Practical Assessment, Research & Evaluation

مجلة إلكترونية تخضع لمراجعة الأقران



العدد 16، الإصدار 1، يناير 2011 الرقم الدولي 1351-4177

# إطار عمل لبناء الاختبارات التكيفية المحوسبة

ناثان أ. طومسون، شركة Assessment Systems

ديفيد ج. ويس، جامعة مينيسوتا

على مدى الأربعين عامًا الماضية، تم إجراء العديد من البحوث لدراسة الجوانب التنقية المتعلقة بالاختبارات التكيفية المحوسبة (CAT)، حيث شمل ذلك على سبيل المثال دراسة خوارزميات اختيار مفردات الاختبار، واستراتيجيات التحكم في عرض المفردات الاختبارية، ومعايير إنهاء الاختبار. في المقابل، لا نجد سوى عدد محدود من الدراسات التي تقدِم إرشادات عملية حول كيفية بناء وتطوير الاختبارات التكيفية المحوسبة. تهدف هذه الورقة البحثية إلى الجمع بين عدد من منهجيات البحث الموجودة الخاصة ببناء الاختبارات التكيفية المحوسبة، وتقديم إطار عام يفيد في تطوير أي أنظمة تقييم تعتمد على الاختبارات التكيفية المحوسبة.

الاختبار التكيفي المحوسب (CAT) هو عبارة عن طريقة متطورة لإجراء الاختبارات. خضعت الاختبارات التكيفية المحوسبة للبحث التقني المتخصص على مدى ما يقارب الأربعون عامًا. بالإضافة إلى البحوث التقنية، تم إجراء عدد آخر من الدراسات حول سياق استخدام الاختبارات التكيفية المحوسبة، مثل دراسات مقارنتها بالاختبارات الورقية أو الاختبارات الإلكترونية غير التكيفية (Vispoel, Rocklin, & Wang, 1994)، وتطبيقها على اختبارات بعينها (Sands, Waters, & McBride, 1997; Gibbons et al., 2008). إلا أنه لا توجد دراسات تناولت عملية بناء أو تطوير الاختبارات التكيفية المحوسبة باستثناء أجزاء صغيرة في الكتب التقنية المتخصصة كتناول Flaugher (2000) لبنوك الأسئلة، وتناوُل Wise و Kingsbury (2000) و Parshall, Spray, Kalohn, and Davey (2006) للمشاكل العملية المتعلقة بتطوير وصيانة الاختبارات التكيفية. فضلًا عن ذلك، لم تُقَدِم نتائج وتوصيات البحوث السابقة نموذج عام يُمكِن الاسترشاد به عند بناء الاختبارات التكيفية المحوسبة. لذا، فإن الغرض من هذه الورقة البحثية هو تقديم مِثل هذا النموذج، الذي يُسهِم في إعداد برامج التقييم التي تعتمد على الاختبارات التكيفية المحوسبة، ولكن بحيث يكون في الوقت ذاته إطارًا عامًا ينطبق على جميع برامج التقييم إلا أنه يُركز بشكل أساسي على تقديم التوجيه والإرشاد للجهات التي تتجه حديثًا إلى استخدام الاختبارات التكيفية المحوسبة. يركز البحث بشكل خاص على أهمية دراسات المحاكاة للإجابة على الأسئلة التي تواجه المختصين عند تطوير الاختبارات التكيفية المحوسبة.

يهدف إطار العمل المُقتَرَح (جدول 1) إلى تغطية جميع مراحل عملية بناء الاختبارات التكيفية المحوسبة (CAT)، بدءًا من إعدادها وحتى نشرها. وبالتالي، فإن هذه الإطار لا يُرَكِز على تناول النواحي الخاصة بالقياس النفسي فحسب بل يتجاوز ذلك بكثير. يبدأ الإطار بالإجابة على السؤال المتعلق بما إذا كانت الاختبارات التكيفية المحوسبة ستكون خيارًا مناسبًا لإجراء الاختبارات داخل نظام تقييم معين، ولا يبدأ من قرار المؤسسة بتطبيق الاختبارات التكيفية المحوسبة، لأن مِثل هذا القرار لابد أن تسبقه عمليات بحث ودراسة. فضلًا عن ذلك، توجد العديد من الأسئلة التي يجب الإجابة عليها قبل البدء بتطوير بنك الأسئلة أو منصة إجراء الاختبار. عندما يتم الانتهاء من الإجابة على هذه الأسئلة، يُمكِن الانتقال إلى الخطوات التالية التي يوضحها الجدول رقم (1).

جدول (1) إطار عمل مقترح لبناء الاختبارات التكيفية المحوسبة

| الخطوة | المرحلة | وصف العمل المُتَضَمَن |
|---|---|---|
| 1 | دراسات الجدوى، والتخطيط، وقابلية التطبيق | إجراء محاكاة (تحليل) بطريقة مونتي كارلو؛ تقييم دراسة الجدوى |
| 2 | تطوير محتوى بنك الأسئلة والاستفادة من البنك الموجود حاليًا | كتابة الأسئلة ومراجعتها |
| 3 | الاختبار القبلي ومعايرة المفردات الاختبارية ببنك الأسئلة | الاختبار القبلي؛ تحليل المفردات |
| 4 | تحديد المواصفات الخاصة بالصورة النهائية للاختبار التكيفي المحوسب | المحاكاة اللاحقة المُخَصَصة أو المحاكاة المختلطة |
| 5 | النشر الفعلي للاختبارات التكيفية المحوسبة | النشر والتعميم؛ تطوير البرمجيات |

ستتناول هذه الورقة البحثية بعض المشاكل المتعلقة بكل مرحلة من هذه المراحل. إلا أنه ينبغي الأخذ في الاعتبار أن هذا التناول لن يكون مفصلًا وشاملًا تمامًا، حيث أن برامج التقييم تختلف عن بعضها البعض، حيث يكون لكل منها خصائصه الفريدة وبالتالي مشكلاته الخاصة. علاوة على ذلك، فإننا أولينا اهتمامًا خاصًا لبعض الجوانب الفردية التي تناولتها مصادر أخرى كالبحث الذي أجراه كل من Georgiadou و Triantafillou و Economides (2007) حول استراتيجيات التحكم في عرض مفردات الاختبار. إذًا، يُمْكِن للمؤسسات المختلفة الاستفادة من هذا الإطار العام – بصورته العامة وليس كإطار شامل- في تحديد المشاكل التي تواجههم ونوع الدراسات والقياس النفسي المطلوب، وبالتالي توفير الإرشاد والتوجيه لاتخاذ القرارات المناسبة.

تنبع أهمية هذا الإطار ليس فقط من الناحية العملية ولكن من كونه يوفر أساسًا للتحقق من صدق الإجراءات التي نقوم بها عند بناء الاختبارات التكيفية المحوسبة (CAT). إن تطوير الاختبارات التكيفية المحوسبة دون القيام بعمليات البحث الكافية يجعل هذه الاختبارات أقل كفاءة وأقل فعالية وأقل قوة. فعلى سبيل المثال، قد يؤدي التحديد العشوائي لمواصفات الاختبار التكيفي المحوسب (معيار الإنهاء، الحد الأدنى من مفردات الاختبار) إلى عدم دقة نتائج الممتحنين وبالتالي تصبح هذه النتائج غير مُعَبِرة عن مستوى الطلاب، مما يقلل من صدق التفسيرات التي توفرها هذه النتائج.

## خلفية البحث

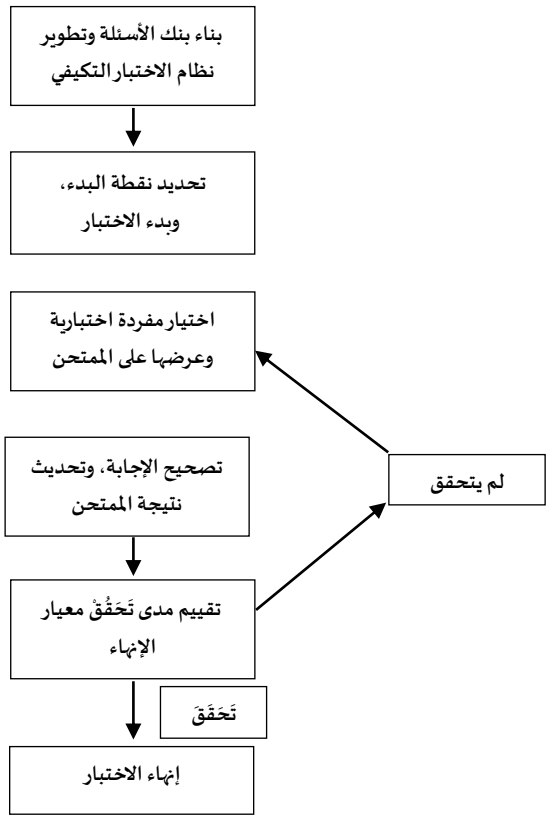
بالرغم من أن العديد من الأدبيات السابقة (Lord, 1980; Wainer, 2000, van der Linden and Glass, 2010) تناولت التفاصيل الخاصة بالاختبارات التكيفية المحوسبة (CAT) كخوارزمية لإجراء الاختبار، إلا أننا سنتناول هنا بعض المعلومات التي نراها ضرورية لتوفير إطار مرجعي للبحث.

