الفصل الخامس

تقويم اختبارات اللغة المحوسبة
Evaluating CALT

إذا كان اختبار اللغة المحوسب مختلفًا عن صور الاختبارات الأخرى، فهل ينبغي تقويمه استنادًا إلى مجموعة مختلفة عن تلك المعايير المستخدمة في تقويم الاختبارات الأخرى؟ استفاض الفصل الثالث في المخاوف المحتملة التي قد تحد من توافر عنصر الصدق في النتائج وفي استخدام اختبارات اللغة المحوسبة. ولا يخفى أن هذه المخاوف قد شغلت بال الجميع، غير أن هذا الفصل يتناول القضايا الفنية تناولاً مركزاً؛ بحثًا عن الكيفية المناسبة والمتكاملة لتقويم اختبارات اللغة المحوسبة. لا شك في أن تقويم عنصر الجودة في أي اختبار، ينبغي ألاّ ينحصر كليةً في جوانبه التي تستقطب انتباه المطّلعين من غير المتخصصين. ومن ثمّ فإن هذا الفصل يتناول تقويم اختبارات اللغة المحوسبة في ضوء مقترحات وآراء بعينها صدرت عن معدي هذه الاختبارات، علمًا بأن هذه المقترحات على وجه الخصوص وتقويم هذا النوع من الاختبارات على وجه العموم سيُبحثان بالنظر في وجهات نظر المحترفين إزاء الكيفية المناسبة لتقويم جميع اختبارات اللغة. وسنستهل الحديث بنظرة في الإرشادات والنصائح المطروحة للارتقاء بمستوى الجودة عند إعداد اختبارات اللغة المحوسبة وتقويمها، متبوعةً بملخص بالبحوث التي عنيت بجوانب محددة من الجودة في تلكم الاختبارات. بعد ذلك سنعرّج على قسم يوضح أنواع النتائج المترتبة على الأخذ بما يُعرف باسم "إطار فعالية من الاختبار" حسب طرح كل من باخمان وبالمر (١٩٩٦م) (Bachman and Palmer's (1996) test usefulness framework). أما القسم الأخير من الفصل، فيتناول تقويم اختبارات اللغة المحوسبة من وجهات النظر الحديثة في القياس

التربوي (Kane, 2001) ؛ وذلك بغية الوصول إلى برهان جديد على الصدق من خلال استنتاجات تُوَظّف في بيان الأساس السليم لاستقراء درجات واستخدامها. ونرى أن القضايا المحددة ذات الصلة بالتقنية، إنما ينبغي أن توضع ضمن إطار عملي أوسع لتقويم الاختبارات.

الإرشادات والنصائح

منذ بداية مرحلة التأسيس للاختبارات المحوسبة، في سبعينيات القرن المنصرم، تصدى الباحثون في القياس التربوي لمجموعة من القضايا الفريدة التي فرضها هذا الوسَط الجديد لتقديم الاختبارات. وقد شهد عام ١٩٨٤م نشر موضوع في "دورية القياس التربوي" (Journal of Educational Measurement) بعنوان "الإرشادات التقنية لتقويم الاختبارات الحاسوبية المتكيفة" (Technical guidelines for assessing computer-adaptive tests)، وتولى حينها كتابته كل من قرين، وبوك، وهمفريز، ولين، وريكيس. وبعد ذلك بخمس سنوات، نشرت الجمعية الأمريكية للصحة النفسية (American Psychological Association) موضوعًا بعنوان "إرشادات بشأن الاختبارات المعتمدة على الحاسوب وتفسير نتائجها" (Guidelines for Computer-Based Tests and Interpretations). غير أن هذه المنشورات إلى جانب ما نشره قرين (Green, 1988) تحت عنوان "صدق البناء في الاختبارات المعتمدة على الحاسوب" (Construct validity of computer-based tests) في مجلد استهلالي بهذه الشأن يحمل اسم "صدق الاختبار" (Test Validity)، وما نشره بيكر (Baker, 1989) تحت عنوان "تقنية الحاسوب في بناء الاختبارات ومعالجتها" (Computer technology in test construction and processing) في الإصدار الثالث المعتمد من مطبوعة "القياس التربوي" (Educational Measurement) قد تنبأت بالاتجاه السائد في أوساط المتخصصين في القياس التربوي: ألا وهو ضرورة التعامل مع تقويم الاختبارات المحوسبة بأسلوب مختلف، أو وفقًا لمعايير مختلفة، عن تلكم الأساليب المتبعة في تقويم الاختبارات الأخرى.

وقد تجلى هذه المنظور للقياس التربوي في التقويم اللغوي إلى حد ما. وفي هذا الإطار اقترح نويجونز (Noijons, 1994) مجموعة من العوامل المرتبطة بالاختبارات والواجب مراعاتها في تقويم اختبارات اللغة المحوسبة. علمًا بأن هذه العوامل بدت كأنها إضافة إلى متطلبات الصدق والثبات التي تناولها نويجونز في قسم منفصل من بحثه الذي استند فيه على معايير تقويم للاختبارات المحوسبة اقترحها كل من أليسي وترولوب (Alessi and Trollop, 1985)، وهما مؤلفان لكتاب قيم جداً يعنى بالتعليم بمساعدة الحاسوب. ويميز هذا التصور بين نوعين من العوامل: يتعلق أولهما بمحتوى الاختبار، ويتعلق ثانيهما بأداء الاختبار كما هو مقدم بإيجاز في الجدول رقم (٥,١).

تتجلى بعض الخصائص الفريدة لاختبارات اللغة المحوسبة في نقاط تناولها نويجونز؛ منها على سبيل المثال، توفير معلومات عن المساعدة المتاحة عبر الإنترنت للطلاب، والتنبيه الفوري عند اختيار وظائف غير مناسبة، والتنبيه إلى موعد انتهاء الوقت المخصص للاختبار، وكذا إنهاء الاختبار دون فقدان استجابات الطالب في حال تعطل المعدات أو البرمجيات، بالإضافة إلى الاسترجاع الآمن للبيانات. ذلك أن هذه الجوانب المرتبطة بتصميم واجهة الاستخدام، تضيف عوامل جديدة إلى العوامل المأخوذ بها فعلا من المقيّمين لدى تحديد العوامل التي قد يكون لها تأثير سلبي في أداء الاختبار (ومن ثمّ في نتائجه) إذا لم تؤخذ على محمل الجد في الدراسة والاهتمام. وثمة عوامل أخرى كالتعقيب الذي يتلقاه الطلاب – سواء أكان ذلك مع تسريع مهام الاختبار أم لا – وكيفية عرض النتائج لهم، وهي قضايا موجودة في كل اختبارات اللغة، لكن عدد الخيارات المتاحة في اختبارات اللغة المحوسبة لا حد له. فعلى سبيل المثال، تتألف خيارات التعقيب المتوفرة للطلاب في الاختبارات الورقية من تقديم معلومات بشأن صحة الاستجابات من خلال النتائج الإجمالية والجزئية بعد استكمال الاختبار. أما في الاختبارات المحوسبة، فقد بحثنا عددًا من الخيارات في الفصول السابقة حول كيفية تقديم التعقيب في أثناء أداء الاختبار، وكذلك تقديم تعقيب مفصل وفوري عند نهايته. وينبغي

على معد الاختبار تحديد كيفية تضمين التعقيب تضمينًا واضحًا في عملية أدائه، كما على المقوم أن يضع مدى مناسبة الخيارات المتاحة في الحسبان.

الجدول رقم (٥,١). ملخص نويجونز لنقاط تقويم اختبارات اللغة المحوسبة (١٩٩٤م).

التوقيت	محتوى الاختبار	خوض الاختبار
قبل	وظيفة/غرض الاختبار: ما الهدف التربوي؟ (الإتقان – تحديد المستوى – التصنيف – الانتقاء)	دخول الاختبار: ما البيانات المسجلة؟ (هوية المرشح – رقم الاختبار – استجابات الطالب)
	هدف الاختبار: ما الذي يسعى الاختبار إلى تقويمه؟	تعليمات الاختبار: ماذا يقال للمختبر؟ (إجراءات أداء الاختبار – المكون محل الاختبار – التعليمات المتاحة عبر الإنترنت)
	طول الاختبار: كم عدد بنود الاختبار؟	أمثلة للبنود: ما أمثلة البنود المقدمة المتاحة؟ هل للطلاب فرصة تجريب كيفية الاختبار وتأكيد فهمهم لكيفية الاستجابة؟
	انتقاء العناصر: كيف تُنتقى العناصر من بنك العناصر؟	فحص الاختبار: هل خضع الاختبار لفحص شامل على يد معديه وغيرهم من الخبراء؟
أثناء	نوع العنصر: ما الشكل الموضوع لتقديم الاستجابات؟	الغش: هل ثمة إمكانية لإنهاء الاختبار دون فقدان إجابات الطالب في حال اكتشاف تورطه في الغش؟
		الأعطال: هل هناك إجراءات مناسبة في حال تعطل المعدات أو حدوث أي خلل في البرامج؟ (حفظ البيانات – إصدار تعليمات للطلاب)
	التعقيب: ما الرسائل التي يتلقاها الطلاب بشأن إجراءات أداء الاختبار أو بشأن محتوى أدائهم؟	التعقيب: هل يصدر تعقيب للطلاب عندما يؤدون وظائف غير مسموح بها؟
	الزمن: هل الاختبار اختبار سرعة؟	نهاية الاختبار: هل يُطلَع الطالب على الزمن المتبقي إذا كان الاختبار مقيدًا بزمن؟ هل توجد تعليمات بشأن ما ينبغي فعله في نهاية الاختبار؟
	تسجيل الاستجابات: كيف يتم تحصيل الإجابات وتخزينها؟	تخزين البيانات: هل البيانات محمية على نحو يحول دون وصول الطلاب إليها؟
بعد	التقويم: كيف يصنع القرار في ضوء نتائج الاختبار؟	طباعة البيانات: هل تُطبَع البيانات كإجراء لدعم المعلومات؟
	إعلان نتائج الاختبار: كيف تُقدّم النتائج للطلاب؟	

