المعلومات لجميع النماذج.

واستخدمت الدراسة الحالية طريقة المعلومات القصوى التقريبية (AMI)، وطريقة فئات المجتمع المحددة (DPI) لإيجاد نقط التوجيه، التي تحدّد مسار المفحوص عبر المراحل داخل اللوحة، وطريقة (AMI) تستخدم دالّة معلومات الاختبار لتحديد نقط التوجيه من خلال تحديد نقطة التقاطع بين الوحدة التي تقدم للمفحوص والوحدة التي ستقدم له في المرحلة التالية (Luecht et al., 2006).

وتعمل طريقة (DPI) على توجيه نسبة محددة من المفحوصين إلى الوحدات المختلفة في المراحل اللاحقة. وكان من المتوقع في هذه الدراسة أن يتم توجيه المفحوصين إلى كل وحدة في المراحل اللاحقة بأعداد متساوية تقريبًا. ولذلك تم تحديد النسب (33%) و(67%) كنقط توجيه لتصميم اللوحة 3-3-1، بينما تم تحديد النسبة (50%) كنقطة توجيه لتصميم اللوحة 1-2-2. وبما أن توزيع القدرة للمفحوصين يتبع التوزيع الطبيعي وبالاعتماد على التوزيع التراكمي للقدرة، فان قيم القدرة المقابلة للنسب السابقة تساوي (0.43 -) و(0.43) و(0) على التوالي.

المعالجة الإحصائيّة

تمّ الحكم على دقّة قياس تصاميم MST المختلفة من خلال إيجاد وسط مربّعات الخطأ ومتوسّط التّحيّز بعد خضوع كلّ مفحوص من أفراد العينة للاختبار على كل تصميم من تصاميم MST وتكرار الاختبار (10) مرات على جميع أفراد العينة، ومعادلة وسط مربعات الخطأ هي:

$$MSE = \sum_{i=1}^{N} \frac{(\theta_i - \hat{\theta}_i)^2}{N} \quad \ldots\ldots\ldots\ldots\ldots (12)$$

ومعادلة متوسّط التّحيّز هي:

$$\text{Mean Bias} = \sum_{i=1}^{N} \frac{(\theta_i - \hat{\theta}_i)}{N} \quad \ldots\ldots\ldots\ldots (13)$$

حيث N عدد المفحوصين، θ_i تمثل القدرة الحقيقية للمفحوصين، $\hat{\theta}_i$ تمثل القدرة المقدرة للمفحوصين. وتزداد دقة القياس كلما اقتربت (MSE) ومتوسط التحيز من الصفر، ويمكن النظر إلى الفروق في قيم (MSE)، وقيم متوسّط التّحيّز التي تقل عن (0.03) على أنها فروق ضئيلة، وليست ذات أهمية من الناحية العملية بناءً على نتائج الدراسات والأدب السابق المُشابه للدراسة الحالية (Chang & Ying, 1999; Wang, 2017).

نتائج الدّراسة

أولاً: النتائج المتعلقة بسؤال الدراسة: "*هل تختلف دقّة قياس تقديرات القدرة للاختبارات التّكيفية متعدّدة المراحل في ظلّ شروط مختلفة من تصميم الألواح، وثلاثة مستويات من طول الاختبار، ومستويين لطول وحدة التوجيه، واستراتيجيتين للتوجيه (AMI, DPI)***؟"**

للإجابة عن سؤال الدراسة؛ تمّ تقييم دقة القياس من خلال حساب وسط مربعات الخطأ لكل شرط من شروط MST ومقارنة النتائج، كما يظهر في الجداول (3،4،5).

الجدول 3

وسط مربّعات الخطأ لتقدير القدرة لاختبار بطول 36 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
36	18	1-3-3	0.1123	0.1114
36	18	1-2-2	0.1109	0.1107
36	9	1-3-3	0.1125	0.1095
36	9	1-2-2	0.1133	0.1114