من الناحية البنائية، تشتمل عملية تطوير الاختبار التكيفي المحوسب على خمسة مكونات أو مراحل أساسية (Weiss & Kingsbury, 1984; Thompson, 2007). المكون الأول في هذه العملية هو بنك الأسئلة المُعايَر، والذي يتم تطويره في شكل محتوى الاختبار (مفردات الرياضيات لامتحان مادة الرياضيات)، أما المكونات الأربعة المتبقية فتتعلق جميعها بالقياس النفسي لا بالمحتوي ذاته وهي ترتبط بخوارزميات نظام الاختبار التكيفي المحوسب (CAT). هذه المكونات هي:

1. بنك الأسئلة المُعايَر
2. تحديد نقطة البدء
3. خوارزمية اختيار المفردة
4. خوارزمية التصحيح (تحديد الدرجة)
5. معيار إنهاء الاختبار

يبدأ الاختبار التكيفي المحوسب بالتفاعل بين العنصرين الأوليين معًا، ثم يتم الانتقال إلى العنصر الثالث فالرابع فالخامس حتى يتم استيفاء معيار إنهاء الاختبار (شكل 1). فعلى سبيل المثال، يجلس الممتحن على جهاز الكمبيوتر لإجراء الاختبار. يكون جهاز الكمبيوتر محملًا بالفعل بمفردات بنك الأسئلة (الذي يشمل بارامترات لكل مفردة)، ويتم تحديد نقطة بدء معينة للممتحن. يتم اختيار مفردة لنقطة البدء هذه، وتكون هذه المفردة هي السؤال الأول في الاختبار. بعدما يجيب الممتحن على السؤال، يتم تصحيحه ويتم تقدير مستوى الممتحن ($\theta$). يتم تقييم مدى تحقق معيار الإنهاء؛ إذا لم يكن قد تحقق، فيظهر سؤال آخر للممتحن (المكون 3) وبناءً على إجابة الممتحن يتم تحديث الدرجة ($\theta$) (المكون 4)، ويتم تقييم معيار الإنهاء (المكون 5).

وحيث أن إجراء الاختبار التكيفي المحوسب (CAT) يعتمد على التفاعل بين هذه الخوارزميات، فمن الهام تحديد المواصفات المناسبة لكل خوارزمية حيث أن ذلك يُسهم في بناء بنك الأسئلة بشكل صحيح. إن عملية البحث والدراسة لتحديد مواصفات الخوارزميات ليست مفهومة على نطاق واسع، وعادة ما تُترَك هذه الخطوة كليًا لأخصائيّ القياس النفسي المسؤولين عن برنامج التقييم. إن هذه الورقة البحثية لا تقدم فحسب نموذجًا يمكن لأخصائيّ القياس الاسترشاد به، ولكنها تُبَسط بعض المسائل لغير المتخصصين في القياس النفسي من المعنيين بعملية التقييم والمسؤولين عن بعض الأعمال في هذه العملية.




يتم بناء معظم الاختبارات التكيفية المحوسبة (CAT) على أساس نظرية الاستجابة للمفردة [المفردة الاختبارية] (IRT)، وهي عبارة عن نموذج فعال للقياس النفسي يوفر العديد من المزايا لعملية تطوير الاختبارات، وتحليل المفردات الاختبارية، وتحديد درجات الممتحنين. بالنسبة للاختبارات التكيفية المحوسبة، فإن الميزة الأكبر والأهم لاستخدام نظرية الاستجابة للمفردة تتمثل في وضع مفردات الاختبار والممتحنين على نفس المقياس، مما يحقق مطابقة المفردات للممتحنين، وعرض المفردات الأكثر ملاءمة لكل ممتحن. بالرغم من أنه يُمْكِن تصميم الاختبارات التكيفية المحوسبة على أساس نظرية القياس الكلاسيكية (Frick, 1992; Rudner, 2002; Rudner & Guo) [البحث الثاني لا زال قيد الطبع]، إلا أن هذه الميزة التي توفرها نظرية الاستجابة للمفردة الاختبارية تعني أن معظم الاختبارات التكيفية المحوسبة يتم بناءها على أساس هذه النظرية. وبالتالي، فمن الضروري الإلمام بنظرية الاستجابة للمفردة (IRT) حتى يتسنى لنا فهم عملية بناء وتطوير الاختبارات التكيفية. على الرغم من سعينا إلى تقديم إطار عام وشامل لعملية تطوير الاختبارات التكيفية إلا أننا اقتصرنا في المقام الأول على الاختبارات التكيفية المُصَمَمة على أساس نظرية الاستجابة للمفردة وذلك للمزايا التي توفرها

وشيوع استخدامها في هذا المجال. لذا، قد يكون من الضروري تعديل هذا الإطار عند الاسترشاد به في بناء الاختبارات التكيفية المبنية على نظرية القياس الكلاسيكية، أو الاختبارات غير التكيفية بشكل كامل مثل الاختبارات متعددة المراحل أو الاختبارات المبرمجة الثابتة، إلا أن الأسس العامة تظل قابلة للتطبيق على جميع هذه الحالات.

## الخطوة الأولى: دراسات الجدوى والتخطيط وقابلية التطبيق

إن الخطوة الأولى في تطوير الاختبارات التكيفية المحوسبة (CAT) تتمثل في تحديد ما إذا كان استخدام الاختبارات التكيفية المحوسبة (CAT) في برنامج التقييم سيكون ذا جدوى. نظرًا لأن خوارزمية الاختبارات التكيفية جذابة مفاهيميًا وتوفر مزايا مؤكدة؛ لذا، نجد أن بعض المعنيين بأنظمة التقييم من غير المتخصصين في القياس النفسي قد يتحمسون للفكرة ويرغبون في تطبيقها دون الإلمام الكافي بها وبكيفية تطويرها ومتطلباتها. فقد يتنامى إلى علم المسؤول التنفيذي للمؤسسة أن عدد مفردات الاختبار التكيفي لا تتجاوز نصف عددها في الاختبار التقليدي (Weiss & Kingsbury, 1984)، هذا يعني توفيرً في الوقت وبناءً على لذلك يتخذ قرارًا بالانتقال ببرنامج التقييم الخاص بالمؤسسة إلى الاختبارات التكيفية المحوسبة. إن مِثل هذا القرار قد يكون بالغ الخطورة، ليس فقط من حيث صعوبة ذلك من ناحية القياس النفسي، ولكن أيضًا من الناحية التجارية. إن قرار الانتقال من برنامج التقييم الذي يعتمد على الاختبارات ثابتة الشكل إلى برنامج تقييم يعتمد على الاختبارات التكيفية المحوسبة لابد أن يكون مدروسًا.

لذا، ينبغي أولًا دراسة الاعتبارات التجارية والعملية المرتبطة بتطبيق الاختبارات التكيفية. فهل لدى المؤسسة خبراء في القياس النفسي، وإذا لم يكن لدى المؤسسة خبراء في هذا المجال، هل ستكون المؤسسة قادرة على تحمل تكاليف استشارة خبير خارجي؟ هل لدى المؤسسة القدرة على تطوير بنوك أسئلة شاملة؟ هل يتوافر لدى المؤسسة منصة لإجراء الاختبارات التكيفية المحوسبة، أو هل تمتلك المؤسسة الموارد الكافية لتطوير منصة خاصة بها؟ هل التحول إلى استخدام الاختبارات التكيفية المحوسبة سيؤدي إلى تقصير الاختبار؟ هل قِصَرْ طول الاختبار يمكن أن يُسهِم في توفير الوقت الذي يستغرقه الممتحن، والذي يعني بدوره التوفير الفعلي في النفقات؟ أو إذ كانت تكلفة الاختبارات التكيفية المحوسبة أعلى وإذا كانت لا تؤدي إلى تقليل وقت الاختبار، فهل هذا يقابله زيادة في دقة النتائج وزيادة في الأمان مما يجعلها خيارًا مجديًا للمؤسسة؟

لحسن الحظ، يُمكِن الإجابة على الكثير من هذه الأسئلة ليس عن طريق التخمين ولكن عن طريق بحوث القياس النفسي. إن دراسات المحاكاة (التحليل) بطريقة مونتي كارلو تُمَكِن الباحثين ليس فقط من تقدير طول الامتحان ودقة الدرجة التي يعطيها الاختبار التكيفي، ولكنها تمكنهم أيضًا من تقييم بعض الجوانب الهامة مثل طرق التحكم في عرض المفردات والحجم المطلوب لبنك الأسئلة بما يساعد في تحقيق الدقة المطلوبة في نتائج الاختبار. تعتمد هذه الدراسات على محاكاة الاختبارات التكيفية المحوسبة تحت ظروف مختلفة لعدد كبير من الممتحنين الوهميين، ثم تتم مقارنة النتائج لاتخاذ القرارات المناسبة.

