الفصل الثالث

المخاوف المترتبة على اختبارات اللغة المحوسبة
The Threat of CALT

توفر الخيارات الجديدة التي استحدثتها التقنية للتعاطي مع مهام اختبارات اللغة، التي يضطلع بها المعلمون ومعدو الاختبارات والباحثون، طيفاً واسعاً من الإمكانات التي يستطيعون من خلالها تقييم القدرات اللغوية. بيد أن هذه الخيارات تثير في الوقت نفسه أسئلة عن حقيقة تقيسه هذه الاختبارات الجديدة، وعن المجالات التي يمكن الاستفادة منها بالنتائج المستقاة من اختبارات اللغة المحوسبة. ومن واقع تجربتنا، نعلم أن مجرد ذكر التقنية في مجال اللغة، يثير سيلا من التعليقات المفعمة بالشكوك من لدن كل من: المعلمين، ومعدي الاختبارات، والباحثين الذين يضطلعون بعملية التقييم. ولقد عبّر أحد الباحثين المعنيين باختبارات اللغة عن هذه الشكوك، بما كتبه عما ينطوي عليه هذا النوع من الاختبارات من مهددات محتملة، وكذلك ما يحمله في طياته من آفاق واعدة. وقد تناول الفصل الثاني الخيارات الجديدة التي تكفلها تقنية الحاسوب في إطار الخصائص المعروفة لطرائق الاختبار. أما هذا الفصل فيتناول المخاوف المذكورة من خلال توضيح أوجه الارتباط بين المخاوف المنبثقة عن استخدام التقنية في تقييم القدرات اللغوية من جانب، ومدى الصدق فيها من جانب آخر. وتتعلق بعض المخاوف المرتبطة بالاختبارات المحوسبة، كما أوردنا في الفصل

الثاني، بما يلي: ١- الاستنتاجات التي يمكن استقاؤها من العلامات المسجلة من الاختبارات المحوسبة، ٢- أوجه استخدام تلك العلامات لأغراضٍ نحو: منح الشهادات والدرجات، واتخاذ قرارات بالقبول، وكذا إرشاد المتعلمين، أي: أننا أمام مسألتين تتعلقان بعنصر الصدق تشكلان الأساس الذي تتفرع منه –على الأقل– ستة مباعث للقلق التي يعبَّر عنها عادةً كمخاوف محتملة للاختبارات المحوسبة. ومن ثمّ سنتطرق في هذا الفصل إلى تلك المخاوف المحتملة، وسنوضح السبل التي يمكن للجهود البحثية من خلالها التعامل مع تلك المخاوف. وقد يُلاحظ أن هذا الفصل يثير أسئلة أكثر مما يجيب عليها، غير أن هذا الملحظ يعكس حالة معرفية تمثل ما انتهت إليه معارفنا بشأن الاستدلالات والاستخدامات المرتبطة باختبارات اللغة المحوسبة، وما من شك في أن الوقوف على تلك القضايا يمثل الخطوة الأولى في سبيل المضي قدما في هذا الصدد.

يلخص الجدول رقم (٣.١) المخاوف المحتملة لعنصر الصدق في الاستدلالات والاستخدامات المرتبطة باختبارات اللغة المحوسبة، علمًا بأننا جمعنا محتوى القائمة الواردة في العمود الأيمن من جملة التعقيبات والمخاوف التي أعرب عنها الطلاب، والمعلمون، ومعدو الاختبارات، والباحثون طوال فترة احتكاكهم التدريجي بهذا النوع من الاختبارات على مدار العقدين الماضيين. وبعض هذه المخاوف – كعنصر تأمين الاختبار وسريته – حظي بالاهتمام في الخطاب العام بشأن التقييم، في حين لم تحظ جوانب أخرى –كتسجيل الإجابات اللغوية– بهذا القدر من الاهتمام في الأوساط العامة كاللجان المعنية بإعداد الاختبارات والمشاريع البحثية.

الجدول رقم (٣,١). المخاوف المحتملة للصدق في اختبارات اللغة المحوسبة.