يندرج لكثير من المخاوف الفريدة المرتبطة باختبارات اللغة المحوسبة دون غيره، حسبما ورد في تحليل نويجونز، تحت قضايا واجهة المستخدم حسب وصف فولتشر (Fulcher, 2003) الذي يتبنى منظور معد الاختبار؛ وبذلك يرى ضرورة مراعاة هذه القضايا أثناء إعداد اختبارات اللغة المحوسبة. كما يقسم القضايا إلى ثلاث مراحل: ١- التخطيط والتصميم المبدئي، ٢- اختبار قابلية الاستخدام، ٣- التجربة العملية والضبط. ويقدم الجدول رقم (٥,٢) ملخصًا للاعتبارات التي ينبغي أن يأخذها معدو الاختبار في الحسبان في المرحلة الأولى، ويدخل في ذلك التوجيه بشأن العوامل التي يراها معدو الاختبارات — لا سيما المبتدئون منهم — أنها عوامل شديدة مربكة صعبة. وكما أسلفنا في الفصل الرابع، عادة ما يعمل معدو الاختبارات مع أنظمة التوليف الاختيارية التي تتيح للمؤلف تحديد بعض هذه العوامل (كاختيار أنواع البنود)، غير أنهم يقدمون نماذج محددة سلفًا للآخرين (الكلمات والرموز المحجوزة للتحكم في واجهة المستخدم).

ومع أن تحليل فولتشر لا يدخل في عداد أوجه التقويم، فإن التوجيه المقدم للمصممين يتضمن نقاطًا يدور التقويم في فلكها على ما يبدو. ولا شك في أن الحافز التقويمي حاضر بقوة في الأسلوب الذي يصف فولتشر به الغرض من تحديد إرشاداته ممثلة في العمل على إنشاء "واجهة مستخدم حاسوبية لا تتدخل في التقويم". ومن الضروري في هذا الصدد، أن يكون أسلوب تقديم الاختبار "بريئًا من عيوب التلاعب بالنتائج وإلا ضاع الاستقراء الصحيح للاستجابة" (ص ٣٨٧). ويتمثل الهدف من ذلك في أن تكون واجهة المستخدم غير مرئية؛ لأن أية إسهام منها في نتائج الاختبار من شأنها التأثير سلبًا في استقراء نتائج الاختبار باعتباره دلالة تقويمية للقدرة اللغوية. وكما أسلفنا في الفصل السابق، يتوقف مدى كون الواجهة غير مرئية في أحد الاختبارات

على مدى وجود – أو انتفاء – المخوف الناجم عنها ممثلاً في كونها مصدرًا غير مقصود للتفاوت في الأداء، الأمر الذي يتوقف على المكوّن محل القياس.

ويلخص الجدول رقم (٥,٣) أنشطة المرحلتين الثانية والثالثة من تحليل فولتشر اللتين تتألفان من عمليات عادةً ما توجد عند إعداد الاختبارات المقدمة على نطاق واسع – خاصة عند كتابة البنود وحفظها في البنوك – إضافة إلى الكثير من العمليات التي تنفرد بها اختبارات اللغة المحوسبة، ومن ذلك البحث عن المشكلات والحلول في واجهة المستخدم، والتحقق من فعالية ذلك في كل المواقع والبرامج ذات الصلة، بالإضافة إلى البحث عن مصادر التفاوت ذات الصلة بواجهة المستخدم وعملها.

الجدول رقم (٥,٢). اعتبارات المرحلة الأولى من العملية لدى تصميم واجهة مستخدم لاختبار القدرات اللغوية عبر الحاسوب. (فولتشر Fulcher, 2003).

فئة قضية التصميم	الاعتبارات والعمليات العامة	نقاط الاهتمام والنصائح المقدمة للمصممين
١- تصميم نماذج أولية	اعتبارات الأجهزة	مواصفات الحاسوب، ودقة الشاشة، وزمن التنزيل، وقدرات برنامج المتصفح، وبرامج الأطراف الثالثة، وبرامج التوليف.
٢- تصميم جيد التنقل لواجهة المستخدم	اعتبارات البرامج	أزرار التنقل وأيقوناته، وأزرار نظام التشغيل وتعليماته، وسهولة التنقل وسرعته، ووضوح عناوين الصفحات، ووضوح القرارات بشأن إعادة زيارة الصفحات أثناء خوض الاختبار، وإجراء الحماية للحد قدر الإمكان من سوء استخدام خاصية التنقل.
	المصطلحات	تخصيص كلمات لتدفق عمليات واجهة المستخدم.
	تصميم الصفحة	تجنب تزاحم النصوص، والحد قدر الإمكان من التنقل، وتجنب الكتابة بالحروف الكبيرة.

تابع الجدول رقم (٥,٢).

فئة قضية التصميم	الاعتبارات والعمليات العامة	نقاط الاهتمام والنصائح المقدمة للمصممين
	النص	تجنب الرسوم المعقدة والمشتتة للانتباه، واستخدام خط بحجم يزيد على ١٠، واتخاذ قرارات واضحة بشأن صلاحية الطلاب في تعديل حجم الخط من عدمه، واستخدام خط معروف، وتجنب الخلط بين الخطوط.
	لون النص	زيادة التباين بين الألوان إذا كان لذلك أهمية (كأن يخصص لون للتعليمات وآخر لقراءة النص).
	أشرطة الأدوات وعناصر التحكم	عرض أقل عدد ممكن من النقاط وتقديم الأهم منها على ما سواه.
	الأيقونات والرسوم	الحد قدر الإمكان من عدد الأيقونات المستخدمة، والوصول إلى أفضل حجم للمعروض المرئي وزمن التنزيل، وتجنب الصور المتحركة والوامضة، وعرض الأيقونات بهدف إعلام الطالب بالجديد في حالات التأخير.
	تسهيلات التعليمات	اتخاذ قرارات واضحة بشأن إتاحة التعليمات.
	خارج الاختبار	مراعاة استخدام مصادر خارجية لتقويم مكوّن بعينه.
	أنواع البنود	الوصول إلى الوضع الأمثل فيما يتعلق بعدد الأنواع المطلوبة في ضوء المكوّن وقدرة الطالب على تغيير الأنواع.
	الوسائط المتعددة	ضمان تقديم الاختبارات على نحو متناسق واستغلال وقت التنزيل استغلالاً كاملاً.
	نماذج الكتابة والمهام ذات الإجابات قصيرة التعقيب.	إتاحة مساحة مناسبة للاستجابة، ومحاذاة المربعات النصية الكثيرة وضبطها، وترتيب المربعات الكثيرة ترتيبًا رأسيًا ضرورة التخطيط للوقوف على نوع التعقيب ووقته وموضعه.

تابع الجدول رقم (٥,٢).

فئة قضية التصميم	الاعتبارات والعمليات العامة	نقاط الاهتمام والنصائح المقدمة للمصممين
٣- الأنشطة المتزامنة بالمرحلة الأولى	استحداث نظم لتقديم الاختبارات، والتحقق من خطوات استعادة النتائج والدرجات والتخزين في قاعدة البيانات، وجوانب التوزيع والاستعادة فيما يتصل بالأقسام التي صححها المصححون من البشر، ولوغاريتمات تسجيل الدرجات وقواعدها، إلى جانب دراسات الشيوع والدراسات التقنية ودراسات التحقق من الصدق البنائي، وإمكانية التجربة على نطاق محدود.	(التوجيه المحدد في المجالات العامة خارج نطاق دراسة فولتشر).

الجدول (٥,٣). اعتبارات المرحلتين الثانية والثالثة من العملية في إطار تصميم واجهة مستخدم للاختبارات اللغوية المحوسبة. (عن دراسة فولتشر بإيجاز، Fulcher, 2003).

المرحلة	فئة قضية التصميم
٢- اختبار إمكانية الاستخدام	١- تقصي المشكلات والحلول. ٢- انتقاء الطلاب لدراسة إمكانية الاستخدام. ٣- الأنشطة المتزامنة: كتابة البنود ووضعها في البنك، والاختبار القبلي، وقواعد تسجيل نتائج التجارب، ودراسة المكونات التي تتألف منها بنية الاختبارات.
٣- الاختبار الميداني والضبط النهائي	١- التحقق من عمل واجهة المستخدم في كل المواقع وعلى كل البرامج. ٢- البحث عن مصادر التباين المتعلقة بواجهة المستخدم. ٣- الأنشطة المتزامنة: استحداث مواد توجيهية تتمثل بإنتاج اختبارات تدريبية/تجريبية، والاستقرار على قواعد تدريب المقيّمين وإجراء هذه التدريبات، والأخذ بدراسات التدرج والإعلان عن النتائج، والتخطيط لإجراء مزيد من دراسات الصدق.