فعلى سبيل المثال، يُمْكِن محاكاة الاختبارات التكيفية باستخدام بنك أسئلة يحتوي على 300 مفردة، وبنك أسئلة آخر يحتوي على 500 مفردة، ثم مقارنة النتائج لتحديد أي منهما يحقق هدف المؤسسة بشكل أفضل. تتمثل أهمية هذه الطريقة في أن دراسات المحاكاة بطريقة مونتي كارلو يمكن أن تتم قبل كتابة مفردات الأسئلة وقبل توافر أي بيانات حقيقية.

تعتمد المحاكاة بطريقة مونتي كارلو على حقيقة أن نظرية الاستجابة للمفردة الاختبارية تعطي تقديرًا دقيقًا لاحتمال الاستجابة الصحيحة على مفردة ما في ضوء قيمة تقديرية مُعطاة ($\theta$). وهذا يُمَكِن الباحثين من توليد استجابات لمفردات الاختبار في ضوء البارامترات الخاصة بالمفردة وقيمة $\theta$. فعلى سبيل المثال، إذا افترضنا أن الممتحن ذو المستوى المتوسط تكون احتمالية استجابته الصحيحة على مفردة ما هي 0.75. في هذه الحالة يبدأ النظام بتوليد رقم عشوائي من مجموعة الأرقام الحقيقية التي تنحصر ما بين صفر و 1. إذا كان الرقم المُختار يساوي أو يقل عن 0.75 فإن استجابة الممتحن على المفردة ستكون "صحيحة" على الأرجح. وإذا كان الرقم العشوائي الذي تم اختياره يزيد عن 0.75 فإن استجابة الممتحن ستكون "غير صحيحة" على الأرجح. وعند توافر بارامترات خاصة بالمفردات الاختبارية الموجودة ببنك الأسئلة وقيم $\theta$ للممتحنين، يمكن إنشاء مجموعة كبيرة من الاستجابات الصحيحة/ غير الصحيحة لجميع المفردات. يمكن أن يتم توليد البارامترات الخاصة بمفردات الاختبار والممتحنين اعتمادًا على بيانات حقيقية أو بشكل عشوائي، وذلك اعتمادًا على مدى إتاحة البيانات في كل مرحلة من مراحل تطوير الاختبارات التكيفية. إذا تم توليد البارامترات عشوائيًا، فإن الاعتماد على البارامترات المتوقعَة كأساس لعملية توليد البارامترات يجعل المحاكاة أكثر قوة. إذا وجِد أن اختبارات مماثلة في البحوث المنشورة لها متوسط تمييز قدره 0.7، يصبح من المنطقي إنشاء بنك أسئلة يعكس هذا المُعطى.

يُمْكِن أن تُستخدَم مجموعة البيانات هذه في محاكاة الاختبارات التكيفية المحوسبة. تكون محاكاة الاختبارات التكيفية مشابهة تمامًا للاختبارات التكيفية الحقيقية باستثناء أن الاستجابة على المفردة لا تتم من قِبَل ممتحن فعلي في الوقت الحقيقي، ولكن يتم البحث عنها في جدول الاستجابات المولَدة أو المُولَدة في الوقت الحقيقي. إذا تم اختيار مفردة بعينها في الاختبار التكيفي المحوسب لعرضها على الممتحن، فإن برنامج المحاكاة يقوم ببساطة بتوفير البارامترات الخاصة بالاستجابة من مجموعة البيانات.

وحيث أن محاكاة الاختبارات التكيفية المحوسبة بطريقة مونتي كارلو لا يمكن أن تتم إلا باستخدام برامج متخصصة، فإن الخطوة الأولى هنا تتمثل في الحصول على البرامج اللازمة. من الضروري وجود نوعين من البرامج: النوع الأول يقوم بتوليد مجموعات البيانات بناءً على المواصفات التي توفرها، أما النوع الثاني فيقوم بمحاكاة الطريقة التي ستتم بها الاختبارات التكيفية. يتيح كل من برنامج WINGEN (Han, 2007) وبرنامج PARDSIM (Yoes, 1997) إمكانية محاكاة مجموعات البيانات على أساس نظرية الاستجابة للمفردة الاختبارية (IRT) (Embretson & Reise, 2000) وذلك في ظل وجود مجموعة كبيرة من المواصفات المُحَدِدة. ثم بعد ذلك يُمْكِن محاكاة الاختبارات التكيفية المحوسبة باستخدام

برنامج FireStar (Choi, 2009) أو برنامج CATSim (Weiss & Guyer, 2010). يتميز برنامج CATSim بالجمع بين وظائف هذين النوعين من البرامج، كما يُمكنه محاكاة مجموعات البيانات الخاصة به بطريقة مونتي كارلو، والاستفادة من مجموعات البيانات الحقيقية، أو دمجهما معا بالتوافق مع محاكاة الاختبار التكيفي المحوسب. إذا كان القائمين على برنامج التقييم في مؤسستك لديهم خبرة كبيرة في القياس النفسي، فيمكن تطوير برنامج المحاكاة داخليًا، إلا أن تكلفة ذلك قد تتجاوز -على الأرجح- تكلفة الحصول على البرامج الموجودة بالفعل.

توجد العديد من المتغيرات التابعة الهامة التي يجب أخذها بعين الاعتبار عند إجراء المحاكاة بطريقة مونتي كارلو. أهم هذه المتغيرات هما متوسط طول الاختبار ودقته، التي تُحَدَد كميًا بالخطأ المعياري في القياس. في الاختبارات التقليدية، يكون طول الاختبار ثابتًا ولكن تكون دقة الاختبار متغيرة، وبالتالي فالممتحنين في منتصف التوزيع عادة ما يكون الخطأ المعياري في حالتهم أقل وذلك عندما يتعلق الأمر بقياس قدراتهم الكامنة وذلك لأن مفردات الاختبار متوسطة الصعوبة هي الأكثر شيوعًا. أما في حالة الاختبارات التكيفية، فإن طول الامتحان عادة ما يكون متغيرًا، ولكن هذه الاختبارات تم تصميمها لتحقيق دقة مكافئة لجميع الممتحنين، وذلك إذا تم تصميم مفردات الأسئلة بشكل صحيح وهو مما يتطلب دراسات محاكاة فعالة.

الخطوة التالية في هذه المرحلة هي تقييم دراسة الجدوى اعتمادًا على نتائج المحاكاة بطريقة مونتي كارلو. فعلى سبيل المثال، إذا افترضنا أن لدينا برنامج تقييم حالي يعتمد على استخدام أربعة نماذج من الاختبارات التقليدية ثابتة التسلسل يتكون كل منها من 100 مفردة، مع وجود 20 مفردة مشتركة لتحقيق التكافؤ بين نسخ الاختبار. هذا يعني أن لدينا بنك أسئلة يضم 340 مفردة. ربما كان يُعتقَد في السابق أن التحول إلى الاختبارات التكيفية المحوسبة يتطلب وجود بنك أسئلة يتألف من 1.000 مفردة على الأقل إلا أن محاكاة مونتي كارلو أثبتت أن بنك أسئلة مُكَوَن من 500 مفردة قد يكون كافيًا. وباعتبار أن بنك الأسئلة يضم حاليًا 340 مفردة، فإن التكلفة الإضافية لإعداد المفردات الجديدة المطلوبة ستكون أقل بكثير من المتوقع. علاوة على ذلك، أثبتت تجارب المحاكاة أن بنك الأسئلة المُكَون من 500 مفردة قد يسهم في إعداد اختبارات لها نفس مستوى الدقة الذي تتمتع به الاختبارات الحالية، ولكن بمتوسط 55 مفردة أو سؤال لكل امتحان. ولكن، هل سيتم تعويض التكلفة التي ستتحملها المؤسسة لإضافة 160 مفردة إلى بنك الأسئلة، وإجراء الدراسات اللازمة للاختبارات التكيفية المحوسبة؟ وهل الانتقال للاعتماد على الاختبارات التكيفية سيكون ذا جدوى لتوفيرها الوقت الذي يستغرقه الممتحن في الاختبار بتقليل مفردات الاختبار إلى 45 مفردة، وإتاحتها مزيدًا من الأمان بالاعتماد على نُسَخ متعددة بدلًا من أربعة نسخ؟ تلك هي نماذج للأسئلة التي تُشكل أساس هذه المرحلة، ولكن يجب أيضًا أن نضع في الاعتبار المزايا غير التجارية لاستخدامها، مثل التمكن من قياس قدرة جميع الممتحنين بدقة متساوية وتحسين تجربة الممتحن وزيادة دافعيته بعرض المفردات الملائمة له فقط.