المخوف المحتمل للصدق	توجهات للتعامل مع المخوف
اختلاف معدلات الأداء في الاختبار. قد لا يكشف الأداء في اختبار اللغة المحوسب عن القدرة ذاتها التي تُقاس بطرق التقييم الأخرى.	دراسات التحقق من مقارنات الأوساط والأنماط عقد مقارنات بأسلوب المجموعة المغايرة (المقابلة) الوقوف على آثار طريقة الاختبار بالحاسوب
أنواع المهام الجديدة. تختلف أنواع البنود التي يمكن إعدادها بصيغ حاسوبية عن الأنواع الممكنة بوسائط أخرى.	الوقوف على أنواع البنود الجديدة ودراستها
القصور الناتج عن الاختيار التكيّفي للبنود. قد لا يسفر اختيار البنود المقرر تضمينها في اختبار تكيّفي بلوغاريثم ما عن عينة مناسبة لمحتوى الاختبار، وقد يسبب قلقاً لدى الطلاب.	استكشاف سبل أخرى غير مرتبطة بالتكيف على مستوى البند التجربة مع مراعاة الاختلاف في تقديم البنود والتحكم فيها
عدم الدقة في تصحيح الاستجابات إلكترونيًا. قد يخفق تصحيح الاستجابات حاسوبياً في تحديد الدرجة لخصائص الأجوبة المرتبطة بالمكوّن الذي يُناط بالاختبار قياسه.	تطبيق معايير مناسبة متعددة الجوانب لتقويم عملية التصحيح الآلي مقارنة التصحيح التفصيلي (التقدير الجزئي) بطرائق التصحيح الأخرى
الإخلال بعنصر تأمين وسرية الاختبار. تنطوي اختبارات اللغة المحوسبة على مخاطر تتهدد أمن وسرية الاختبارات.	عقد مقارنات بين التقديرات الآلية والتقديرات البشرية استحداث معايير تصحيح مناسبة (ذات صلة بالمكونات)
التبعات السلبية. قد تنطوي اختبارات اللغة المحوسبة على آثار سلبية تطال المتعلمين والعملية التعليمية، وصفوف الدراسة، والمجتمع.	الوقوف على المقدار المناسب من بنود الاختبارات الحاسوبية المتكيفة تأسيس مراكز اختبارات استخدام التقنيات المتطورة لتحديد الهويات التخطيط وإعداد الميزانيات على نحو سليم إعداد المتعلمين لاختبارات اللغة المحوسبة فحص الآثار السلبية للاختبارات

وعلى الرغم من الأساس الذي تنبعث منه هذه المخاوف، فإن كل مخوف محتمل مما سبق ذكره، يمس أي فرد مهتم بمستقبل التقييم اللغوي؛ لأن هذه القضايا ستمثل دورًا محوريًا على صعيد البحث والممارسة العملية.

اختلاف معدلات الأداء في الاختبارات

لعل أوسع مباعث القلق التي تثار حول استخدام التقنية في التقييم اللغوي، هو ما قد ينطوي عليه أداء الطلاب في اختبارات اللغة المحوسبة من احتمالِ الإخفاق في الكشف عن القدرة ذاتها التي تُقاس بصور أخرى من التقييم، علمًا بأن المشكلة المحتملة تخص الاستدلالات التي توصل إلى قدرة الطلاب على أساس أدائهم في الاختبار. ومن المسلم به أنه إذا أسفر اختبار محوسب عن نتيجة تختلف في المتوسط عن النتيجة المتوقعة، ولو كان الاختبار مماثلاً لاختبار ورقي، فإن المخوف يعد خطيراً إذا ما اعتبر مستخدمو هذه النتائج أن كلتا النتيجتيْن متكافئتان. وهذا التوقع الضمني باعث أكيد على تناول هذا المخوف المحتمل بالنقاش؛ ولذا كان ما يلي أحد الأساليب التي عبر بها متخصصون في القياس التربوي عن هذه المشكلة: "إذا كان تقديم البنود على شاشة حاسب – بدلاً من ورقة – من شأنه أن يغير من العمليات العقلية المطلوبة للإجابة الصحيحة عن البند، فإن الصدق في الاستدلالات الذي استند إلى تلك النتائج قد يتغير هو الآخر" (Wainer, Dorans, Eignor, Flaugher, Green, Mislervy, Steinberg & Thissen, 2000، ص ١٦).