على شاكلة أنشطة المرحلة الأولى، تُدرج هذه الأنشطة في قائمة العمليات الموصى بها مع مراعاة أن تقويم الاختبارات المُستَكملة سينطوي على تقويم مدى الالتزام بهذه الخطوات.

وثمة اتجاه أكثر عمومية إزاء التقويم يتمثل في تحديد الخصائص المطلوبة في اختبارات اللغة المحوسبة، وذلك استنادًا إلى الإمكانات التي تتيحها التقنية، والاحتياجات المنظورة في اختبارات اللغة. وبالحديث عن الباحثين، تناولت مونير (Meunier, 1994) هذا الاتجاه في دراسة عدّدت فيها المزايا والمساوئ المرتبطة بالاختبارات الحاسوبية المتكيفة، وبعد تحديد ما أسمته "الشكوك التي تحيط بعنصر صدق البناء" بكونه أحد المساوئ، ذكرت خمس خواص ذات صلة بهذا العنصر في إجراءات تقويم الصدق في هذا النوع من الاختبارات، وبيانها كالتالي: صدق المحتوى، والصدق التلازمي، والصدق التنبؤي، وصدق البناء، والصدق الظاهري، إضافة إلى عنصر التحيّز في الاختبارات. وتناولت في مناقشاتها طرائق البحث المعنية باستكشاف هذه الخواص إلى جانب البحوث التي أُجريت بالفعل على هذا النوع من الاختبارات. وبذلك نجد أن ما انتهت إليه الباحثة في مناقشة جوانب تقويمية يفيد "باستبدال نماذج الاختيار من متعدد ونماذج اختبارات التتمة في القراءة، والأخذ بإمكانات اختبارات اللغة المحوسبة لإنشاء صور محاكية لإجراءات (حية) مباشرة؛ وذلك (من أجل) تقويم الطلاب من حيث قدرتهم على الأداء في مواقف ذات مستويات مختلفة من الصعوبة" (Meunier, 1994، ص ٣٧). وبالمثل، يمضي تحليل لورير (Laurier, 2000) في الاتجاه ذاته: بمعنى أن إمكانات الحاسوب قد يسّرت إعداد اختبارات ذات سمات إيجابية، مثل: الموثوقية، علمًا بأنه يولي اهتمامًا بالغًا لهذه السمة. وبناء على ذلك، تركز ورقتا البحث المذكورتان على الجوانب المثالية، وما يمكن النهوض به في هذا الصدد من أجل مواجهة التحديات المستقبلية التي تعترض طريق معدي الاختبارات ومطوري البرامج.

وعلى الرغم من أن التوجيهات المفصّلة والتوجيه العام يشيران إلى خصائص حظيت – أو ينبغي أن تحظى – باهتمام خاص على يد المطورين، ومعدي الاختبارات، ومستخدمي اختبارات اللغة المحوسبة، فإن كلا النوعين من التوجيهات لا يقدم وصفًا للوسائل التجريبية المعنية بالتحقق من جوانب هذا النوع من الاختبارات، بل لم ينطويا على توجيه الباحثين إلى المسائل التجريبية ذات الصلة. غير أن هذه المسائل وتلك الوسائل بادية في بعض البحوث التي استكشفت أنماطًا بعينها من الاختبارات المحوسبة، وهذا ما نحن بصدد الحديث عنه.

البحوث التقويمية في اختبارات اللغة المحوسبة

تناولنا في الفصل الثالث بعض البحوث التي عنيت بالمخاوف التي تؤثر في الصدق. ويركز هذا الاتجاه في اختبارات اللغة المحوسبة على الأخطاء المحتملة في النتائج والدرجات المستقاة من هذه الاختبارات. ومع تفاوت مباعث القلق الواردة في الفصل الثالث، فمن الثابت أن أكثر الجوانب المرتبطة باختبارات اللغة المحوسبة وتحظى بجهود البحث، إنما تتمثل في التحقق من عنصر الجودة والفعالية في الاختبارات الحاسوبية المتكيفة. وثمة جوانب أخرى (مثل: جودة تصحيح الاستجابة حاسوبياً، والتأثير الناجم عن إجراء اختبارات اللغة المحوسبة في أداء المتعلمين وجهودهم) فقد تناولتها البحوث أيضاً، ولكن على نحو أقل من ذلك الجانب سالف الذكر.

الجودة والفعالية

تتناول الدراسات المعنية بإعداد الاختبارات الحاسوبية المتكيفة وتقويمها قضايا أفرزتها الحاجة إلى بنود مستقلة (انظر على سبيل المثال Kaya-Carton, Carton & Dandonoli, 1991) وأوجه الارتباط بالاختبارات الأخرى (انظر على سبيل المثال Stevenson & Gross, 1991)، وتتناسب استجابات الطلاب مع أنموذج نظرية الاستجابة للبند. غير أن توحيد موضوع

الدراسات المعنية بهذه الاختبارات، يعكس ما يشير إليه تشالو – ديفيل بكونه "أهم أسباب الانجذاب إلى الاختبارات الحاسوبية المتكيفة على ما يبدو"؛ بمعنى إمكانية إدارة كم كبير من الاختبارات على نحو يحقق التصور الأمثل "في أوضاع الاختبارات من خلال استهداف مستوى قدرة كل طالب على حدة" (١٩٩٩م، ص ٩). ومن ثمّ نجد أن أهم البيانات في مثل هذه السياقات تتمثل في تسجيلات المقدار الفعلي من الوقت الذي يستغرقه كل مختبر حال أدائه الاختبار، والأنموذج التنبئي النفسي المستخدم في تحصيل النتائج، ومدى موثوقية النتائج المستقاة من الاختبار. وفي هذا الإطار، نرى أن بحوث تشالو – ديفيل (١٩٩٩م) المصنفة في الإصدار المنقح لها، تتناول مسألة إعداد الاختبار (انظر على سبيل المثال: Dunkel, 1999; Laurier, 1999) عارضةً أمثلة للقضايا المعقدة التي تنجم عن العمل أثناء تنفيذ القرارات التي يتحتم على معدي الاختبارات اتخاذها لدى تحديد الغرض من الاختبار، ومحتواه، ونماذجه السيكومترية، إضافة إلى البرامج والأجهزة التي سيُقدم الاختبار من خلالها. ثمة بحوث أخرى تناولت قضايا من قبيل اختيار الأنموذج السيكومتري (Blais & Laurier, 1993) ومناسبة الأنموذج للطلاب من ذوي اللغات الأم المختلفة (Brown & Iwashita, 1996) وغير ذلك من القضايا التي تمس الجانب العملي من الاختبار (Burston & Monville-Burston, 1995; Young, Shermis, Brutten.& Perkins, 1996).

تقويم الاستجابات اللغوية

تناول الفصل الثالث التقويم الآلي للاستجابات اللغوية من حيث الحد من المخاوف التي تحيط بصدق البناء، غير أن الأعمال السابقة في هذا المجال، تشير إلى بعض القضايا الأخرى، إذ طرح كل من مولهولت وبريسل (Molholt and Presler, 1986) دراسة استكشافية تناولت التقدير الآلي للاستجابات الشفهية في الاختبارات، مدفوعين بدافعين لاستكشاف هذا الجانب هما:

١- التقدير الآلي أكثر فعالية بكثير من التقدير البشري، لاسيما عندما يتعلق بمقادير هائلة من البيانات.

٢- يمكن استخدام التقدير الآلي لتحديد وتصنيف نقاط الضعف الخاصة بالنطق بكونها جزءًا من التعليم المحوسب في مكون النطق (ص ١١٣).

اتجه البحث الذي أجرياه إلى تقويم جودة التحليل الحاسوبي لعشرين مختبرًا في اختبارات الإنجليزية بوصفها لغة ثانية؛ وتحديدًا في مهارة قراءة النص قراءة جهرية، وكان هذا مجرد مهمة واحدة في "اختبار تحدث الإنجليزية" (Test of Spoken English) (خدمة الاختبارات التربوية Educational Testing Service, 1986) وقت إجراء البحث. وفي هذا الإطار، حُسِبت النتائج المستقاة من الحاسوب من خلال المطابقة بين التحليل المعروف باسم "التحليل التكراري الأساسي " في استجابات الطلاب الشفهية وبين التحليل الخاص بقالب الأنموذج وأخطاء التصحيح والانحرافات، ويدخل في عداد ذلك الفونيمات المُسقطة (الفونيم أصغر وحدة صوتية تمييزية منطوقة) والفونيمات الخطأ، والفونيمات الجزئية، والزائدة، إضافة إلى مواضع النبر غير الصحيحة. وتألف التحليل من استخدام مُعامل سبيرمان رو (Spearman rho) (.93) وذلك بين النتائج المأخوذة من المقيّمين من بني البشر وبين النتائج المستقاة آليًا. ثمة دراسات أخرى على التحليل اللغوي للاستجابات لم تسفر دومًا عن نتائج إيجابية كهذه (انظر مثلا Reid, 1986)، غير أن المشكلة الأوسع نطاقًا – كما ورد في الفصل الثالث – على ما يبدو تتمثل في الحاجة إلى وسائل أخرى لحساب قيمة التحليل اللغوي؛ بحيث تكون أثرى وأكثر كفاءة من ارتباطات المجموع الكلي اعتمادًا على معيار واحد.