## الخطوة 2: تطوير محتوى بنك الأسئلة

عندما يتم اتخاذ قرار بشأن التحول إلى الاختبارات التكيفية المحوسبة، فإن الخطوة التالية تتمثل في إعداد بنك الأسئلة. يجب الإشارة هنا إلى أن هذه الخطوة يجب أن تتم استنادًا إلى أدلة تجريبية قدر المستطاع. فدراسات المحاكاة التي تم إجراؤها في الخطوة السابقة يمكن الاستفادة منها أو ربما توسيع نطاقها لتقدِم إرشادات حول كيفية تطوير بنك الأسئلة. يشير Veldkamp و van der Linden (2010) إلى أن دراسات المحاكاة هامة ومفيدة في هذه الخطوة ولا تقتصر الاستفادة منها على التجريب القبلي فقط كما ذكر Flaugher (2000).

عند إعداد بنك الأسئلة، لا يتم التركيز فقط على عدد المفردات التي سيضمها البنك، ولكن يمتد ليشمل تناوُل محددات توزيع المفردات، والاعتبارات العملية المتعلقة بتوزيع المحتوى، والمشاكل المتعلقة بالتحكم في عرض المفردات. عند إجراء المحاكاة، لابد أن تتعدد وتتنوع الظروف المرتبطة بها، كأن يتم استخدام بنك أسئلة يضم مجموعة كبيرة من الأسئلة ذات مستويات صعوبة مختلفة ومقارنتها ببنك به عدد أقل من الأسئلة، أو به أسئلة ذات مستويات صعوبة غير متكافئة، أو استخدام بنك أسئلة يحتوي على أسئلة عالية التمييز مقارنة بأسئلة أخرى منخفضة التمييز. قدم كل من Veldkamp و van der Linden بحثًا حول الممارسات المثلى في عملية البحث المتعلقة ببنوك الأسئلة، كما قدم Reckase (2003) معلومات قيمة أيضًا في هذا الصدد.

إن أحد الاعتبارات الهامة التي يجب مراعاتها عند تصميم دراسات المحاكاة للاختبارات التكيفية هو ضرورة أن تتطابق وظيفة معلومات الاختبار مع الهدف من الاختبار (Embretson & Reise, 2000). فإذا كان الاختبار سيُستخدَم لأغراض تصنيف الطلاب على أساس درجة قطعية (ناجح/راسب) فإن الاختبار لابد أن يتضمن مزيدًا من المفردات التي تقترب من هذه الدرجة، لا المفردات التي تقيس الحدود العليا للمهارة. حيث أننا في هذه الحالة لا نحتاج إلى الحصول على درجات دقيقة تعبر عن مستوى اتقان الممتحنين للحدود العليا للمهارة، وبالتالي تصبح الأسئلة عالية الصعوبة غير مطلوبة. وعلى العكس من ذلك، إذا كنا بحاجة إلى الحصول على درجات دقيقة لجميع الممتحنين، ذوي القدرات العالية والمنخفضة على حد سواء، فيصبح من الضروري إدراج أسئلة عالية الصعوبة وأخرى منخفضة الصعوبة، بالإضافة إلى أسئلة متوسطة الصعوبة بحيث نغطي جميع مستويات الطلاب. في هذه الحالة، يصبح من الضروري وجود مجموعة كبيرة من الأسئلة عالية الصعوبة وأخرى منخفضة الصعوبة.

ولحسن الحظ، في كثير من الحالات، لا يكون من الضروري بناء أو إعداد بنك أسئلة جديد، حيث يمكن الاستفادة من بنك الأسئلة القديم. حتى أنه من المفيد القيام بذلك لأغراض الاستمرارية. فمن خلال الربط ودمج الأسئلة المضافة حديثًا مع تلك الموجودة بالفعل، يُمْكِن ضمان ثبات مقياس نظرية الاستجابة للمفردة أثناء تحولنا إلى الاختبارات التكيفية المحوسبة. كما أن ذلك يسهم في تقليل عدد المفردات المطلوب تطويرها أو إضافتها إلى بنك الأسئلة الموجود بالفعل.

وسواء كان بنك الأسئلة سيتألف من مفردات جديدة تمامًا أو سيضم مزيجًا من

المفردات الجديدة والقديمة، فمن المهم التركيز على الاختبارات الإحصائية التي تتطلبها المفردات الاختبارية. فإذا كانت المعايير الخاصة ببرنامج التقييم عالية وعادة ما يتم استبعاد نسبة مئوية كبيرة من المفردات أثناء عملية التطوير، فإن ذلك لابد أن يؤخذ بعين الاعتبار في هذه المرحلة.

## الخطوة 3: الاختبار القبلي، التدريج (المعايرة)، والربط

عند الانتهاء من عملية تطوير وبناء مفردات الاختبار، يجب أن يتم اختبارها قبليًا. إن هذا الأمر ضروري للغاية لأن مفردات الاختبار يتم مطابقتها مع مستوى الممتحنين بناءً على البارامترات الخاصة بالمفردة وذلك في ضوء نظرية الاستجابة للمفردة الاختبارية، كما أن البارامترات الخاصة بالمفردة يتم تقديرها عن طريق التحليل الإحصائي للاستجابات الصادرة من ممتحنين فعليين. يتحدد حجم العينة المطلوبة للاختبار القبلي على أساس نموذج نظرية الاستجابة للمفردة الاختبارية الذي يتم توظيفه. فعلى سبيل المثال، يقترح Yoes (1995) أن العينة يجب أن تتراوح ما بين 500-1000 ممتحن لكل مفردة وذلك عند استخدام النموذج ثلاثي البارامترات لنظرية الاستجابة للمفردة. يُمْكِن الرجوع إلى بحث Flaugher (2000) للحصول على وصف تفصيلي للمواضيع المرتبطة بهذه الخطوة.

توجد طريقتان لإجراء الاختبارات القبلية لمفردات الاختبار، وذلك اعتمادًا على ما إذا كانت مفردات بنك الأسئلة للاختبارات التكيفية المحوسبة جديدة جميعها أم سيتم دمج المفردات الجديدة مع القديمة، وما إذا كان من المتوجب إبقاء الاختبارات الحالية قيد التشغيل أثناء عملية تطوير المفردات وإجراء الاختيارات القبلية أم لا. إذا كانت جميع مفردات الاختبار ببنك الأسئلة جديدة، فيُمكِن ببساطة عرض عدد كبير من مفردات الاختبار في كل امتحان. فمثلًا إذا كننا بصدد إعداد بنك أسئلة يضم 400 مفردة، فإن كل ممتحن قد تُعرَض ليه 100 مفردة. وفي المقابل، إذا تم دمج المفردات الجديدة والقديمة، وكان من الضروري إبقاء الاختبارات الحالية قيد التشغيل أثناء عملية الاختبار القبلي، فإن مفردات الاختبار الجديدة قد يتم دمجها في الاختبارات الحالية قيد التشغيل. ودعونا نستخدم نفس المثال السابق، حيث كان من المتطلب إضافة 160 مفردة جديدة إلى المفردات الموجودة ببنك الأسئلة والبالغ عددها 340 مفردة. وحيث أن بعض مفردات الاختبار لن تكون على الأرجح جيدة ودقيقة كما نأمل، فلنفترض أننا سنقوم بإجراء اختبار قبلي لعدد 200 مفردة. إذا كان الممتحن سيجرى اختبار ثابت الشكل مكون من 100 سؤال، فإن عرض الـ 200 مفردة الجديدة سوف يزيد من طول الاختبار بمقدار ثلاثة أمثال، الأمر الذي سيستغرق وقتًا طويلًا من الممتحنين. وحيث أننا نريد اختبار 200 مفردة، ولدينا 4 نماذج للاختبار، فسيكون من المناسب إضافة 50 مفردة جديدة لكل ممتحن. يمكن اختيار الـ 50 مفردة الجديدة عشوائيًا أو وضعها في مجموعات مُحَدَدة مسبقًا باستخدام برامج مختلفة (Verschoor, 2010). إن الهدف من ذلك هو ترتيب مفردات الاختبار الخاضعة للاختبار القبلي بحيث يتم عرضها على عدد كافي من الممتحنين لتوفير الحد الأدنى من الاستجابات المطلوبة على كل سؤال.