وتشير المقارنة المنشودة إلى اختبار محوسب مقابل اختبار مقدم بأي شكل آخر؛ سواء أكان ورقيًا أم شفهيًا على هيئة حوار، أم حديثاً مسجلاً لشخص على شريط تسجيل. ولا شك أن مبعث القلق هذا إزاء كل الاختبارات المحوسبة، يرتبط باختبارات اللغة الثانية التي تتطلب من الطالب في المعتاد قراءة النصوص والتعامل معها على

شاشة الحاسوب، أو الاستماع إلى مجموعة من الأصوات، ومشاهدة مجموعة من الصور المقدمة في صورة مدخلات اختبارية كما أسلفنا في الفصل السابق. وقبل أكثر من خمسة عشر عاماً، أوضح أحد الرواد في مجال الاختبارات المحوسبة، عددًا من الأسباب المحتملة لتوقع معدلات أداء مختلفة في هذه الاختبارات، واشتملت هذه الأسباب على اختلافات في سهولة التراجع، وسعة استيعاب الشاشة، وسهولة الاستجابة، واستخدام عنصر التحكم في الوقت التي تخضع للتحكم الفردي (Green, 1988). ثمة تغييرات طرأت على بعض هذه التفاصيل اليوم، ويعزى ذلك، في جانب منه، إلى التطورات التقنية، ولكن يعزى أيضاً إلى الألفة الكبيرة التي اكتسبها الطلاب من استخدامهم للحاسوب.

لكن اعتياد الطلاب على البحث عن المعلومات على شبكة الإنترنت، واستخدام برنامج لمعالجة النصوص، لا يحول دون تعرضهم لتحديات تعترضهم في التعامل مع الحاسوب والتحكم من خلال اللغة الثانية، وهي تحديات يرى النقاد أنها أقرب صلة إلى مهارات التعامل مع الحاسوب منها إلى القدرات اللغوية. وبالمثال يتضح المقال؛ فالطلاب الخاضعون لاختبار في قواعد اللغة — يتطلب منهم النقر على عبارات، ومن ثم سحبها إلى أماكنها الصحيحة لتكوين جملة سليمة — قد لا يعرفون كيفية التعامل مع هذا النوع من البنود الاختبارية إذا لم يسبق لهم التعامل معه قبل الخضوع للاختبار. وقد يسفر هذا النقص المعرفي — في بيئة الاختبار — عن اهتزاز الثقة والتأثير سلبًا على أداء مهام الاختبار، حتى وإن كان الطالب حاذقاً ولديه معرفة قوية بالقواعد. والأدهى من ذلك أن يسفر هذا النقص عن مجموعة من البنود غير المكتملة جرّاء إهدار الوقت في محاولة التعرف على كيفية الاستجابة. وفي كلتا الحالتين، سيكون الاستدلال الذي نستقيه من نتيجة الاختبار هو ضعف الطالب في القواعد، في حين أن الاستدلال الصحيح والحقيقي هو عدم كفاءة هذا الطالب في النقر على البنود وسحبها إلى

أماكنها. إن المشكلة العامة – المتمثلة في تأثير الحاسوب على الأداء – ينبغي أن تحظى باهتمام بالغ لدى المعنيين باستكشاف كل أنواع اختبارات اللغة، علمًا بأنه يمكن استخدام عدد من السبل للتعاطي مع هذه المشكلة، بيد أن فحص الأمثلة القليلة التي وردت في ثنايا البحوث، تشير إلى أن معدي اختبارات اللغة المحوسبة، قلما تطرقوا بالبحث إلى هذه المسألة الحساسة.