أما الهدف الثاني من دارسة مولهولست وبريسلر (١٩٨٦م)، فيتطلب تحليلاً أدق، لا مجرد ارتباط بين التقديرات البشرية الصرفة وبين التقديرات الآلية، علمًا بأن هذه التحليلات مطلوبة للوصول إلى الغاية المنشودة المتمثلة في استحداث ممارسات

تقويمية متغيرة من حيث الكيف لتأخذ صورًا من شأنها إفادة المتعلمين. إن الهدف المرجو من استخدام التحليل اللغوي لأغراض التشخيص يعني أن الحاسوب سيفوق ما يمكن إنجازه في المعتاد على يد المقدرين من بني البشر، ولذا لا يكفي إجراء البحوث المنهجية المقتصرة على المطابقة بين نتائج التقويم الآلي والبشري. ومما يُذكر أن هذه القضايا قد يتطرق الباحثون إليها في إطار القياس التربوي لبعض الوقت (انظر مثلاً Birenbaum & Tatsuoka, 1987; Mislevy, 1993) لكن ليس في إطار التقويم اللغوي.

التأثير على العاطفة والأداء والفعالية

تدعم البحوث – المعنية بالنظر في الفوائد والأضرار النسبية المترتبة على تنظيم اختبارات اللغة المحوسبة – بالتوثيق مدى كفاءة هذه الاختبارات إذا ما قورنت بالاختبارات الورقية، وما يترتب على ذلك من آثار على الطلاب. وفي هذا الصدد، نجد أن الدراسة الموسعة لشيوع الحاسوب في برنامج توفل (TOEFL) في منتصف العقد الأخير من القرن المنصرم، تشكل مثالاً مناسبًا تناولناه في الفصل الثالث، لكن سبق ذلك بعشرة أعوام بحث قضايا مماثلة في جامعة "بريجام يونج" (Brigham Young University) عندما أعدَّ أحد أوائل الاختبارات الحاسوبية المتكيفة موضع التنفيذ في مجال اكتساب الإنجليزية بوصفها لغة ثانية (Madsen, 1986). وبعد محاولة في ديسمبر عام ١٩٨٥م، طبقٌ اختبار حاسوبي متكيف لتحديد المستوى، تناول عنصري القراءة والقواعد وتألف من ٣٠٠ بند طبق على ٤٢ طالبًا يدرسون الإنجليزية بوصفها لغة ثانية. ولم يُغفَل حينها جمع بيانات ديموغرافية عن الطلاب، والوقت المستغرق لإكمال كل قسم من أقسام الاختبار، إضافة إلى الاستجابات بمقياس "ليكرت" (Likert)، وهي إجابات عن أسئلة تناولت ما يلي:

- الخبرة السابقة بالحاسوب.
- ردود الفعل العاطفية قبيل الاختبار.

- ردود الفعل العاطفية عقب الاختبار.
- الصعوبة النسبية للاختبارين.
- وضوح التعليمات.
- تقويم طول الاختبار.
- وضوح شاشة الاختبار (Madsen, 1986، ص ٤٣).

إضافة إلى ما سبق، طُلِب من الطلاب استكمال أسئلة مفتوحة تتعلق بما استحسنوه وما لم يستحسنوه في الاختبار، إن وُجِد شيء من هذا القبيل. وعلى الرغم من وجود اختلاف طفيف بين الناطقين بلغات أم مختلفة، انتهى الأمر في مجمله إلى أن "برامج الحاسوب الدقيقة التي تستعين بأنموذج "راش" (Rasch) تنطوي على آفاق واعدة بتوفير اختبارات تكيفية أو مفصلة دقيقة لأغراض خاصة، حيث يقل فيها الضغط النفسي الذي ينتاب معظم الطلاب، فضلاً عن انطوائها على إجراءات أكثر كفاءة". إن هذه الدراسة و دراسة اختبار التوفل – التي تناولهما الباحثون منذ ثلاثة عقود على التوالي – تستلزم إعادة النظر فيها حاليًا في ضوء حقيقة مفادها أن الكثير من الطلاب قد ألفوا التعامل مع التقنية الآن، لكنهما تنطويان على أساليب لتقويم الاختبارات تراعي التأثيرات المحتملة للتقنية في العملية الاختبارية.

ملخص

ركزت البحوث السابقة التي تناولت اختبارات اللغة المحوسبة على المخاوف الرئيسة التي كانت تعتري معدي الاختبارات، كما تناولت كل دراسة بمفردها جانباً من جوانب تقويم هذا النوع من الاختبارات. وعلى غرار الاتجاه الذي انتهج في الفصل الثالث، أولت كل دراسة اهتمامها إلى ما اعتبرته مخاوف محددة معلومة، أو مشكلات محتملة، أو فوائد مرجوة من اختبارات اللغة المحوسبة لدى النظر في موقف اختباري بعينه. وبناء على ذلك، تناولت معظم البحوث التي أُجريت على اختبارات اللغة

المحوسبة أموراً ظاهرة على شاكلة توافق البيانات مع الأنموذج، وتأثير هذا النوع من الاختبارات على الطلاب. تلكم القضايا – مضافًا إليها القضايا التي تناولها الفصل الثالث – تشكل جوهر البحوث المعنية بكل من الصدق والثبات، وذلك إذا ما نظرنا إلى الصدق من منظور موسع بكونه منتميًا إلى تفسير نتائج الاختبارات، والنتائج المترتبة على استخدام الاختبار. بيد أن الهدف في اختبارات اللغة يتمثل في الوقوف على تلك المهارات و الخصائص الأخرى المطلوبة، وجمع كل الأدلة المطروحة في هذا الصدد في نظرية واحدة تتعلق بفائدة الاختبار لغرض بعينه. علاوة على ذلك، ووفقًا لما يُستقى من الكثير من الجوانب التي طرحها فولتشر، ينبغي ألا يقتصر تقويم الاختبار على البحوث التجريبية التي أجريت بعد إكمال الاختبار وانتهاء استخدامه.

تبرير الجدوى الإجمالية

قدم تشابيل (Chapelle, 2001) تقويم اختبارات اللغة المحوسبة في إطار أكثر شمولاً للاستفادة من الاختبار (Bachman & Palmer, 1996)، ويضع فيه ستة معايير تهدف إلى توجيه عملية التقويم لدى إجراء التقويم بعينه، وهي: الثبات، وصدق البناء، والموثوقية، والتفاعلية، وسهولة التطبيق. ويمكن تقويم هذه الخصائص المثالية، على أساس يستند إلى رأي تحكيمي وبيانات تجريبية. وقد قُدّمت لذلك أمثلة في كتاب باخمان وبالمر. علمًا بأن تقويم اختبارات اللغة المحوسبة في هذا الإطار، قد خضع للتوضيح من أجل توظيفه في كثير من اختبارات اللغة المحوسبة (Chapelle, 2001; Chapelle, Jamieson & Hegelheimer, 2003). والهدف من وراء ذلك كله هو استحداث وسائل لتقويم اختبارات اللغة المحوسبة؛ بحيث تتسم بالتوافق مع الممارسات القائمة بالفعل في هذا الميدان، بدلاً من تركها نهباً لمخاوف غير مترابطة ومتباينة تحكم على مصير تطورها. وبتطبيق تحليل الجدوى على أربعة تقويمات محوسبة، حدد تشابيل بعض

السمات الإيجابية والسلبية المرتبطة باستخدام التقنية والمتعلقة بالمخاوف التي تكتنف التقويم اللغوي. ويشتمل الجدول رقم (٥,٤) على ملخص لهذه التحليلات التي تستند إلى التحليل الذي أوجزه تشابيل في دراسته عام ٢٠٠١م.

تستند هذه العوامل الإيجابية والسلبية معاً إلى تحليل أربعة اختبارات، أحدها اختبار حاسوبي متكيف لمهارة القراءة؛ يُعنى باستقاء الاستدلالات بشأن عنصر الفهم في القراءة لدى متعلمي الإنجليزية بوصفها لغة ثانية، بغرض استخدام النتائج في اتخاذ قرارات تحديد المستوى للانخراط في برنامج مكثف لتعليم الإنجليزية بوصفها لغة ثانية (Madsen, 1991)، فيما كان الثاني اختباراً لمهارة الاستماع في إطار اكتساب الإنجليزية بوصفها لغة ثانية، وهو اختبار خضع للدراسة في برنامج التوفل خلال استكشاف القائمين عليه لإعداد الاختبار في صورته المحوسبة (Taylor, Kirsch, Eignor & Jamieson, 1999). أما الاختبار الثالث فكان اختباراً لمهارة الكتابة للناطقين بالإنجليزية بوصفها لغة أولى أم، وهو اختبار خضع لدراسة مبدئية على يد باحثين متخصصين (Powers, Fowles, Farnum, and Ramsey, 1994)، وأما الاختبار الرابع فكان مستمداً من مشروع بحثي يبحث في برنامج لتحليل الاستجابات بغية الوقوف على استجابات الطلاب اللغوية في اختبار لمهارة القراءة في إطار اكتساب الإنجليزية بوصفها لغة ثانية (Henning, Anbar, Helm & D'Arcy, 1993). وإضافة إلى الملاحظات المأخوذة من الاختبارات الأربعة، فقد أُضيفت نقاط أخرى إلى هذا الملخص على أساس خبرتنا في التعامل مع الاختبارات المحوسبة.

الثبات

جاءت الخاصية الإيجابية الأولى للثبات من تحليل اختبار مهارة القراءة المشتمل على أسئلة مفتوحة تطلب من الطالب إنشاء إجابة، وإذا كانت الإجابة المنشأة (المكتوبة) غير صحيحة، فإنها قد تكون صحيحة جزئيًا، وبذلك تتكون قاعدة تسفر عن نتائج ذات مدلول جزئي من تنفيذ الحاسوب ينبغي أن توفر قياسًا أكثر دقة وتحديدًا.