بعد الانتهاء من الاختبار القبلي، يتم تقدير البارامترات الخاصة بالمفردات الاختبارية عن طريق استخدام برنامج معايرة (تدريج) لنظرية الاستجابة للمفردة. الخطوة الهامة في هذه المرحلة هي ***الربط***، الذي يضمن تدريج جميع البارامترات الخاصة بالمفردات الاختبارية على مقياس واحد. توجد العديد من الطرق لتحقيق ذلك، ولكن يجب التمييز بين الطرق التي تقوم بمعايرة المفردات الجديدة على مقياس موجود بالفعل (Stocking & Lord, 1980) وبين تلك التي تعتمد على بناء مقياس جديد (Lee & Weiss, 2010). وبطبيعة الحال، إذا كننا بصدد بناء بنك أسئلة جديد، فنادرًا ما تكون هناك حاجة إلى ربط المفردات الجديدة بمقياس موجود بالفعل. وبالمثل، إذا كنا سنقوم بإضافة بعض المفردات الاختبارية الجديدة إلى اختبار موجود بالفعل ومن الضروري الإبقاء على المقياس المستخدَم، فلن يكون من المناسب بناء مقياس جديد. يُرجى الاطلاع على ما قدمه Kolen و Brennan (2004) في هذا الصدد.

تنطوي خطوة المعايرة هذه على عمليات تحليل إحصائي إضافية. ففي معظم الأحيان، تتم مراجعة البيانات الإحصائية الخاصة بمعامل التمييز والصعوبة لكل مفردة لتحديد المفردات التي يجب حذفها أو مراجعتها أو إعادة اختبارها. حتى إذا كان برنامج الاختبار مبنيًا بالفعل على نظرية الاستجابة للمفردة، فستكون العمليات الإحصائية القياسية مفيدة جدًا لهذا الغرض. بالإضافة إلى ذلك، توجد عملية إحصائية إضافية يتم إجراؤها على مستوى كل مفردة، ألا وهي ***تحليل مطابقة النموذج***، والذي يُقصَد به مدى مطابقة البيانات لنموذج الاستجابة للمفردة الذي تم استخدامه للمعايرة. بالنسبة للمفردات التي لها مشاكل كبيرة تتعلق بالتحرر من السرعة والقابلية للتخمين، فعادة ما يكون من غير المناسب تضمينها في الاختبارات التكيفية المحوسبة حيث أن بارامترات الاستجابة للمفردة الخاصة بها لا تتمتع بثبات كافٍ لاستخدامها في الاختبارات التكيفية المحوسبة.

وأخيرًا، من الضروري إجراء تحليل الأبعاد في هذه المرحلة. تفترض نظرية الاستجابة للمفردة أن الاختبار يكون أحادي البعد (ما لم يتم استخدام نماذج الاستجابة للمفردة متعددة الأبعاد)؛ لذا، فإن مفردات بنك الأسئلة التي تخضع للاختبار القبلي لابد أن يتم تحليلها عامليًا لضمان تحقق ذلك. إن الإجراء المناسب هنا هو إجراء التحليل العاملي باستخدام الارتباط الرباعي الذي يمكن قياسه باستخدام برنامج MicroFACT (Waller, 1997) أو إجراء *التحليل العاملي لجميع المعلومات* باستخدام TESTFACT 4 (Bock et al., 2003). يقترح Bejar (1980؛ 1988) طريقة بديلة لتقييم الأبعاد داخل إطار نظرية الاستجابة للمفردة.

## الخطوة 4: تحديد المواصفات للاختبار التكيفي المحوسب النهائي

في هذه المرحلة، نكون قد انتهينا بالفعل من إعداد بنك الأسئلة ومعايرة المفردات الاختبارية باستخدام نظرية الاستجابة للمفردة، إلا أن هذه الخطوات السابقة لا تمثل سوى المُكَوِن الأول من إجمال خمسة مكونات رئيسية في عملية بناء الاختبارات التكيفية المحوسبة. لذا، يجب تحديد المكونات المتبقية ودراستها جيدًا قبل نشر وتعميم الاختبارات التكيفية المحوسبة. يجب ألا تبنى هذه المكونات على قرارات عشوائية ولكن ينبغي أن تُبنى على دراسات محاكاة مثلما حدث عند إعداد بنك الأسئلة (Flaugher, 2000). إلا أن هناك اختلافًا واحدًا هامًا في هذه المرحلة: فلدينا الآن بنك أسئلة تم إعداده بالفعل ولدينا بيانات تتضمن استجابات

لمتحنين حقيقيين على المفردات الاختبارية. وبطبيعة الحال، دائمًا ما تكون البيانات الحقيقية أفضل من البيانات المُولّدة عشوائيًا إذا كان الهدف منها هو تحديد الكيفية التي ستتم بها الاختبارات التكيفية المحوسبة مع الممتحنين الحقيقيين في المستقبل. وبالتالي فإن هذه البيانات يُمْكِن الاستفادة منها في دراسات محاكاة جديدة، تُسَمى *المحاكاة اللاحقة المُخَصَصة أو المحاكاة المعتمِدة على بيانات حقيقية*.

في المحاكاة المعتمِدة على بيانات حقيقية، مثلما هو الحال في محاكاة مونتي كارلو، تتم محاكاة الاختبارات التكيفية المحوسبة لكل ممتحن على أساس الاستجابات الخاصة بكل مفردة في بنك الأسئلة. الاختلاف الوحيد بينهما هو أن محاكاة مونتي كارلو تقوم بتوليد الاستجابات الخاصة بكل ممتحن على كل مفردة، بينما في المحاكاة اللاحقة المُخَصَصة يتم الاستفادة من البيانات الحقيقية المتوافرة لدينا. فعلى سبيل المثال، إذا كانت محاكاة الاختبار التكيفي للممتحن الأول قد حددت المفردة الـ 19 على أنها يجب أن تكون المفردة الأولى التي يتم عرضها للممتحن، فستقوم محاكاة مونتي كارلو بتوليد استجابة لهذه المفردة اعتمادًا على بارامتر المفردة، وبارامتر الممتحن (θ)، ونموذج نظرية الاستجابة للمفردة المُفترَض. وعلى العكس من ذلك، في المحاكاة اللاحقة لن تكون هناك حاجة لتوليد استجابة للمفردة، لأن خوارزمية المحاكاة تقوم بالبحث عن استجابة الممتحن الأول على المفردة الـ 19.

من عيوب هذا النوع من المحاكاة مع تصميمات الاختبار القبلي أن الممتحن تُعرَض عليه نسبة محدودة من المفردات الموجودة ببنك الأسئلة. ففي المثال السابق، سوف يُعرض على كل ممتحن 150 مفردة فقط من البنك الذي تم تطويره والذي يشتمل على 450 مفردة (إذا افترضنا عدم استبعاد أي مفردة): 100 مفردة من النماذج الموجودة و50 مفردة جديدة. إذا تم تطبيق المحاكاة المعتمدة على بيانات حقيقية على مجموعة البيانات هذه، فلن تتوافر استجابة لكل ممتحن على عدد 390 مفردة. لحل هذه المشكلة، تم تطوير نوع ثالث من أنواع المحاكاة يُسمى ***المحاكاة المختلطة*** (Weiss & Nydick, 2009; Weiss & Guyer, 2010)، حيث يتم فيها استخدام البيانات الحقيقية في حال توافرها، ولكن يتم أيضًا توليد الاستجابات غير المتوفرة باستخدام طريقة مونتي كارلو وذلك بالاستناد إلى درجة (θ) التي يتم تقديرها بناءً على استجابة الطالب/الطالبة على الأسئلة التي أجابها. وهذا يجعل الاختبارات التكيفية المحوسبة أكثر فعالية مع بنوك الأسئلة الحقيقية والممتحنين الحقيقيين.

تُعَد عمليات المحاكاة المعتمدة على بيانات حقيقية والمحاكاة المختلطة هامة وأساسية لمقارنة وتقييم الطرق والمواصفات المختلفة الخاصة بالمكونات الخوارزمية الأربع للاختبار التكيفي المحوسب عند استخدامها مع بنك أسئلة حقيقي. توجد مجموعة من الأسئلة الهامة المرتبطة بكل مكون من هذه المكونات والتي ينبغي الإجابة عليها، مثل الأسئلة المتعلقة بمقارنة الطرق المختلفة للتحكم في عرض المفردة، أو المتعلقة بتطبيق قيود اختيار المحتوى على خوارزمية اختيار المفردة. توجد بعض البرامج المُصَمَمة للإجابة على مِثل هذه الأسئلة مثل برنامج CATSim (Weiss & Guyer, 2010). إن نشر وتعميم الاختبارات التكيفية المحوسبة دون إجراء الدراسات الكافية يجعل هذه الاختبارات أقل كفاءة. فعلى سبيل المثال، قد يكون محتوى بنك الأسئلة غير كافيًا لتحقيق المتطلبات الخاصة بالتحكم في اختيار المفردة، أو الموازنة بين المحتوى المعروض، أو تحديد خوارزميات إنهاء الاختبار. بدون دراسات المحاكاة، قد لا يُمْكِن تحقيق ذلك حتى بدء إجراء الاختبارات الفعلية.