الدراسات المقارنة بين الاختبارات

لعل أوضح السبل لتقصي مدى كفاءة أداء الطلاب في الاختبارات المحوسبة وعزو ذلك إلى السبب غير الصحيح (بمعنى الأداء التفاضلي في الاختبارات بسبب عوامل لا علاقة لها البتة بالاختلافات في القدرة المطلوب قياسها) يتمثل في إجراء دراسةٍ قادرة على المقارنة بين أداء الطلاب في اختبارين متماثلين في كل شيء ماعدا الوسيلة (الوسيط) التي يقدم بها الاختبار؛ بمعنى أن تكون إحداهما ورقية والأخرى محوسبة. وإن شئنا النظر في أمثلة واقعية، فسنجد أن واحدًا من أولى برامج الاختبارات الكبرى بالولايات المتحدة قد حول الاختبارات إلى منظومة اختبارية حاسوبية – وهو البرنامج المعروف اختصارًا باسم "GRE" "فحص نتائج الخريجين" (Graduate Record Examination)– حيث عقد عددًا من المقارنات بين بنود الاختبارات، وأقسامها، ونتائجها الإجمالية في إطار برنامج بحثي يهدف إلى الوقوف على أوجه التشابه والاختلاف بين صورتي الاختبارات المحوسبة والاختبارات الورقية المقدمة في إطار البرنامج. وفي العديد من الدراسات التي حصلت على بيانات من الطلاب حول أدائهم في اختبارات محوسبة وأخرى ورقية في إطار البرنامج المذكور، وجد الباحثون اختلافات طفيفة للغاية، يُتوقع أن تتطلب المزيد من البحث والتقصي للآثار المترتبة على اختلاف وسيلة الاختبار (الوسيط)، لكنهم لم ينتهوا إلى آثار جلية واضحة وثابتة تؤكد على ضرورة اختلاف الاستدلالات والنتائج باختلاف نوعي الاختبارات في إطار البرنامج (Schaeffer, Reese, Steffen, McKinley & Mills, 1993).

تجدر الإشارة إلى أن اختبارات هذا البرنامج تتضمن قسماً لتقييم القدرة على التحدث، لكن الجانب الأقرب إلى القضايا المرتبطة باختبارات اللغة المحوسبة تمثل في البحث الذي أُجري على اختبار "التوفل" لتوفير معلومات حول أوجه التشابه بين الأنموذج الورقي وأنواع المهام المزمع إدراجها في اختبار "التوفل" المحوسب، وهو الاختبار الذي طُرِح عام ١٩٩٨م (Taylor, Kirsch, Eignor & Jamieson, 1999). وقد طبق اختبار "التوفل" بنوعيه الورقي والمحوسب على الطلاب، ووجد الباحثون معامل ارتباط قدره ٠,٨٤ بين الاختبارين. ورغم أن هذا الارتباط ليس بالقدر المخصوص من القوة لدى النظر إليه باعتباره أنموذجا موازيًا يشكل معاملاً لعنصر الثبات، فإن الباحثين فسروه على أنه مؤشر إلى عدم التأثير القوي للحاسوب إذا ما استخدم وسيلة للمادة موضوع الاختبار. غير أن التفسير الذي يمكن استقراؤه من ارتباط ذي متغيّرين في اختبارين متماثلين لقياس القدرات اللغوية، إنما هو تفسير محدود بطبيعته. ومن أجل السعي إلى عزل الكيفية التي يمكن أن تؤثر بها خصائص المهمة الاختبارية في الأداء في الاختبار، يأمل المرء في إجراء بحث مماثل للبحث الذي أُجري في إطار برنامج GRE ؛ أي إلى عقد مقارنة دقيقة بين البنود الاختبارية المحوسبة والبنود التقليدية بهدف تقييم القدرات اللغوية نفسها. وثمة أمثلة على ذلك تناول أحدها بالفحص والتدقيق اختلاف معدلات الأداء في اختبار استماع قُدّم مرة بمدخلات صوتية فقط، ومرة أخرى بمدخلات صوتية وصورية (Coniam, 2001). ولم يخرج البحث باختلافات كمية ذات دلالة بين نوعي الاختبار في هذه الحالة، لكن المسألة برمتها تستحق مزيدًا من البحث والتقصي.