الجدول رقم (٥,٤). الخصائص التقنية الإيجابية والسلبية لاختبارات اللغة بالحاسوب.

(مأخوذة من Chapelle, 2001، ص ١١ بتصرف).

الخاصية	السمات الإيجابية	السمات السلبية
الثبات	ينبغي أن توفر النتائج ذات المدلول الجزئي من تنفيذ الحاسوب قياسًا أكثر دقة، وتباينًا أكبر في الدرجات، ومعامل ثبات عالية بمعامل ألفا. تقوم قواعد تسجيل النتائج المحوسبة بأداء متسق من مناسبة لأخرى. يأتي بعد ذلك اختبار دائم للوغاريتمات الاختبارات التكيفية بالحاسوب حتى الحصول على نتيجة ذات مستوى منشود من الثبات.	
صدق البناء	يمكن أن تنعكس المكونات الأكاديمية للقراءة والاستماع الإنشاء (الكتابة) عبر الإنترنت في مهام الاختبارات المحوسبة. احتمال تأثر الاستجابات المفتوحة باستراتيجيات أداء الاختبار أقل من احتمال التأثر الموجود في الاختيار من متعدد.	تختلف أنواع البنود التي يمكن استحداثها حاسوبياً عن الأنواع الممكنة بوسائط أخرى. قد لا يكشف الأداء في اختبار اللغة المحوسب عن القدرة ذاتها التي تُقاس بطرق التقويم الأخرى. قد لا يسفر اختيار البنود – المقرر تضمينها في اختبار تكيّفي بلوغاريثم ما – عن عينة مناسبة لمحتوى الاختبار. قد يخفق تسجيل الاستجابات بالحاسوب في تحديد الدرجة لخصائص الاستجابات المرتبطة بالمكوّن الذي يقصد أن يقيسه الاختبار.
الموثوقية	مهام الاختبارات المحوسبة تحاكي بعض المهام الخاصة بمجال استخدام اللغة المستهدفة.	مهام الاختبارات المحوسبة لا تماثل بعض المهام الخاصة بمجال استخدام اللغة المستهدفة.

تابع الجدول رقم (٥,٤).

الخاصية	السمات الإيجابية	السمات السلبية
التفاعلية	يمكن أن تسهم مدخلات الوسائط المتعددة بفرض للنهوض بمستوى التفاعلية.	البنود القصيرة المستخدمة أحيانًا في الاختبارات التكيفية بالحاسوب قد تؤدي إلى محدودية التفاعل.
التأثير	الاستعداد لاختبارات اللغة المحوسبة ينبغي أن يحفز العمل مع تقنية الحاسوب في صفوف تعلم اللغة الثانية، الأمر الذي سيعين متعلمي اللغة الثانية على اكتساب مهارات مهمة. يمكن تجهيز برامج تعليم اللغات لإتاحة الحواسيب للمتعلمين والمعلمين.	قد تسبب الاختبارات القلق للطلاب الذين يفتقرون للخبرة الكافية في التعامل مع التقنية. قد تكون الاختبارات باهظة التكلفة بالقدر الذي يحول دون أداء الطلاب لها. يمكن أن تؤدي البحوث وإعداد الاختبارات والبنية التحتية لاختبارات اللغة المحوسبة إلى تضييق نطاق موارد الاختبار لتنحصر في التقنية بدلاً من مراعاة أية جوانب أخرى بحثية لا تقل أهمية عن عنصر التقنية.
القابلية للتطبيق	يجعل التحليل المحوسب للاستجابات وجود أسئلة مفتوحة أمرًا ممكنًا في برنامج عملي للاختبارات. تضيف الاختبارات المقدمة عبر شبكة الإنترنت المرونة من حيث الزمان والمكان لتطبيق الاختبارات.	قد ينطوي الأمر على تكاليف باهظة للوصول بصفة منتظمة إلى نتائج ذات مدلول جزئي للبنود. يمكن أن تتمخض الاختبارات المقدمة عبر الإنترنت عن قضايا تتعلق بعنصر الأمان والسرية في الاختبارات المصيرية.

تعكس هذه الزيادة النظرية في معامل الثبات – واتساع الأخذ بها مقارنة بنتائج العناصر المقيّمة على مرحلتين (بشري وآلي) – مدى الدقة في القياس؛ لأن النتائج

تعكس معارف الطلاب على نحو أفضل، وهي المعارف التي نفترض ضعفها مراراً في مرحلة ما من عملية مستمرة: بدايتها حد الصفر ونهايتها حد الاكتمال، بدلاً من افتراض ذلك عند أحد الحدين فقط في العملية المستمرة. أما عمليًا، فنجد أن اتساع التباين جرّاء النتائج ذات المدلول الجزئي، سيسفر عن معدل أعلى من الثبات بمعامل ألفا. وفي الاختبارات ذات الاستجابات المنشأة (المكتوبة) المدخلة، يعد الثبات في عملية التقدير على قائمة المخاوف أيضًا، لكن عندما يتصحح وتسجل النتائج ببرنامج حاسوبي، فإن الشك يقل كثيرًا من حيث اتساق التقدير من عملية إلى أخرى، وفي ذلك نقطة إيجابية ثانية تُضم إلى خاصية الثبات.

ثمة نقطة إيجابية أخرى خاصة بالثبات، وهي الآفاق التي تعد بها الاختبارات الحاسوبية المتكيفة إزاء استمرار الاختبار لحين الحصول على نتيجة يمكن الاعتماد عليها، فالهدف من خاصية التكيف في الاختبارات الحاسوبية المتكيفة، يتمثل في إمداد الطالب ببنود اختبارية تناسب مستواه من حيث الصعوبة والسهولة، وذلك للحصول على تقدير موثوق به عن قدرة الطالب بأقل عدد ممكن من البنود. ومن المعلوم أن إجراء هذا النوع من الاختبارات يمثل تحديًا سافراً للمفهوم السائد الذي يقول: "كلما طال الاختبار، ارتفع معدل الثبات"، وذلك بإنتاج اختبارات أقصر ذات معدل مساو، أو أعلى من الثبات مقارنة بالاختبارات الطويلة. ولا يمكن التعبير عن المفهوم الجديد إزاء إعداد الاختبارات ومعدل الثبات فيها استنادًا إلى عنصر الطول في الاختبار، بل يمكن صياغته كالتالي: "كلما لاءمت البنود الطلاب، ارتفع معدل الثبات".

صدق البناء

تناول الفصل الثالث صدق البناء من منظور المخاوف المرتبطة باستخدام التقنية، ويقدم الجدول رقم (٥,٤) ملخصًا لهذه المخاوف في الجانب السلبي لصدق البناء. ولعل النقاط السلبية المحتملة التي حددها تشابيل في الأمثلة التي فحصها، تتمثل في

القيود المفروضة على الاختبارات الحاسوبية المتكيفة لما تنطوي عليه من فرض لاتجاه بعينه في المهام الاختبارية، مما يجعلها مقاييس وقاصر لقياس الكفاءة النصية. وكان من الأمثلة مثال من اختبار لمهارة القراءة لاكتساب الإنجليزية بوصفها لغة ثانية، وكان اختبارًا حاسوبياً متكيفاً غير مشتمل على نصوص للقراءة (أي لم تشتمل على فقرة كاملة أو نص مترابط) لكنه تألف – عوضًا عن ذلك – من بنود مكونة من جملة إلى ثلاث جمل يطلب من الطلاب قراءتها والاستجابة لها باختيار واحدة من أصل أربعة خيارات (Madsen, 1991). وبصفة أكثر عمومًا، تختلف أنواع البنود التي يمكن استحداثها بصيغ حاسوبية، عن تلك الأنواع التي بمقدورنا استحداثها بوسائط أخرى، وفي هذه الحالة، تمثل الغرض من البنود القصيرة في الالتزام بافتراض الاستقلال المشروط من قبل أنموذج نظرية الاستجابة للبند المستخدمة في حساب مَعْلَمات (بارمترات) البنود في الاختبار، غير أن مثالا آخر من الفصل الثالث، اشتمل على مهام للقراءة تتطلب من الطالب قراءة نص من الشاشة ليس من ورقة، أو الاستماع بالنقر فوق زر على الشاشة، بدلا من التحكم في المدخلات الصوتية تحكمًا خارجيًا.

من النقاط السلبية المحتملة لاستخدام التقنية أيضاً، والمرتبطة أيضاً بالاختبارات الحاسوبية المتكيفة، فقد لا يسفر اختيار البنود – المقرر تضمينها في اختبار متكيف بلوغاريثم ما – عن عينة مناسبة لمحتوى الاختبار. وفيما يتعلق بتحليل الاستجابات، قد يخفق تصحيح الاستجابات حاسوبيًا في تحديد الدرجة لخصائص الاستجابات المرتبطة بالمكوّن الذي يُناط بالاختبار قياسه. وأخيرًا، قد لا يكشف الأداء في اختبارات اللغة المحوسبة عن القدرة ذاتها التي تُقاس بطرق التقويم الأخرى. كل هذه القضايا – إلى جانب البحوث المعنية بتخفيف حدة تلك المشكلات – سبق تناولها في الفصل الثالث، لكن تحليل الجدوى، يدفع الباحثين إلى النظر في الجانب الآخر من تحليل صدق البناء: أي الجانب الإيجابي منه.