يظهر الجدول (3) وسط مربعات الخطأ لجميع شرّوط MST لاختبار بطول (36) فقرة. ويتضح من الجدول أنّ هناك فروقًا ضئيلة في قيم MSE لطريقتي التّوجيه AMI و DPI، حيث كانت الفروق أقل من (0.003) باختلاف تصميم اللوحة (3-3-1، 2-2-1) وطول وحدة التّوجيه (18,9) فقرة. كما يتضح من الجدول أن هناك فروقًا ضئيلة في قيم MSE لتصميمي اللوحة (3-3-1، 1-2-2) باختلاف طريقتي التوجيه (DPI,AMI) وطول وحدة التوجيه (18,9) فقرة، حيث كانت الفروق أقل من (0.002). ويتضح من الجدول أيضًا أن هناك فروقًا ضئيلة في قيم MSE لطول وحدتي التّوجيه (18) فقرة و (9) فقرات باختلاف طريقتي التّوجيه (DPI,AMI) و تصميم اللوحة (3-3-1، 2-2-1)، حيث كانت الفروق أقل من (0.002).

الجدول 4

وسط مربّعات الخطأ لتقدير القدرة لاختبار بطول 24 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
24	12	1-3-3	0.1435	0.1432
24	12	1-2-2	0.1415	0.1413
24	6	1-3-3	0.1416	0.1434
24	6	1-2-2	0.1382	0.1421

يظهر الجدول (4) وسط مربعات الخطأ لجميع شرّوط MST لاختبار بطول (24) فقرة، ويتضح من الجدول أن هناك فروقًا ضئيلة في قيم MSE لطريقتي التوجيه AMI و DPI، حيث كانت الفروق أقل من (0.004) باختلاف تصميم اللوحة (3-3-1، 2-2-1) وطول وحدة التوجيه (12,6) فقرة. كما يتضح من الجدول أن هناك فروقًا ضئيلة في قيم MSE لتصميمي اللوحة (3-3-1، 1-2-2) باختلاف طريقتي التوجيه (DPI,AMI) وطول وحدة التوجيه (12,6) فقرة، حيث كانت الفروق أقل من (0.0034). ويتضح من الجدول أيضًا أن هناك فروقًا ضئيلة في قيم MSE لطول وحدتي التوجيه (12) فقرة و (6) فقرات باختلاف طريقتي التوجيه (DPI,AMI) و تصميم اللوحة (3-3-1، 2-2-1)، حيث كانت الفروق أقل من (0.0033).

الجدول 5

وسط مربّعات الخطأ لتقدير القدرة لاختبار بطول 12 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
12	6	1-3-3	0.2478	0.2322
12	6	1-2-2	0.2245	0.2251
12	3	1-3-3	0.2220	0.2220
12	3	1-2-2	0.2224	0.2231

يظهر الجدول (5) وسط مربعات الخطأ لجميع شروط MST لاختبار بطول (12) فقرة، ويتضح من الجدول أن هناك فروقًا ضئيلة في قيم MSE لطريقتي التّوجيه AMI و DPI، حيث كانت الفروق أقل من (0.004) باختلاف تصميم اللوحة (3-3-1، 2-2-1) وطول وحدة التّوجيه (6،3) فقرات. كما يتضح من الجدول أن هناك فروقًا ضئيلة في قيم MSE لتصميمي اللوحة (3-3-1، 2-2-1) باختلاف طريقتي التّوجيه (DPI,AMI) وطول وحدة التوجيه (6،3) فقرات، حيث كانت الفروق أقل من (0.016). ويتضح من الجدول أيضًا أن هناك فروقًا ضئيلة في قيم MSE لطول وحدتي التوجيه (6) فقرات و(3) فقرات باختلاف طريقتي التوجيه (DPI,AMI) و تصميمي اللوحة (3-3-1، 2-2-1)، حيث كانت الفروق أقل من (0.026).

وأظهرت النّتائج في الجداول (5، 4، 3) أن قيم MSE تقل بزيادة طول الاختبار في مختلف تصاميم MST المستخدمة في الدّراسة الحالية، حيث تراوح مقدار الانخفاض في قيمة MSE بين (0.034) و(0.1383). ويتضح أيضًا أنّ حجم الانخفاض في قيمة MSE يكون ضئيلاً وغير مهم من النّاحية العمليّة عندما يزيد طول الاختبار من (24) إلى (36) فقرة، حيث تراوحت قيمة الانخفاض بين (0.025) و (0.034) باختلاف تصميم اللوحة وطريقة التوجيه وطول وحدة التوجيه، بينما كان حجم الانخفاض في قيمة MSE يتراوح بين (0.0785) و (0.1383) عندما زاد طول الاختبار من (12) فقرة إلى (24) فقرة.