هناك دراسة مقارنة أخرى تناولت أوجه التماثل والاختلاف استنادًا إلى مجموعة من الطرائق التجريبية التي أتاحت عددًا من المنظورات المتممة. فقد سعى كل من Choi وKim وBoo (2003) إلى إيجاد دليل على أوجه المقارنة بين اختبارات اللغة الورقية ونظيراتها المحوسبة في اختبار الكفاية في اللغة الإنجليزية (Test of English Proficiency)

الذي أعدته جامعة سيول الوطنية، وذلك من خلال تحليل المحتوى وتحليلات أوجه الارتباط بين النماذج، وبطريقة "أنوفا" البحثية (ANOVA) وتحليل عوامل البرهان. ومجمل القول أن النتائج أكدت وجود أوجه تشابه قوية بين نوعي الاختبار (الورقي والمحوسب) في كل قسم من أقسام الاختبار، في حين حظيت أقسام قواعد اللغة بأكبر نسبة من أوجه التشابه، أما أعلى نسبة من أوجه التباين فسُجِلت في أقسام القراءة.

كما نجد أن النتائج التي أكدت على عدم وجود أوجه اختلاف ذات مغزى بين نوعي الاختبار، ينبغي أن تطمئن المستخدمين بأنه لا يوجد تأثير كبير للحاسوب باعتباره وسيلة للاختبار في نتائج الاختبارات. لكن إنعام النظر في هذه النتائج، يؤدي إلى طرح عدد من الأسئلة أولها: إذا كان تقديم الاختبار بالحاسوب ينتج عنه اختبار يتماثل في جله مع الاختبار المقدم بوسيلة أخرى، فهل يستحق الأمر بذل التكلفة اللازمة لتغيير وسيلة الاختبار؟ وإذا كان الاختبار المحوسب أكثر فعالية، فهل الفعالية وحدها في حد ذاتها سبب كاف للتغيير؟ أولا ينبغي السعي – والحال هذه – إلى وضع اختبار أفضل بدلا من السعي إلى التغيير؟ وبعد أن بذل معدو الاختبارات الوقت والمال في إعداد اختبار لمهارة الاستماع مصحوبًا بالصور (فيديو) – على سبيل المثال – فإننا نصبو إلى الحصول على نتائج أفضل في الاختبارات يؤخذ بها كمؤشرات دالة على مدى إجادة الاستماع باعتباره المهارة التي وُضِع الاختبار لقياسها. ولا يخفى أن وجه الأفضلية هنا يميل إلى الاختبار المصور مقارنة بالاختبار الذي يقتصر على مادة سمعية ليس إلا. وثانيها: هل الإسهام في نتائج الاختبار – نعني إسهام قدرة الطلاب على التعامل مع الحاسوب – غير ذي صلة فعلاً بقدراتهم اللغوية كما ينبغي تحديدها وتعريفها في القرن الحادي والعشرين؟ فعلى سبيل المثال، إذا كان الطلاب الذين سيلتحقون بإحدى الجامعات في الولايات المتحدة غير قادرين على التنقل وإدخال إجاباتهم في الاختبار المحوسب المنعقد لتحديد المستوى – الاختبار المعروف باسم "WebLAS" – فهل هم بذلك مستعدون

فعلاً لاستخدام اللغة الإنجليزية في الوسط الأكاديمي بالولايات المتحدة؛ حيث تتم الإجراءات الإدارية والأكاديمية في جانب منها من خلال التعاطي مع وسائل التقنية؟ وثالثها: هل يمكن اعتبار التماثل العام في نتائج نوعي الاختبار مؤشرًا يؤكد أن قلة من الأفراد، ضعاف الخبرة في استخدام الحاسوب، لا يتأثرون سلبًا بتقديم الاختبار عبر وسيلة اختباريه بعينها؟ سنتناول القضيتين الأولى والثانية في الفصول اللاحقة، أما القضية الثالثة فسيتم تناولها من خلال بحوث تعقد مقارنات بين معدلات الأداء في الاختبارات التي خضعت لها مختلف المجموعات قيد البحث.