يتجلى الجانب الإيجابي من تحليل صدق البناء، في حال مراعاة الحقيقة التي مفادها أن التقنية هي الوسيط الذي يوجد الكثير من مهام استخدام اللغة التي ينخرط فيها الطلاب. وعند التعامل مع بيئة يقضي فيها مستخدمو اللغة الكثير من وقتهم في القراءة والكتابة أمام الحاسوب، فإن أفضل السبل لتحقيق المرجو من مهام القراءة والكتابة الأكاديميتين والإنشاء (الكتابة) عبر الإنترنت، يكمن في المهام الاختبارية المحوسبة. ومن ثمّ، يؤدي الحاسوب دورًا تكامليًا مع عملية الكتابة لدرجة أن فكرة تقويم القدرة على الكتابة من خلال مهمة اختبارية في اختبار ورقي، سينظر إليها معظم الأكاديميين على أنها أداة تفتح الباب للتحيز في القياس. أما القدرات الأخرى فليست قابلة للعرض بمستوى واحد من خلال التقنية، وبناء على ذلك فإن المكون محل الاهتمام، قد حظي بالتحليل الدقيق، لكن الأكثرية ستتفق على أن استخدام الاستجابات المفتوحة إلى جانب الاستجابات القائمة على الاختيار من متعدد، إنما هي ضرورية لضمان صدق البناء في الكثير من مواقف التقويم اللغوي، ويرجع السبب في ذلك – جزئيًا – إلى أن الاستجابات المفتوحة، أقل عرضة للتأثر – مقارنة باستجابات الاختيار من متعدد – بالاستراتيجيات المنتظمة لأداء الاختبارات (أي الحنكة في التعامل مع الاختبار). إن استخدام لوغاريتمات حاسوبية لتصحيح هذه الاستجابات قد يجعل منها صيغة قابلة للاستخدام، وبذلك يزداد عنصر صدق البناء ارتفاعاً.

الموثوقية

تتجلى بعض القضايا ذات الصلة بصدق البناء عند تحليل خاصية الموثوقية، وعلى شاكلة الملاحظة الإيجابية التي تفيد بأن أفضل سبل قياس الكتابة الإنشائية عبر الإنترنت تتمثل في مهام الكتابة المحوسبة، فإن الجانب الإيجابي لتحليل الموثوقية يشير إلى حقيقة مفادها أن كثيراً من المهام الاختبارية المحوسبة تحاكي بعض المهام الأكاديمية؛ ذلك أن تحليل الموثوقية يتطلب من الباحث النظر في أنواع المهام والاستخدامات اللغوية

التي سينخرط فيها الطلاب بعد إجراء الاختبار. وإذا كان كثير من استخدامات اللغة يحدث عبر وسائط تقنية، فهذا يعني أن المهام الواقعية لتقويم القدرات اللغوية، تحتاج هي الأخرى إلى إشراك الطلاب في استخدامات اللغة عبر الوسائط التقنية. لكن في المقابل نجد أن استخدام اللغة مستمر بدون تدخل التقنية؛ فمستخدمو اللغة ما يزالون بحاجة إلى التحدث للآخرين وجهًا لوجه، وعبر الكتب الورقية، والاستماع إلى المحاضرات المباشرة. ومن ثمّ فإن المهام التي تعكس هذه الاستخدامات اللغوية، لم تصبح مقطوعة الصلة بالتقويم اللغوي. وما صنعته التقنية هو زيادة الخيارات المتاحة لمستخدمي اللغة لاختيار وضع اتصال أكثر فعالية من وجهة نظرهم. ولما كان من المتعذر أن يكون الخاضع للاختبار هو القادر دومًا على ممارسة الاختيار، ترتب على ذلك ضرورة إيجاد مجموعة من الأشكال التي تقدم بها اختبارات اللغة إذا أردنا توفير خاصية الموثوقية فيما يتعلق بسياق الاستخدام اللغوي.

التفاعلية

بما أن خاصية التفاعلية تنضوي على التعامل مع استراتيجيات التواصل ولغة المستخدم والمعارف المواضيعية والمتغيرات المؤثرة على نحو متكامل، فإن المهام الاختبارية التي يمكن بناؤها من خلال عروض تفاعلية على وسائط متعددة، تصبح من الخصائص الإيجابية المحتملة في هذا الصدد. ومن المعلوم أن المقدار المطلوب من التفاعلية في أي اختبار، يتوقف على المكون المنوط بالاختبار قياسه، أما الناحية السلبية في الأمر، فتتمثل في احتمال محدودية التفاعلية جرّاء البنود القصيرة المستخدمة في الاختبارات الحاسوبية المتكيفة. وحتى لحظة كتابة هذه السطور، لم تخضع حدود خاصية التفاعلية المحتملة للبحث كما هي الحال في المجالات الأخرى: كألعاب الفيديو مثلاً، وهي مجالات تعد مشاركة المستخدم فيها أمرًا أساسيًا.

التأثير

من المخاوف المرتبطة باختبارات اللغة المحوسبة،والتي نوقشت في الفصل الثالث، احتمال التأثير السلبي لهذا النوع من الاختبارات في كل من: المتعلمين، والتعلم، وصفوف الدراسة، والمجتمع ككل. وقد تأتي هذه التبعات السلبية في صورة قلق لدى الطلاب ذوي الخبرة المحدودة في استخدام التقنية. وقد كان هذا الأمر محور اهتمام مادسن (Madsen, 1986) منذ ما يزيد على ٢٠ عامًا عندما استُحدثت الاختبارات الحاسوبية المتكيفة في إطار اكتساب الإنجليزية بوصفها لغة ثانية. كما قد تكون تكلفة الاختبارات المحوسبة – التي تتضمن الكثير من السمات الإيجابية لخاصيتي الموثوقية والثبات – باهظة التكلفة إلى الدرجة التي تمنع بعض الطلاب من أدائها. فمن ناحية يُنظَر إلى هذه الجوانب على أنها قضية من قضايا التوجه العملي من منظور معد الاختبار، لكن للشخص غير القادر على التقدم للالتحاق بإحدى الجامعات لأداء الاختبار نظراً لضعف إمكاناته المادية، فقد يكون للتبعات السلبية اعتبارات وحضوراً أقوى من ذلك بكثير. وعلى مستوى برنامج الاختبار، يمكن أن تؤدي البحوث، وإعداد الاختبارات، والبنية التحتية لاختبارات اللغة المحوسبة، إلى تضييق نطاق موارد الاختبار لتنحصر في التقنية بدلاً من مراعاة أية جوانب أخرى بحثية لا تقل أهمية عن عنصر التقنية. أما من الناحية الإيجابية لمسألة التأثير، فنجد فيها أن الاستعداد لاختبارات اللغة المحوسبة، ينبغي أن يدفع باتجاه العمل المحوسب في صفوف تعلم اللغة الثانية، الأمر الذي من شأنه مساعدة المتعلمين في تلك الصفوف لاكتساب مهارات مهمة في ضوء الحقائق الثابتة في مجال استخدام اللغة عبر التقنية السائدة في عالم اليوم. كما أنه يمكن تجهيز برامج تعليم اللغات لإتاحة الحواسيب للمتعلمين والمعلمين.

القابلية للتطبيق

تشير خاصية القابلية للتطبيق إلى الحاجة للنظر في الإمكانات التقنية لدى معدي الاختبارات في ضوء المطلوب لإعدادها وتعديلها. وهنا يقدم فولتشر قائمة مفصلة من الاعتبارات المطلوبة للإسهام في رسم صورة دقيقة للأمر؛ مفادها أنه يتطلب قدرًا كبيرًا من الخبرة في أوساط أعضاء هيئة التدريس؛ من أجل إعداد اختبارات اللغة المحوسبة ووضعها قيد التطبيق، فالخبرة تقتضي تكلفة مالية، شأنها في ذلك شأن المعدات ومقتضيات الاستدامة والصيانة. وثمة قاعدة شبه ثابتة في المشاريع التقنية؛ مفادها أنها تستلزم وقتًا ومالاً أكثر من المتوقع. وعلى ذلك فإن أفضل توجيه في هذا الصدد هو الشروع في مشروع محدود، والتخطيط بأكبر قدر ممكن من الحرص. وقد لاحظنا أن مشاريع اختبارات اللغة المحوسبة عادةً ما تدشَّن دون توفير الخبرة اللازمة نظرًا للصعوبة البالغة في إيجاد هذا النوع من الخبرة في هذا المجال، لكن حتى عندما تكون التقنية متاحة، فإنه لا بد من اتخاذ قرارات بشأن تطوير برمجيات التوليف من عدمه، ومدى تعقيد المدخلات وخيارات الاستجابة والتفاعل بينها وبين المدخلات والاستجابة، وكذا بشأن ماهية الإجراءات الواجب اعتمادها للحفاظ على دفع العمل قدماً إلى لأمام على نحو متناسق. ومن نافلة القول أن يطلب معدو اختبارات اللغة المحوسبة المبتدئون بعض التوجيهات والنصائح إزاء الوقت والتكلفة وأنواع الخبرة المطلوبة. لكن الواقع يشهد بأن المتغيرات كثيرة للغاية، وتتفاعل بصور مختلفة في مختلف السياقات لدرجة يُحتمل معها فشل التوجيه العام في تفسير العوامل ذات الصلة. ولذلك نرى أن المطلوب هو تطوير خبرات قائمة على الممارسة في استخدام هذا النوع من الاختبارات وعند إعدادها على يد طلاب مرحلة الدراسات العليا المتخصصين في علوم اللغويات التطبيقية، بما يتيح الأخذ بمعارفهم والاعتداد بها فيما يتصل بقضايا التقنية دائمة التغير.