### *بنك الأسئلة*

لا يجب بالضرورة استخدام بنوك الأسئلة كما هي. حينما يشتمل بنك الأسئلة على 500 مفردة، قد تكون جودة المفردات أعلى قليلًا من المتوقع، وقد يكون بنك الأسئلة المكون من 400 مفردة كافيًا، مع وجود 100 مفردة تتم مناوبة عرضها في وقت ما أثناء الاختبار لتحقيق التكافؤ بين نسخ الامتحان. يُمْكِن من خلال المحاكاة المقارنة بين بنوك الأسئلة التي تضم 500 مفردة وتلك التي تشتمل على 400 مفردة.

### *نقطة البدء*

هناك العديد من الخيارات المتاحة لتحديد قيمة (θ) لنقطة البدء الخاصة بكل ممتحن قبل عرض مفردة الاختبار الأولى. إن الطريقة المباشرة والأكثر بساطة للقيام بذلك هي تعيين قيمة ثابتة تكافئ الدرجة المتوسطة. مع نظرية الاستجابة للمفردة الاختبارية (IRT)، غالبًا ما تكون هذه القيمة هي (0.0) لأن المقياس يتمحور حول الممتحنين.

عند بدء الاختبار لجميع الممتحنين انطلاقًا من قيمة (θ) الأولية ذاتها، فإننا ندرك أن هناك عيبًا واضحًا لهذه الطريقة. فحيث أن خوارزمية الاختبار التكيفي المحوسب تقوم باختيار المفردة الأنسب للممتحن بناءً على قيمة (θ) التقديرية، فإذا كان لكل ممتحن نفس القيمة التقديرية الأولية، فسيُعرَض لكل الممتحنين نفس المفردة الأولى. إذا كان ذلك سيؤدي إلى مشكلة في أمان الاختبار أو مشكلة في كثرة عرض المفردة، فيُمْكِن تنفيذ هذه الخطوة عن طريق التوزيع العشوائي. فعلى سبيل المثال، يُمْكِن تحديد القيمة التقديرية بحيث تقع في النطاق من -0.5إلى +0.5 أو تطبيق طريقة عشوائية لاختيار المفردة، حيث أن هذه الحلول ستساعد في إتاحة العديد من المفردات كنقاط بداية.

وحيث أن الهدف من الاختبارات التكيفية المحوسبة هو ملائمة أو تكييف الاختبار حسب قدرة كل ممتحن، فيُلاحظ أن كلا الطريقتين السابقتين لتحديد نقاط البدء تفترضا عدم توافر أي معلومات مُسبَقَة عن الممتحن. إلا أنه في معظم الحالات تتوافر معلومات عن الممتحنين، كدرجات الممتحنين في الاختبارات السابقة. فإذا تم استخدام الاختبارات التكيفية المحوسبة في المدارس كجزء من برنامج التقييم التكويني، فسيتم استخدامهم مرات عدة على مدار العام الدراسي. وفي مِثل هذه الحالة، تُمَثِل درجة الطالب في الاختبار الأول نقطة بداية مثالية للاختبارات التالية، لأن مستوى الطالب سيقع على الأرجح ضمن نطاق مشابه، إلا أننا نأمل بالطبع أن يزداد مستوى الطالب.

يوجد خيار آخر لتحديد نقطة البدء وهو يتمثل في استخدام المعلومات الظاهرية كالدافعية والحالة الاجتماعية والاقتصادية لتقدير مستوى أو قدرة الممتحن،

(Castro, Suarez, & Chirinos, 2010). في السياق التعليمي، قد تكون نتائج التقييم لمختلف الجوانب الأخرى أو المعلومات المدرسية عن الطلاب مفيدة في هذا الصدد. فعلى سبيل المثال، في حالة الاختبارات المترتب عليها منح ترخيص مهني أو شهادة إتمام مرحلة تعليمية ما، يمكن استخدام مؤشرات الأداء كمتوسط الدرجات الدراسية كنقطة بداية إذا أظهر البحث والدراسة وجود علاقة ارتباط بينهما. وعلى الرغم من أن ذلك لا يقدم تنبؤًا دقيقًا عن مستوى كل مُمتحِن، إلا أن هذا من شأنه أن يزيد من كفاءة الاختبار، وهو ما يعني توفير الوقت والتحكم في عرض المفردات مما يعني توفير النفقات على المدى الطويل. بالنسبة للأقلية من الممتحنين الذين لا تعطي هذه الطريقة تنبؤًا دقيقًا عن مستواهم، فسيسهم الاختبار التكيفي المحوسب في سد الفجوة في هذا الجانب.

### خوارزمية اختيار المفردة

إن خوارزمية اختيار المفردة هامة للغاية ليس فقط لارتباطها بالإحصاءات الخاصة بتحديد المفردة الأنسب للممتحن، ولكن لدورها في التحكم في بعض القيود العملية. تُبنى خوارزمية اختيار المفردة على مفهوم ***معلومات المفردة***، والذي يهدف إلى التحديد الكمي لمدى مناسبة مفردات معينة دون غيرها لموقف معين. فمن غير المنطقي على سبيل المثال أن نقوم بعرض أو تقديم مفردة سهلة للغاية لممتحن متفوق، نكاد نجزم أنه سيجيب عليها بشكل صحيح. والعكس صحيح في حالة الممتحنين الأقل قدرة.

من الاعتبارات الهامة التي يجب مراعاتها عند بناء خوارزمية اختيار المفردة هي أن نحدد ما إذا كان الغرض من الاختبار هو الحصول على تقديرات دقيقة ($\theta$) لمستوى اللطلاب أم اتخاذ قرارات عامة. إذا كان الهدف من الاختبار هو تقدير قيمة ($\theta$) بمستوى دقة معين، فمن المناسب في هذه الحالة تقديم المفردات الاختبارية التي تقيس أكبر قدر من المعلومات عند القيمة التقديرية ($\theta$) الخاصة بالممتحن. وفي المقابل، إذا كان الغرض من الاختبار هو تصنيف الممتحنين على أساس درجة قطعية، باستخدام اختبار نسب الاحتمال (Reckase, 1983)، فإن تصميم خوارزمية اختبار المفردة على أساس تقييم المعلومات عند الدرجة القطعية يجعلها أكثر كفاءة وفعالية (Eggen, 1999; Eggen & Straetmans, 2000; Thompson, 2009).

توجد عدة طرق لحساب معيار المعلومة في نظرية الاستجابة للمفردة والذي يُستخدَم لاختيار المفردة، بالرغم من أن نسبة كبيرة من بحوث الاختبارات التكيفية المحوسبة هي عبارة عن دراسات محاكاة تم تصميمها لمقارنة الطرق المختلفة لاختيار المفردة (مثل Eggen, 1999; Weissman, 2004)، وبالرغم من أن مؤتمر الجمعية الدولية لتكنولوجيا الاختبار التكيفي المحوسب خصص جلستين كاملتين لتناول البحوث التي تم إجراءها فيما يتعلق بخوارزميات اختبار المفردة، إلا أننا نجد أن الفروق بين هذه الطرق ليست ذات دلالة؛ لذا، ذهب البعض إلى ضرورة تقييم الطرق الأخرى التي قد تؤثر في معيار اختيار المفردة لجعل الاختبار أكثر كفاءة (Thompson, 2009; van der Linden, 2010).

ولنفس السبب، غالبًا ما يكون من المهم تقييم تأثير المحدِدات أو القيود العملية التي تؤثر في عملية اختيار المفردة. أكثر أنواع هذه القيود شيوعًا هي القيود المتعلقة بعرض المفردة والقيود المتعلقة بخصائص المفردة. القيود المتعلقة بعرض المفردة هي عبارة عن خوارزميات فرعية مدمَجة في خوارزمية اختيار المفردة للتغلب على تكرار اختيار المفردات الأفضل التي يكون لها معامل تمييز عالي بصورة أكبر من غيرها. وبالتالي، فغالبا ما يتم عرض المفردات ذات معامل التمييز الأعلى أكثر من تلك التي لها معامل تمييز منخفض أو متوسط. لحل هذه المشكلة، يتم عادة استخدام التوزيع العشوائي. انظر Economides و Georgidou و Triantfillou (2007) لمراجعة هذه الطرق.

إضافة إلى ذلك، في كثير من برامج الاختبار قد تُفرَض مجموعة أخرى من القيود متمثلة في بعض الخصائص غير النفسية. فمثلًا، توجد القيود الخاصة بالمحتوى، كما هو الحال في اختبارات مادة الرياضيات التي تتطلب تغطية الاختبار لفروع الجبر والهندسة والاحتمال بنسب معينة. توجد أيضًا بعض القيود التي تتعلق بالجانب المعرفي، بما في ذلك تصنيف بلوم (1956) الذي يتطلب اشتمال الاختبار على عدد محدود من الأسئلة التي تقيس التذكر.