الدراسات التقابلية للمجموعات

استُخدمت مجموعات التقابل للوقوف على مدى تأثر الطلاب قليلي الخبرة بالحاسوب، سلبًا عند الخضوع للاختبارات المحوسبة، ذلك أن هذا التأثير السلبي قد يظهر في صورة علاقة أقل مما يتوقع المرء بين الاختبارات المحوسبة والورقية لدى التعاطي معها في مجموعة كبيرة من الأفراد، لكن هذه النتيجة لا يمكن عزوها إلى تأثير سلبي أو نقص لدى أفراد بأنفسهم. وفي هذا السياق، يشير علماء من أمثال Steinberg, Thissen, and Wainer (2000) إلى المشكلة المحتملة بمسمى "الصدق التفاضلي":

> من المعلوم أن (الاختبار المحوسب) يكون أكثر عرضة لمشكلات التفاضل في عنصر الصدق إذا ما قورن بالاختبارات الورقية، ويعزى ذلك إلى احتمال الاختلافات الكائنة بين المجموعات العرقية أو الجنس من حيث ألفتهم بالحاسوب بصفة عامة، الأمر الذي يؤثر في نتائج الاختبارات المحوسبة (ص ٢٠٤).

يحظى هذا الأمر بأهمية خاصة في اختبارات اللغة الثانية (L2) عالمياً؛ هذا إذا ما علمنا أن بعض الأفراد في العالم تكون لديهم خبرة ودراية أقل من غيرهم في استخدام الحاسوب. وقد توصلت البحوث التي طبقت لمعرفة مدى الألفة بالحاسوب لدى أكثر من ١٠٠,٠٠٠ شخص في (اختبار الإنجليزية بوصفها لغة ثانية ESL) على

مستوى العالم عام ١٩٩٧م، إلى وجود اختلافات جوهرية باختلاف المناطق، فقد أفادت نسبة قدرها ٥,٦٪ من الطلاب في الولايات المتحدة وكندا وعددهم ٢٧,٩٨٨ شخصاً من الذين طبق عليهم البحث، أنهم لم يستخدموا الحاسوب قط، في حين أفادت نسبة قدرها ٢٤,٥٪ ممن شملهم البحث في إفريقيا — وكان عددهم ١,٦٥٠ شخصاً — بالأمر ذاته (Taylor, Jamieson, & Eignor, 2000). وتجدر الإشارة في هذا المقام، إلى أن هذه الدراسة المسحية قد أجريت على أولئك الذين كانوا ينوون الانخراط في الدراسات الجامعية في الولايات المتحدة أو كندا، وهو ما يعلل الرغبة في تعميم النتائج على مجموع المعنيين في أي من المنطقتين. وعلى الرغم من أن بعضهم يرى في الأفق مجالاً للانتشار السريع للتقنية واتساع استخدامها، فإن التوقعات تشير إلى بقاء الاختلاف القائم بين المناطق في المستقبل المنظور، وهو توقع تدعمه مصادر البيانات على مستوى العالم (GeoHive, 2005).