تكمن قيمة تحليل الجدوى تجاه تقويم اختبارات اللغة المحوسبة في أن هذا التقويم يغطي الجوانب نفسها كما هي الحال في تقويم أي اختبار آخر. فهو لا يتيح للتقنية أن تصبح مشتتة للجهود ناهيك عن أصول ممارسات التحقق والتثبت. بيد أن المشكلة في ذلك هي احتمال الإغفال – جرّاء استخدام تقنية الحاسوب – في خطة التقويم جوانب في اختبارات اللغة المحوسبة تتطلب الاهتمام. وإلى حد ما، يُغفل تحليل الجدوى، بخصائصه المذكورة، بعض أهم النقاط التي أوردها نويجونز وفولتشر في أعمالهما، كما أنه قد يعجز عن التوسع في هذا المجال على النحو الذي حاوله مونير ولورير لتسليط الضوء على الإمكانات التقنية الفريدة بدلا من التركيز على معايير التقويم القياسية. المطلوب إذًا خطة لا تراعي مباعث القلق لدى الباحثين في مجال القياس التربوي على نحو أكثر عمومية فحسب، بل تراعي كذلك مع مباعث القلق التي أفرزتها الإمكانيات الفريدة للتقنية.

تحقيق الصدق واختبارات اللغة المحوسبة

يأمل المرء في أن تشتمل نظرية تحقيق الصدق على التعامل مع مباعث القلق الناجمة عن استخدام التقنية في تقويم اختبارات اللغة المحوسبة. لكن هذه النظرية بذاتها هدف متحرك لحد ما (Kane, 1992, 2001; Kane, Crooks & Cohen, 1999). وعلى غرار المشهد التقني الآخذ في التطور، بل والمدفوع بالمعطيات التقنية إلى حد ما، تخضع نظرية تحقيق الصدق للتغيير بصور من شأنها التأثير في الممارسة. ثمة جانبان من هذه التغييرات قد يكونان مهمين لوضع الإطار المفاهيمي لتقويم اختبارات اللغة المحوسبة في المستقبل وهما: التكوين المعتمد على البرهان للتعبير عن جوانب التقويم، والإطار القائم على الاستخدام لإظهار الاستخدام المناسب.

التكوين المعتمد على البرهان

يقدم العمل الحالي في القياس التربوي، الذي أثر في اختبارات اللغة، طرحًا مفاده تأطير تحقيق الصدق في تكوين معتمد على البرهان بغية توضيح الأساس المُعول عليه في معنى النتائج واستخدامها (Kane, 1992). وبالمثال يتضح المقال: هل تعني النتيجة المتدنية في اختبار القراءة ذي الأهمية أن الطالب لا يملك إلا مستوى ضعيفًا من القدرة على القراءة؟ إن شئنا استخدام لغة التنظير، فإن تدني مستوى القدرة في القراءة لدى الطالب عبارة عن ادعاء قد يرغب المرء في إصداره استنادًا إلى نتيجة اختبار القراءة. لكن النتيجة ذاتها لا تثبت – بكل تأكيد – ضعف مستوى مهارة القراءة لدى الطالب، بل إن مستخدم الاختبار يستنتج – استنادًا إلى نتيجة الاختبار – ضعف القدرة على القراءة، وأن البرهان أو الحجة تظهر عندما يقدم الباحث في الاختبار الدعم المطلوب للوصول إلى هذا التفسير والاستنتاج. ومن الاتجاهات القائمة على طرح البرهان، ما أوضحه باخمان (قيد الطبع) في العبارة التالية:

> ينبغي استنتاج النتيجة الظاهرة بحيث تعني أنها (ادعاء)، ما لم (يكن هناك سبب قوي).

يمكن استخدام هذا الأنموذج العام للتعبير عن معنى النتيجة في اختبار القراءة المحوسب على النحو التالي:

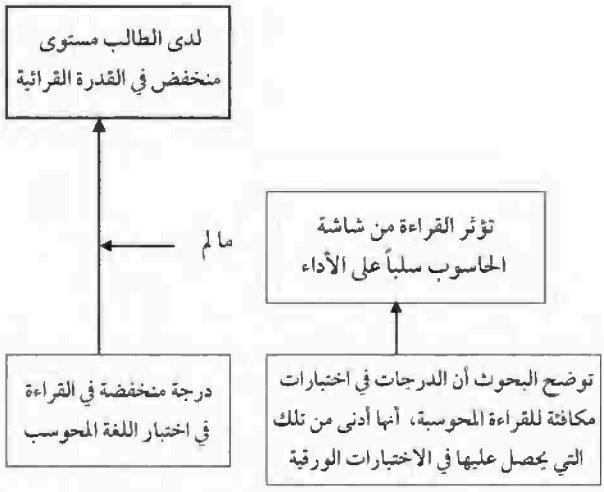
> ينبغي استنتاج النتيجة المتدنية الظاهرة في اختبار القراءة المحوسب، بحيث تعني أن الطالب على مستوى ضعيف من القدرة على القراءة ما لم تكن القراءة على شاشة الحاسوب تؤثر سلبًا في الأداء، بمعنى أن الطلاب دائمًا ما يحصلون على نتائج ضعيفة في اختبارات القراءة المحوسبة مقارنة بالنتائج الأفضل التي يحصلون عليها في المهام الاختبارية التي تعتمد على الاختبارات الورقية.

فيما يلي صياغة لهذا البرهان:

(ادعاء) = ضعف مستوى الطالب في القدرة على القراءة.

(السبب القوي) = القراءة على شاشة الحاسوب تؤثر سلبًا في الأداء بمعنى أن الطلاب دائمًا ما يحصلون على نتائج ضعيفة في اختبارات القراءة المحوسبة مقارنة بالنتائج الأفضل التي يحصلون عليها في المهام الاختبارية التي تعتمد على الاختبارات الورقية.

يلائم التكوين المعتمد على البرهان تمامًا المنطق الذي تقوم عليه البحوث والاستقصاءات التي تركز على المخاوف التي تحيط بعنصر الصدق، كما هو موضح في الفصل الثالث؛ فعلى سبيل المثال، يوضح الشكل رقم (٥,١) تكوين البرهان المقرر صياغته لتفسير البحوث المعنية بالنظر في أحد المخاوف. ويتألف الشكل من نتيجة الاختبار في المربع السفلي، في حين يشار إلى الاستنتاج بالسهم المتجه إلى الادعاء "ضعف مستوى القدرة في القراءة لدى الطالب" (the examinee has a low level of reading ability). أما سهم "ما لم" (Unless) فيشير إلى العبارة التي من شأنها إضعاف الاستنتاج، وهي هنا عبارة "القراءة على شاشة الحاسوب تؤثر سلبًا في الأداء" (reading on a computer screen negatively affects performance). تظهر نتيجة البحث في المربع الواقع أسفل كلمة "ما لم" كونها داعمة لمعنى الكلمة واستنتاجها. وتشير النتائج في هذا المثال إلى بحث تشوي وكيم وبو (Choi, Kirn, and Boo, 2003) الذي سبق عرضه في الفصل الثالث، والذي انتهى فيه الباحثون إلى أن ضعف نتيجة الطلاب في اختبار القراءة المحوسب، وهو الاختبار الذي كان من المفترض فيه أن يسفر عن نتائج مكافئة للاختبار الورقي. وهذا النوع من البحوث يبدو أنه يدعم ويبرهن صحة عبارة "ما لم" سالفة الذكر؛ أي أن عبارات "ما لم" تستمد قوتها وحجتها من هذا الدعم.



الشكل رقم (٥,١). مثال على تكوين برهاني أعد اعتماداً على مخوف للصدق.

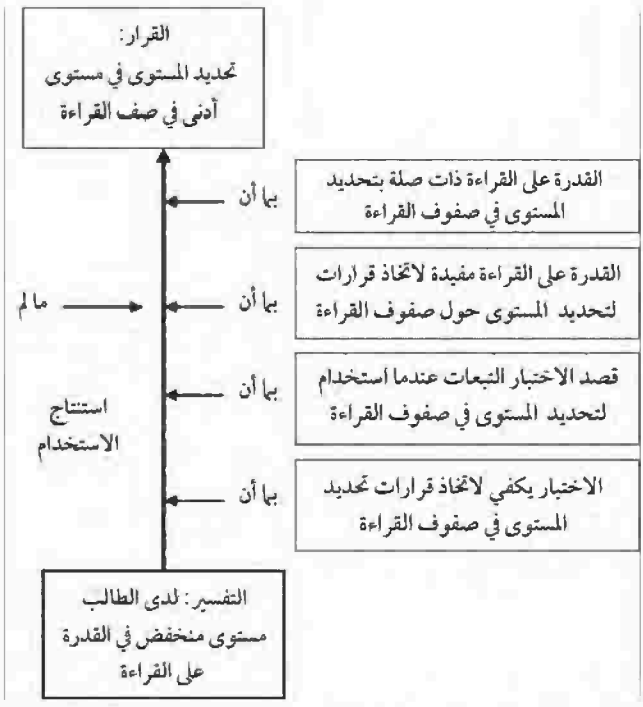
يتفاوت مقدار الدعم الذي تحظى به عبارة "ما لم" سالفة الذكر، وذلك حسب طبيعة نتائج البحث وعدد الدراسات الداعمة، ويتمثل الدعم القوي لهذه العبارة في صورة العديد من المربعات المصحوبة بنتائج البحث التي تدعم عبارة "ما لم". ومن الطبيعي أن الدعم القوي لهذه العبارات، يؤكد السبب في عدم الذهاب للاستنتاج المقصود بشأن معنى درجات الاختبار. وفي هذا أسلوب أكثر رسمية للتعبير عن فكرة المخوف الذي يحيط بالصدق في تفسير نتيجة الاختبار. علاوة على ذلك، يشتمل الدعم على مساحة لعرض الدعم القائم على التجربة لدى التعامل مع المخوف المفترض. ومن ثمّ، فإنه يمكن للمرء – على سبيل المثال – وطرح فكرة مفادها أن القلق من الحاسوب لدى الطلاب يؤثر سلبًا في درجات الاختبار فيما يتصل بوضعية عبارة "ما لم"، لكن دون أي برهان مستند إلى البحث أو الملاحظة لدعم مدلول العبارة، فإن الفكرة لن تضعف الدعم المنصرف إلى الاستنتاج المقصود.