وللإجابة عن سؤال الدراسة أيضًا، تم تقييم دقّة القياس من خلال حساب متوسّط التّحيّز لكلّ شرط من شروط MST، كما يظهر في الجداول (6، 7، 8).

الجدول 6

متوسّط التّحيّز لتقدير القدرة لاختبار بطول 36 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
36	18	1-3-3	-0.001	-0.001
36	18	1-2-2	-0.001	0.001
36	9	1-3-3	-0.002	-0.003
36	9	1-2-2	-0.001	-0.001

يظهر الجدول (6) أن قيم متوسّط التّحيّز متشابهة بشكل كبير ولا يوجد فرق واضح بين تصميم اللوحة وطريقة التّوجيه وطول وحدة التوجيه لاختبار بطول (36) فقرة وجميعها قريبة من الصفر، حيث تراوحت قيم الفرق المطلق لمتوسّط التّحيّز لتصميم اللوحة بين (0.001) و (0.002)، وتراوحت قيم الفرق المطلق لمتوسط التحيز لاستراتيجية التّوجيه بين (0.001) و (0.002)، وتراوحت قيم الفرق المطلق لمتوسّط التّحيّز لطول وحدة التّوجيه بين (0) و (0.001).

الجدول 7

متوسّط التّحيّز لتقدير القدرة لاختبار بطول 24 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
24	12	1-3-3	0.004	0.001
24	12	1-2-2	0.001	-0.001
24	6	1-3-3	0.001	-0.001
24	6	1-2-2	-0.001	0.002

يتضح من الجدول (7) أنّ قيم متوسّط التّحيّز متشابهة بشكل كبير ولا يوجد فرق واضح بين تصميم اللوحة وطريقة التوجيه وطول وحدة التوجيه لاختبار بطول (24) فقرة وجميعها قريبة من الصفر، حيث تراوحت قيم الفرق المطلق لمتوسّط التّحيّز لتصميم اللوحة بين (0.001) و(0.003). وتراوحت قيم الفرق المطلق لمتوسّط التّحيّز لاستراتيجية التوجيه بين (0.001) و(0.003)، وتراوحت قيم الفرق المطلق لمتوسّط التّحيّز لطول وحدة التوجيه بين (0) و (0.001).

الجدول 8

متوسّط التّحيّز لتقدير القدرة لاختبار بطول 12 فقرة.

طول الاختبار	طول وحدة التوجيه	التصميم	AMI	DPI
12	6	1-3-3	-0.002	0.005
12	6	1-2-2	0.001	-0.002
12	3	1-3-3	-0.002	-0.003
12	3	1-2-2	-0.009	-0.001

يظهر الجدول (8) أنّ قيم متوسّط التّحيّز متشابهة ولا يوجد فرق واضح بين تصميم اللوحة وطريقة التوجيه وطول وحدة التوجيه لاختبار بطول (12) فقرة وجميعها قريبة من الصفر، حيث تراوحت قيم الفرق المطلق لمتوسّط التّحيّز لتصميم اللوحة بين (0.001) و(0.007)، وتراوحت قيم الفرق المطلق لمتوسّط التّحيّز لاستراتيجية التّوجيه بين (0.001) و(0.008)، وتراوحت قيم الفرق المطلق لمتوسّط التّحيّز لطول وحدة التّوجيه بين (0.001) و(0.008).

مناقشة النّتائج

بالاعتماد على نتائج سؤال الدراسة التي أظهرت أن قيم MSE ومتوسّط التّحيّز كانت متقاربة لتصميمي اللوحة 2-2-1 و 3-3-1، فإن اختيار وحدتين أو ثلاث وحدات للمرحلتين الثّانية والثّالثة للاختبار لم يكن له تأثير على دقة القياس مع نفس طول الاختبار وطريقة التّوجيه ونفس طول وحدة التّوجيه. وقد تشابهت هذه النتيجة مع نتائج دراسة زنزكي (Zenisky, 2004) ودراسة زنزكي وهامبلتون (Zenisky & Hambleton, 2004) ودراسة كيم وآخرين (Kim et al., 2013) ودراسة وانغ (Wang, 2017)، علمًا بأن تلك الدراسات قد ركزت على أطوال اختبار تتراوح بين المتوسطة والكبيرة، في حين استخدمت الدراسة الحالية أطوال اختبار قصيرة.