إن هذه القيود، من كلا النوعين، تؤدي إلى انخفاض كفاءة الخوارزمية التكيفية، لأنها ببساطة تعوق عملية الاختيار الطبيعية التي تعتمد على اختيار المفردات الأكثر تمييزًا، إلا أنها قد تكون هامة للغاية من منظور أوسع. وبالتالي، ينبغي أن تؤخذ هذه القيود بعين الاعتبار في المحاكاة المعتمِدة على البيانات الحقيقية أو المحاكاة المختلطة وذلك عند تحديد مواصفات الاختبار التكيفي المحوسب. إن المحاكاة لا تفيد فقط في تقييم المشاكل المتعلقة بتطبيق القيود الخاصة بعرض المفردات، ولكنها تفيد أيضًا في مقارنة فعالية الطرق المختلفة التي يمكن استخدامها للتحكم في عرض المفردة.

### خوارزمية التصحيح (تقدير $\theta$)

تستخدِم معظم الاختبارات التكيفية المحوسبة نظرية الاستجابة للمفردة (IRT) لأغراض تصحيح [حساب درجات] الاختبار واختيار المفردات التي تُعرَض على الممتحنين. بالرغم من أن دراسة Rudner (2002) أثبتت كفاءة الاختبارات التكيفية المحوسبة المبنية على نظرية الاختبار الكلاسيكية في تصنيف الممتحنين، إلا أن الاختبارات التكيفية التي تهدف إلى تقدير الدرجة التي تعبر بدقة عن قدرة الممتحن يجب أن تكون مبنية على نظرية الاستجابة للمفردة نظرًا للدقة التي توفرها هذه النظرية. يُمْكِن استخدام دراسات المحاكاة لمقارنة كفاءة الاختبارات التكيفية المحوسبة المُصَمَمة بخوارزميات تصحيح مختلفة. وبالطبع، لا تقتصر هذه الدراسات على مقارنة نظرية الاختبار الكلاسيكية بنظرية الاستجابة للمفردة، ولكن تتسع لتشمل مقارنة النماذج المختلفة لنظرية الاستجابة للمفردة كالمقارنة بين طريقة الاحتمال الأقصى وطريقة بيشان [النظرية الافتراضية]. قد تسفر نتائج مقارنة النماذج المختلفة لنظرية الاستجابة للمفردة عن اختلافات طفيفة في النتائج المرصودة، إلا أن هذه الاختلافات تكون لها دلالات هامة. تعطي طريقة الاحتمال الأقصى تقديرًا أقل تحيزًا (Lord, 1986) إلا أنها تتطلب وجود أنماط استجابة متنوعة (على الأقل استجابة واحدة صحيحة وأخرى غير صحيحة)، وهو

ما لا يتحقق أبدًا بعدما يتم عرض المفردة الأولى. في حال وجود متجه استجابات غير متنوع، يجب تطبيق خوارزمية فرعية. يُمْكِن أن تساعد المحاكاة أيضًا في تعديل المواصفات الخاصة بهذه الخطوة.

### *معيار الإنهاء*

بالرغم من أنه يُمْكِن تصميم الاختبارات التكيفية المحوسبة لتكون ثابتة الطول (كأن يُعرَض على كل ممتحن عدد 100 مفردة، ولكن يتم اختيار المفردات من بنك الأسئلة بشكل تكيفي تبعًا لمستوى الممتحن) إلا أنه يُمْكِن تصميمها بحيث تكون ***متباينة الطول*** أيضًا. إنه هذه الاختبارات تسهم ليس فقط في مطابقة مفردات الاختبار مع مستوى الممتحن، ولكن في عرض المفردات التي نحتاجها فقط. يمكن تنفيذ ذلك بطرق عدة. فبعض هذه الطرق تعتمد على تقدير درجة ($\theta$) الممتحن، وبعضها يعتمد على الخطأ المعياري في القياس، والبعض يعتمد على بنك الأسئلة.

فمثلًا عند الاعتماد على الدرجة التقديرية ($\theta$) للممتحن كمعيار لإنهاء الاختبار، يتم إنهاء الاختبار إذا لم تتغير الدرجة التقديرية ($\theta$) - سوى بالقدر الضئيل- بعد إجابة الطالب على كل مفردة. ذلك لأن الاختبار التكيفي المحوسب عبارة عن عملية تكرارية، حيث يكون هناك تباينًا كبيرًا في تقدير مستوى الممتحن في بداية الاختبار ولكن يثبت مستوى الممتحن عند نقطة معينة. ينطبق نفس الشيء على الخطأ المعياري للقياس، حيث يكون الخطأ المعياري كبيرًا في البداية، ويقل تدريجيًا مع متابعة الاختبار.

أيضًا، يُمْكِن الاعتماد على مفردات بنك الأسئلة كأساس لتحديد معيار إنهاء الاختبار بدلًا من الاعتماد على البارامترات الخاصة بالممتحن. وكمثال على ذلك، يُمْكِن استخدام ***معيار الحد الأدنى من المعلومات*** كمؤشر لتحقق شرط إنهاء الاختبار. فعلى سبيل المثال، إذا لم يتبقى ببنك الأسئلة مفردات يُمْكِن أن توفر على الأقل مستوى أدنى من المعلومات عن مستوى الطالب، كما هو محدد بخوارزمية اختيار المفردة، فيُمكن إنهاء الاختبار لعدم توافر مزيد من المفردات التي يُمْكِن عرضها.

ومع ذلك، نجد أن معيار الإنهاء المُعتَمِد على ***الخطأ المعياري الأدنى*** هو المعيار الأكثر شيوعًا في هذا الصدد. يعتمد هذا المنهج على تصميم الاختبار بحيث يتم إنهاءه عندما يصل الممتحن إلى مستوى خطأ معياري معين أو مستوى دقة مكافئ. فعلى سبيل المثال، يُمْكِن أن ينتهي الاختبار عندما يصبح الخطأ المعياري 0.25 أو أقل. وهذا يعني أن حد الثقة 95% مع ±2 خطأ معياري بالزيادة أو النقصان يجعل درجة الممتحن ($\theta$) أبعد وحدة واحدة عن الخطأ المعياري. يتميز معيار الإنهاء هذا بتوفير ***درجات مكافئة دقيقة*** لجميع الممتحنين، إذا سلمنا بأن بنك الأسئلة تم بناءه بشكل صحيح.

وكما هو الحال مع خوارزمية اختيار المفردة، تخضع هذه الخوارزمية أيضًا لبعض القيود العملية، والتي من أهمها القيود المتعلقة بطول الاختبار، سواء فيما يتعلق بالحد الأدنى أو الأقصى لطول الاختبار. إن تعيين حد أدنى من الأسئلة للاختبار يضمن تَعَرُض الممتحن لعدد مُحَدَد من مفردات الاختبار. فهذه الطريقة تضمن أن الممتحن الذي يرسب في الاختبار لعدم الإجابة الصحيحة على 10 مفردات مثلًا، يكون قد تعرض لمفردات اختبارية لا تقل عن 20 مفردة قبل أن يرسب، وهذا من شأنه أن يقلل شكاوى الطلاب. كما أن تعيين حد أقصى من الأسئلة يضمن عدم عرض جميع مفردات الأسئلة التي يتضمنها البنك. في الاختبار التكيفي المحوسب الذي يقوم على أساس النجاح/الرسوب، لن يكون من الممكن تصنيف الممتحنين الذين تكون درجتهم التقديرية ($\theta$) مساوية للدرجة القطعية؛ وبالتالي، يجب ضبط الاختبار بحيث يتم إنهاءه بعد عرض عدد كبير نسبيًا من مفردات الاختبار، 200 مفردة مثلًا.

إن كل هذه الخيارات تتيح إمكانية التحكم المباشر في إجراء الاختبارات التكيفية وفي عدد المفردات التي تُعرَض على الممتحنين. بصفة عامة، كلما زادت عدد مفردات الاختبار، كلما كانت نتائج الاختبار أدق والعكس صحيح. تُعَد دراسات المحاكاة ضرورية للتحديد الدقيق لمواصفات الاختبار بما تحقق متطلبات برنامج التقييم. إذا تم استخدام معيار الخطأ القياسي الأدنى كطريقة لتحديد معيار إنهاء الاختبار، فسيكون من المفيد إجراء دراسات محاكاة عند مستويات خطأ معياري مختلفة، عند مستوى خطأ معياري 0.25 و 0.30 و 0.35، ثم تقييم العدد الأكبر من المفردات المطلوبة لمزيد من الدقة.