أما في الجانب الأيسر من سطر الاستنتاج، فيمكن إضافة عبارة "بما أن" (since) للنظرية دعمًا للحصول على الاستنتاج. وعبارة كهذه يمكن أن تفيد – على سبيل المثال– بأن " الطالب استكمل التوجيه المبدئي لكيفية الاستجابة لبنود الاختبار"، علمًا بأن دعم هذه العبارة قد يتقدم على البحث بما يفيد أن الطلاب الذين تلقوا توجيهًا مبدئيًا حول كيفية التفاعل بنجاح مع واجهة المستخدم الحاسوبية، قد أصبحوا قادرين على إظهار أداء متكافئ في اختبار مهارة القراءة؛ سواء أكان ذلك في الاختبار المحوسب أم الاختبار الورقي. خلاصة القول: أن التكوين المعتمد على البرهان، يوفر السبل اللازمة للتعبير عن الاستنتاج المقصود إلى جانب عبارتيْن أخريَيْن ("ما لم" و"بما أن") اللتين من شأنهما تقوية الاستنتاج أو إضعافه، وكذا إظهار دعم البحث أو عجزه عن دعم العبارات. وبناء على ذلك، يبين المثال أن هذا التكوين يمكن استخدامه لوضع أنواع المخاوف التي نوقشت وفُحِصت في إطار برهان يخص معنى نتيجة الاختبار، وبذلك نجد أن هذا التوجه يربط بين فكرة النظرة العامة للمهدد، وبين الوسائل الرسمية للتعبير عن معنى نتيجة الاختبار.

الإطار القائم على الاستخدام

يرى باخمان (في دراسة له قيد الطبع) أن البرهان التفسيري، لا يكمّل سوى الجزء الأول من الحجة من خلال توضيح التفسير المنصرف إلى نتيجة الاختبار، غير أن لب الأمر في السياقات التربوية، يتمثل في القرارات المتخذة على أساس التفسير. ولذا يوضح باخمان ما يسميه "برهان الاستخدام" الذي يشكل جزءًا من "حجة استخدام التقويم" التي ينبغي أن تدعم استخدام الاختبار لغرض بعينه. ويصف باخمان في عمله هذا (قيد الطبع) برهان الاستخدام من خلال إطار مستند إلى تفسير النتيجة (بمعنى النتيجة أو الخلاصة من البرهان التفسيري).

وتجدر الإشارة إلى أن بُنية برهان الاستخدام مماثلة للبرهان التفسيري، باستثناء أن نقطة البداية هي تفسير النتيجة، وأن نقطة النهاية هي القرار المتخذ على أساس التفسير. ويظهر تفسير النتيجة في مربع معتم إشارةً إلى تطابقه مع المربع الموضح في الشكل رقم (٥,١)، علمًا بأن الشكل رقم (٥,٢) يبيّن المكونات الرئيسة لبرهان الاستخدام وفقًا لطرح باخمان (المادة قيد الطبع) باستخدام مثال اختبار القراءة المحوسب. وفي الشكل ذاته يربط استنتاج الاستخدام، بين التفسير المطروح لضعف مستوى الطالب في القراءة، وبين قرار تحديد المستوى المفضي إلى إلحاقه بصف دراسي مخصص لذوي المستويات الضعيفة في القراءة. ونرى أن التفسير مدعوم بالبرهان التفسيري، وأن برهان الاستخدام هو الذي يترك لتفسير استخدام الاختبار.



الشكل رقم (٥,٢). مثال على برهان الاستخدام لاختبار القراءة المحوسب.

يقدم باخمان (المادة قيد الطبع) التوجيه بشأن أنواع الاعتبارات التي ينبغي أن تظهر في عبارات "بما أن" في إطار برهان الاستخدام وهي: الصلة، والاستخدام والتبعات الإيجابية، والكفاية. ويوضح الشكل رقم (٥,٢) أمثلة لعبارات ذات صلة بهذه الاعتبارات، لكن الدعم الذي من شأنه تأييد كل اعتبار فيها لم يُذكر إيثارًا للتبسيط في سوق المثال. ولا يخفى أن برهان الاستخدام هذا، يوفر الوسائل اللازمة لتضمين القضايا التي تخص الاستخدام الواقعي للاختبارات في إطار عام وشامل لهذا الاستخدام. من جانب آخر، تقدم عبارات "بما أن" منظورًا مفيدًا فيما يتعلق باختبارات اللغة المحوسبة؛ فمن المعتاد سماع نظرية لاستخدام هذا النوع من الاختبارات ترتكز في المقام الأول على بيان أداة اتخاذ القرار (أي عمليات تحديد المستوى القائمة على النتائج المأخوذة من اختبار قصير ومبسط بعد تصحيحه آليًا). غير أن نظرية تقويم الاستخدام التي يطرحها باخمان، تميط اللثام عن عجز هذه النظرية بما مفاده: تتطلب أي نظرية لتقويم الاستخدام تضمين نظرية تفسيرية توضح معنى (أي تفسير) نتائج الاختبار، وهذا التفسير لا بد أن يكون نقطة بداية في نظرية استخدام تتألف من برهان يدعم بيان هذا الاستخدام جنبًا إلى جنب مع البيانات الأخرى الخاصة بجوانب الصلة والتبعات المنشودة والكفاية.

إن الزمن جدير ببيان مدى التقاطع والالتقاء بين المخاوف التقنية متعددة الأوجه مع جوانب النظرية التفسيرية وبرهان الاستخدام. أما حالياً فيبدو أنه على الرغم من ما تضيفه التقنية من مباعث قلق جديدة، قد تؤثر على عملية الصدق، فإنه لا بد من تضمين هذه المخاوف نهاية المطاف في البرهان أو الحجة الشاملة لتفسير نتائج الاختبارات واستخدامها.

خاتمة

ثبت أن كلا من تحليل الجدوى وبرهان استخدام التقويم، يقدمان طرقًا لعرض تقويم اختبارات اللغة المحوسبة في إطار أشمل وأرحب من التدقيق، وأن نظرية برهان

الجدوى تسلط الضوء على العوامل ذات الاهتمام الخاص في تقويم القدرات اللغوية – ومن ذلك عنصر الموثوقية – التي تطرقت إليها أبحاث مونير ولورير باعتبارها القضية الأهم في إعداد الاختبار وتقويمه. أما نظرية البرهان التفسيري، فتؤسس للمنظومة التي يمكن من خلالها النظر في مباعث القلق التي أثارها كل من نويجونز وفولتشر فيما يتعلق بأداء الاختبار وواجهة المستخدم، إلى جانب المخاوف المختلفة ومباعث القلق التي قد تعزز من الأساس المتبع في تفسير النتائج والدرجات أو تُضعِفه. وأما نظرية برهان الاستخدام، جانب اتخاذ القرار في الاختبار الأولوية من أجل توجُّه مسار تحليل البرهان المعني بالنظر؛ لا في الأدلة الخاصة بالأداة المستخدمة فحسب، بل أيضاً بالصلة والنتائج، علاوة على ملاءمة درجات الاختبار وكفايتها لغرض بعينه.

وفي الوقت نفسه، يبدو أن بعض القضايا المرتبطة باختبارات اللغة المحوسبة، قد مضت دون ملاحظة في الأُطُر الخاصة بأساليب التحقق من الصدق، لا سيما تكامل عملية التقويم مع عملية التعلم واستخدام الاختبارات في بحوث اكتساب اللغة الثانية. وفي هذه السياقات، يظل جانب تفسير النتائج موضع أهمية واعتبار، غير أن الأمر لم يرق إلى مستوى الوضوح الكافي فيما يتعلق بضرورة جعل اتخاذ القرار القوة الموجهة في نظرية برهان الاستخدام أم لا. وبدلاً من ذلك، فقد تشتمل القضايا محل البحث على عوامل مثل: الصلة، والاستخدام، والكفاية، والتبعات لمراعاتها عند تطوير النظرية ووضعها موضع الاختبار الباحثين، أو لنشر فكرة التعلم الذاتي على سبيل المثال. وفي الفصل التالي سنعود إلى قضية التفكير الاستباقي لاستخدام التقنية في التقويم باعتبارها اتجاهًا يستحق تناوله بمزيد من التقصّي والبحث.