ويعزو الباحثان هذه النّتيجة إلى طريقة تحديد نقطة المركز لدالّة معلومات الهدف للوحدات في المرحلتين الثانية والثالثة لكلا التصميمين 2-2-1 و 3-3-1 ، بحيث كانت قيم معاملات الصّعوبة متداخلة مع بعضها البعض عبر الوحدات ولكلّ مرحلة. ففي التصميم 3-3-1 كان مركز دالّة معلومات الهدف للوحدات السّهلة والمتوسّطة والصّعبة ولكلتا المرحلتين الثّانية والثّالثة عند قيم صعوبة (1-)، (0)، (1) على التّوالي، بينما في التصميم 2-2-1 كانت نقط المركز للوحدات السهلة والصعبة ولكلتا المرحلتين الثّانية والثّالثة عند قيم صعوبة (0.7-) و(0.7) على التّوالي. وبمعنى آخر، فقد كانت صعوبة الفقرات للوحدات في المرحلتين الثّانية والثّالثة متداخلة وتغطي مدى واسعًا من متصل القدرة، وهذا ما أكّدت عليه نتائج دراسة كيم وآخرين (Kim et al., 2013).

وأظهرت النتائج أيضًا أنّ قيمة MSE ومتوسّط التّحيّز كانت متقاربة لطريقتي التّوجيه AMI و DPI المستخدمتين في الدّراسة الحالية. وقد اتفقت هذه النتيجة مع نتائج دراسة زنزكي (Zenisky, 2004) ودراسة كيم وآخرين (Kim et al., 2013) ودراسة وانغ (Wang, 2017) ودراسة الغامدي (Alghamdi, 2018).

وأظهرت النتائج أن قيمة MSE ومتوسّط التّحيّز كانت متقاربة باختلاف طول وحدة التّوجيه مع ثبات طول الاختبار. وقد تشابهت هذه النتيجة مع نتائج دراسة باتسولا (Patsula, 1999) ودراسة تشنغ وآخرين (Zheng et al., 2012) ودراسة كيم وآخرين (Kim et al., 2015).

وتعارضت نتائج الدّراسة الحالية مع نتائج دراسة اوزتورك (Öztürk, 2019) التي أظهرت نتائجها أنه كلما زاد طول وحدة التّوجيه تنخفض قيمة الجذر التربيعي لمربعات متوسّط الخطأ وتزداد قيمة معامل الارتباط. ويعزو الباحثان سبب الاختلاف في النتائج بين الدّراسة الحاليّة ودراسة اوزتورك إلى اعتمادها على طول متغيّر للاختبار عند مقارنة تأثير طول وحدة التوجيه على دقة القياس، على عكس الدراسة الحالية التي عملت على تثبيت طول الاختبار عند عملية المقارنة. وبالتالي يعتقد الباحثان أن زيادة دقة القياس كانت بسبب زيادة طول الاختبار وليس بسبب طول وحدة التّوجيه، وهذا ما يتفق مع نتائج الدراسة الحالية من أن زيادة طول وحدة التوجية تؤدي لزيادة في طول الاختبار، وكون الفقرات يتم انتقاؤها بناءً على دالة معلومات الفقرة ربما يكون السبب وراء عدم وجود فروق واضحة في قيمة MSE ومتوسّط التّحيّز.

وأظهرت نتائج الدراسة أن قيمة MSE ومتوسّط التّحيّز تقل بزيادة طول الاختبار، وبالتالي تزداد دقة القياس بزيادة طول الاختبار. وقد اتفقت هذه النتيجة مع نتائج دراسة جودون (Jodoin, 2003) ودراسة جودون وآخرين (Jodoin et al., 2006) ودراسة كيم وآخرين (Kim et al., 2013) ودراسة وانغ (Wang, 2017) ودراسة ساري وهوغنز مانلي (Sari & Huggins-Manley, 2017).