## الخطوة 5: النشر الفعلي الاختبارات التكيفية المحوسبة

بعدما يتم الانتهاء من تحديد المواصفات لجميع مكونات الاختبار وجميع الخوارزميات الإضافية، يُمْكِن حينها نشر الاختبار التكيفي المحوسب. إذا كانت المؤسسة تمتلك برمجية لتطوير وإجراء الاختبارات (كأن تكون المؤسسة قد قامت بشراء منصة أو تمتلك إمكانية الوصول إلى منصة ما) فإن هذه الخطوة لن تكون صعبة مطلقًا. إن معظم الخيارات الموضحة في الجزء السابق تظهر على هيئة أزرار اختيار داخل نظام الاختبار التكيفي المحوسب؛ لذا، إذا كانت المؤسسة بصدد تطوير منصة خاصة بها فستكون هذه هي الخطوة الأكثر صعوبة. حتى إذا كان هذا هو الحال مع مؤسستك، فلحُسن الحظ يمكن القيام بتطوير منصتك الخاصة بالتزامن مع الخطوات الأربع السابقة مما يوفر الكثير من الوقت. إن هذه الخطوة تتضمن الكثير من المشاكل العملية المتعلقة بإجراء الاختبار وتعميم النظام ولكن جميع هذه المشاكل ترتبط بعملية التقييم بشكل عام وليست مقتصرة على الاختبارات التكيفية المحوسبة فقط.

## الخاتمة: صيانة نظام الاختبار التكيفي المحوسب

إن عمليات البحث المتضَمَنة في عملية تطوير الاختبار التكيفي المحوسب (CAT) لا تتوقف بنشر الاختبار. فعملية صيانة نظام الاختبار التكيفي المحوسب تحتاج إلى مزيد من البحث. ربما النقطة الأهم هنا هي التحقق مما إذا كانت النتائج الفعلية للاختبارات التكيفية المحوسبة بعد النشر تتطابق مع النتائج المتوقعة المبنية على دراسات المحاكاة. فعلى سبيل المثال، إذا كانت المحاكاة المستندة على بيانات حقيقية تتنبأ بأننا نحتاج إلى عرض 47 مفردة اختبارية على الممتحن حتى نصل إلى الخطأ المعياري الأدنى الذي مقداره 0.25، فهل تحقق ذلك فعليًا أثناء الشهر الأول

من تطبيق الاختبار التكيفي المحوسب؟

ثمة مسألة أخرى هامة هي صيانة بنك الأسئلة، والتي يطلق عليها أحيانًا "تحديث" بنك الأسئلة. حيث أن بعض مفردات الاختبار قد يكون تكرار عرضها عالي جدًا وخاصة في الاختبارات التي تُجرى على نطاق كبير، قد نحتاج إلى استبعاد هذه المفردات من مجموعة المفردات النشطة وإضافة مفردات جديدة بدلًا منها. ويتم ذلك عادةً عن طريق تغذية بنك الأسئلة بمفردات اختبارية جديدة ليتم اختبارها قبليًا، ثم وضعها مع المفردات النشطة بعدما يتم اختبارها على عينة كافية والحصول على البارامترات الخاصة بها. توجد بعض الأبحاث التي استقصت تطبيق المعايرة الإلكترونية للمفردات الجديدة، حيث تتم المعايرة الفورية لهذه المفردات ببنك الأسئلة أثناء عملية الاختبار القبلي.

إن تحديد مفردات الاختبار التي يجب عزلها هو خيار تحدده الجهة المعنية، إلا أنه توجد العديد من النقاط التي يجب أخذها بعين الاعتبار في هذا السياق. المسألة الأهم هنا تتعلق بتكرار عرض المفردات. فإذا تم عرض مفردة بعينها على نصف الممتحنين، فمن المحتمل أن تتسرب هذه المفردة. إحدى الطرق التي يمكن من خلالها معالجة هذه المشكلة هي قياس الانحراف لبارامترات المفردة. إذا تم تسريب مفردة ما، فستزيد نسبة الطلاب الذين يجيبون عليها بشكل صحيح مقارنة بوقت نشرها لأول مرة. عند تحليل البيانات، فإن البارامترات الخاصة بالمفردة ستكون مختلفة، مما يعني الحاجة إلى عزل المفردة. قد يكون من المفيد في هذا الصدد استخدام أحد برامج حماية الاختبار المُصَمَمة للقيام بالبحث عن مفردات الاختبار على مواقع معينة على الانترنت.

## ملخص

إن تطوير الاختبارات التكيفية المحوسبة (CAT) يتطلب توافر خبرة كبيرة في مجال القياس النفسي. لذا، غالبًا ما تترك عملية تطوير الاختبارات التكيفية المحوسبة للمتخصصين العاملين في هذا المجال. ولكن مع ازدياد انتشار الاختبارات التكيفية المحوسبة، قد لا تكون خبرة القائمين على العمل في مجال القياس النفسي غير كافية لتطوير اختبارات تكيفية محوسبة قوية دون الحصول على بعض التوجيه والإرشاد. عرضت هذه الورقة البحثية إطارًا عامًا لإعداد الاختبارات التكيفية المحوسبة، والذي يُمْكِن أن ينطبق على معظم الحالات. وبالرغم من أن هذا الإطار عام وبالرغم من تناوله العديد من النقاط، إلا أنه ليس شاملًا. فينبغي أن نتذكر أن كل برنامج تقييم له مشاكله الخاصة التي لا يجب تحديدها فحسب ولكن يجب أيضًا عزلها وإخضاعها للبحث التجريبي قدر الإمكان. إلا أنه قبل القيام بالبحث التجريبي لحل هذه المشاكل، يجب تحديد الأسس التي تضع إجابات لهذه المشكلات.

## المراجع

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283-296.

Bejar, I.I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12,* 377-379.

Bloom B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co. Inc.

Bock, R., Gibbons, R., Schilling, S., Muraki, E. Wilson, & Wood, R., (2003) TESTFACT 4 (Computer software). Lincolnwood, IL: Scientific Software International.

Castro, F., Suarez, J., & Chirinos, R. (2010). *Competence´s initial estimation in computer adaptive testing.* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Choi, S.W. (2009). Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT Models (Computer software). *Applied Psychological Measurement, 33,* 644–645.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-261.

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713-734.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A Primer*. Mahwah, NJ: Erlbaum.

Frick, T. (*1992*). Computerized Adaptive Mastery Tests as Expert Systems. *Journal of Educational Computing Research, 8(2),* 187-213.

Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved August 23, 2010 from http://www.jtla.org .

Gibbons, R.D., Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A., Grochocinski, V.J., Bhaumik, D.K., Stover, A., Bock, R.D., Immekus, J.C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Kolen, M.J., & Brennan, R.L.. (2004) *Test equating, scaling, and linking. Methods and practices,* 2nd ed. New York: Springer.

Lee, J, & Weiss, D. J. (2010). *Selection of common items in full metric calibration for the development of CAT item banks*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved August 23, 2010 from www.psych.umn.edu/psylabs/CATCentral.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2006). *Practical considerations in computer-based testing.* New York: Springer.

Pommerich, M., Segall, D.O., & Moreno, K.E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved August 23, 2010 from http://www.psych.umn.edu/psylabs/CATCentral .

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests.* Paper presented at annual meeting of the National Council on Measurement in Education, Chicago, IL.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures.* Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

Rudner, L.M. and Guo, F.M. (in press) Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology.*

Sands, W.A., Waters, B.K. and McBride, J.R. (Eds.) (1997). *Computerized adaptive testing. From inquiry to operation.* Washington: American Psychological Association.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, 12(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1 .

Thompson, N.A. (2009a). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793.

van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*, (Statistics for Social and Behavioral Sciences Series). New York: Springer.

van der Linden, W.J. (2010). *How to make adaptive testing more efficient?* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Veldkamp, B.P., & van der Linden, W.J. (2010). Designing item pools for adaptive testing. In van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*, (Statistics for Social and Behavioral Sciences Series). New York: Springer.

Verschoor, A. (2010). *Optimal calibration designs for computerized adaptive testing*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Vispoel, W.P., Rocklin, T.R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53-59.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15-20.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A Primer*. Mahwah, NJ: Erlbaum.

Waller, N. (1997). *MicroFACT* (Computer software). Saint Paul, MN: Assessment Systems Corporation.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weiss, D. J. & Guyer, R. (2010). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing*. St. Paul MN: Assessment Systems Corporation.

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

Wise, S. G., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologia, 21*(1), 135-155. http://redalyc.uaemex.mx/pdf/169/16921108.pdf

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation.

Yoes, M. (1997). *PARDSIM* (Computer software). Saint Paul, MN: Assessment Systems Corporation.

**توثيق المستند عند الاقتباس:**

Thompson, Nathan A., & Weiss, David A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

**المؤلف (المساهم الرئيسي):**

Nathan A. Thompson, Vice President
Assessment Systems Corporation
2233 University Ave., Suite 200
St. Paul, MN. 55114

Phone: +1 (651) 647-9220
Email: nthompson [at] assess.com