ويعزو الباحثان هذه النّتيجة إلى الصيغة الرياضيّة لدالّة معلومات الاختبار، وهي عبارة عن مجموع دوال معلومات الفقرات المتضمّنة في الاختبار. وبالتالي، من الناحية النظرية، فإن زيادة طول الاختبار تؤدي إلى زيادة كمية المعلومات على طول متصّل القدرة، وزيادة المعلومات تقلّل من قيمة الخطأ المعياري في القياس وتعمل على زيادة دقة تقدير القدرة.

ومن ناحية عملية، فإنّ زيادة طول الاختبار بشكل كبير تؤدّي إلى تحسن قليل في دقّة القياس مقارنة بطول اختبار أقل. وهذا ما يظهر من نتائج الدراسة الحالية. فقد أدّت زيادة طول الاختبار من (24) فقرة إلى (36) فقرة إلى انخفاض ضئيل في قيمة وسط مربّعات الخطأ بلغت قيمته (0.03) عبر جميع شروط MST. وعندما زاد طول الاختبار من (12) فقرة إلى (24) فقرة، انخفضت قيمة وسط مربّعات الخطأ بشكل أكبر، حيث تراوحت قيم الانخفاض بين (0.785) و(0.13). وهذا ما تؤكّد عليه دراسة زنزكي وهامبلتون (Zenisky and Hambleton, 2004)، حيث أشارت إلى أنّ زيادة دالّة معلومات الاختبار الكلية بشكل كبير على (10) تؤدي إلى زيادة ضئيلة في دقّة تقدير القدرة، وهذا الاكتشاف له تأثير على اختيار طول الاختبار أو الاستخدام الأمثل لتجمع الفقرات. وبالتالي، فإنّ عمليّة تحديد طول الاختبار المناسب للاختبارات التّكيفيّة متعدّدة المراحل تعتمد على الاستخدام الأمثل لتجمع الفقرات، ودرجة دقّة القياس التي يرغب مستخدم الاختبار في الحصول عليها. ففي بعض الأحيان، يرغب مطوّر الاختبار في اختبارات قصيرة مع دقّة قياس مقبولة، وفي بعض الأحيان تكون درجة دقّة القياس هي محطّ اهتمام مطوّر الاختبار.

الاستنتاجات والتّوصيات

بيّنت النتائج أنّ عدد فقرات الاختبار هي العامل الأكثر تأثيرًا على دقّة القياس لاختبارات MST بغض النّظر عن الظّروف الاختبارية الأخرى، وقد أشارت النّتائج من خلال إيجاد وسط مربعات الخطأ ومتوسّط التّحيّز بعد خضوع كل مفحوص من أفراد العينة للاختبار على كل تصميم من تصاميم MST وتكرار الاختبار (10) مرات على جميع أفراد العيّنة، إلى أداء أفضل للاختبار الأطول. وبالرجوع إلى دراسة وانغ (Wang, 2017) التي درست دقّة القياس للاختبارات التّكيفيّة متعدّدة المراحل باستخدام مستويين من طول الاختبار (40،60) فقرة، ومستويين من تصميم اللوحة (1-2-2،1-3-3)، ومستويين من استراتيجية التوجيه (AMI، DPI)، يتبيّن اتفاق نتائج هذه الدّراسة مع نتائج دراسة وانغ بالرغم من استخدام وانغ لأطوال أكبر من الاختبار. وعليه، فإنّ الدّراسة الحاليّة توصي بمزيد من الدّراسة للعوامل التي قد تؤثّر على دقّة القياس في اختبارات MST بحثًا عن ظروف اختبارية... قد تسهم في الوصول إلى اختبارات قصيرة نسبيًا وذات دقة قياس عالية. ولأن الطريقة المستخدمة في الدراسة الحالية لتجميع الوحدات هي طريقة البرمجة الخطية (Linear programming integer method) مع استخدام لوحات غير متداخلة، فإن الدّراسة الحالية توصي بإجراء مزيد من الدراسات يتمّ فيها استخدام طرق مختلفة لتجميع الوحدات، وكذلك استخدام تصميم لوحات متداخلة.

وفي ضوء النتائج التي انتهت إليها هذه الدراسة، يوصي الباحثان مستخدمي الاختبارات التّكيفيّة متعدّدة المراحل بمجموعة من التوصيات العملية، منها:

- استخدام طريقة التوجيه (DPI) بدلاً من طريقة التوجيه (AMI)، لسهولة تحديد نقط التوجيه وفق هذه الطريقة مقارنة بطريقة (AMI) التي تحتاج إلى إجراءات تحليل عددي طويلة ومعقدة.
- استخدام تصميم اللوحة 2-2-1، مع وحدة توجيه أطول، بدلاً من التصميم 3-3-1، وذلك لتقليل التكلفة، من خلال بناء عدد أقل من الفقرات. على سبيل المثال، لاختبار بطول (36) فقرة نحتاج إلى (54) فقرة لبناء لوحة ذات تصميم 2-2-1، مقارنة مع (72) فقرة لبناء لوحة ذات تصميم 3-3-1.

References

Alghamdi, H. (2018). *Assessment of multiple-form structure designs of multistage testing using IRT*. Ph.D. Dissertation, University of Denver, U.S.A.

Armstrong, R., & Roussos, L. (2003). *A method to determine targets for multi-stage adaptive tests*. (Computerized Testing Rep. No. 02-07). Law School Admission Council.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211-222.

Diao, Q. & van der Linden, W. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398-409.

Jodoin, M. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1-15.

Jodoin, M., Zenisky, A., & Hambleton, R. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203-220.

Kim, J., Chung, H., Park, R., & Dodd, B. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods*, *45*(4), 1087-1098.

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79.

Leung, C., Chang, H., & Hau, K. (2005). Computerized adaptive testing: A mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology*, *58*(2), 239-257.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc.

Luecht, R. (2000, April). Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high-quality computer-adaptive and mastery tests. In: *Annual Meeting of National Council on Measurement in Education*. New Orleans, LA.

Luecht, R., & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229–249. Retrieved from: https://doi-org.du.idm.oclc.org/10.1111/j.1745-3984.1998.tb00537.x

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Magis, D., Yan, D., & Von Davier, A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

Melican, G., Breithaupt, K., & Zhang, Y. (2009). Designing and implementing a multistage adaptive test: The uniform CPA exam. In: W. J. van der Linden & C.W. Glas (Eds.). *Elements of adaptive testing* (pp. 167-190). Springer.

National Center. for Assessment in Higher Education. (2017). *General aptitude test (GAT)*.

Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in ca-MST. *Universal Journal of Educational Research*, *7*(1), 164-170.

Patsula, L. (1999). *A comparison of computerized adaptive testing and multi-stage testing*. Ph.D. Dissertation, University of Massachusetts, Amherst.

Rudner, L. (2009). Implementing the graduate management admission test computerized adaptive test. In: J. van der Linden & A. W. Glas. (Eds). *Elements of adaptive testing* (pp. 151-165). Springer.

Sari, H., & Huggins-Manley, A. (2017). Examining content control in adaptive tests: Computerized adaptive testing *vs*. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5), 1759-1781.

Swanson, L., & Stocking, M. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*(2), 151-166.

van der Linden, W., Ariel, A., & Veldkamp, B. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, *31*(1), 81-99.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Lawrence Erlbaum Associates, Inc.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing*. (Ph.D. Dissertation, University of Michigan State, U.S.A).

Wang, X., Fluegge, L., & Luecht, R. (2012). A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations. In: *Annual Meeting of the National Council on Measurement in Education*. April 2012, Vancouver, BC, Canada.

Yan, D., Lewis, C. & Von Davier, A. (2014).Overview of computerized multistage test. In: D. Yan, A. von Davier & C. Lewis (Eds.). *Computerized multistage testing: Theory and applications* (pp.3-20). Chapman and Hall/CRC.

Zenisky, A. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment.* (Ph.D. Dissertation, University of Massachusetts Amherst, U.S).

Zenisky, A., & Hambleton, R. (2004). Effects of selected multi-stage test design alternatives on credentialing examination outcomes. Paper presented at the A*nnual Meeting of the National Council on Measurement in Education*, April 2004, San Diego, CA.

Zheng, Y., & Chang, H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly and comparison with other testing modes* (ACT Research Report 6). Retrieved from: https//pdfs. Semanticscholar. org/049b/54fca400fb73116c438a1f834adcc349f07e.pdf.