وعلى الرغم من الأهمية المحتملة لمسألة التفاضل في عنصر الصدق، فإنه لم تُقِمْ — حتى الآن— دراسة واحدة من الدراسات التي تعنى باختبارات اللغة الثانية، بدراسة مباشرة عما إذا كانت الخبرة السابقة بالحاسوبات تؤثر — أو لا تؤثر — في الأداء في اختبارات اللغة الثانية المحوسبة. غير أن دراسة مدى الألفة بالحاسوب، التي أُجريت في إطار اختبار "التوفل"، قد وفرت بعض البيانات وثيقة الصلة بالمشكلة (Taylor *et al.*, 1999)، فالطلاب في تلك الدراسة قُسموا إلى مجموعتين على أساس إجاباتهم عن أسئلة استبيان طلب منهم فيه تقييم مستوى معرفتهم بالحاسوب، ومقدار استخدامهم لتطبيقاته المختلفة؛ كتطبيقات معالجة النصوص والإنترنت، ثم قورنت نتائج المجموعتين للوقوف على مدى تأثير مدى الألفة بالحاسوب في الأداء من خلال الاختبار المحوسب، علمًا بأن الباحثين أخذوا — في أثناء المقارنة بين المجموعتين — بالنتائج المحرزة في اختبار "التوفل" المقدم ورقيًا، بغية الوقوف على مستويات القدرات اللغوية لدى قياسها بعيدًا عن تأثير

الحاسوب باعتباره وسيلة في القياس. وبعد الوقوف على القدرات اللغوية (بقياسها من خلال اختبار "التوفل" الورقي)، لم يقف الباحثون على اختلاف ذي دلالة بين أداء الطلاب ذوي الألفة بالحاسوب وأقرانهم ممن ليس لديهم ألفة به في الاختبار المحوسب. وظاهر الأمر أن الدراسة تفيد بأن مباعث القلق بشأن التفاضل في عنصر الصدق، ليست بالأمر الذي يؤبه له كثيراً كما قد يظنه المرء لأول وهلة.

بيد أن الهدف الحقيقي من وراء دراسة الألفة بالحاسوب في إطار اختبار "التوفل"، هو تحديد ما إذا كانت الاختلافات العملية بيّنة بين المجموعات بعد أن تلقى أفرادها تعليمات لما يقرب من ساعة حول كيفية استخدام الحاسوب. وبمعنى آخر، هدفت الدراسة إلى الوقوف على ما إذا كانت الآثار المحتملة للخبرة السابقة – على اختلافها – يمكن تخفيضها لأدنى حدٍّ من خلال استكمال توجيهات على الشبكة تُتاح في مراكز الاختبار المخصصة لأغراض الاختبارات العملية (Jamieson, Taylor, Kirsch & Eignor, 1998). ولم يكن تصميم البحث موجهاً إلى جمع بيانات عن الجانب النظري لقضية التفاضل في عنصر الصدق، بل كان توجيهه بفعل الحقيقة العملية المتمثلة في تحويل اختبارات "التوفل" إلى اختبارات محوسبة، وهو قرار سابق في وقته على وقت إجراء البحث. وعلى الرغم من أن الدراسة لم تتعرض بالبحث لقضية المزية المحتملة للخبرة السابقة بالحاسوب، فإن تصميم البحث – بطبيعته العملية – يتيح نظرة عميقة بعض الشيء في الغياب المحير للدراسات المقارنة والدراسات المعنية بتناول التفاضل في عنصر الصدق. وتعكس هذه الدراسة حقيقة اجتماعية مفادها أن اتخاذ القرارات بتقديم الاختبارات في صورة محوسبة إنما ترجع – كما أسلفنا في الفصل الثاني – إلى أسباب أخرى غير نتائج البحوث التي تؤكد عدم وجود اختلافات بين الأداء في الاختبارات المحوسبة وغير المحوسبة؛ فالتقنية هي الاختيار المرجَّح باعتبارها الوسط الواضح البيّن لاستحداث صور أخرى من التقييم، ويمكن للمرء أن يتأمل في تأثر الأداء كثيرًا – أو عدم تأثره –

جرّاء ذلك. ولكن إلى أن تقدم البحوث مقارنات بين مستويات الأداء في اختبارات اللغة المحوسبة بين الماهرين والمبتدئين في التعامل مع الحاسوب، يظل الأمر في الغموض إزاء إمكانية دعم هذا التأمل بمعطيات من التجربة العملية. وبالنظر إلى الانتشار الهائل لاستخدامات التقنية، فإن العثور على اختلافات، لا يعني بالضرورة النظر إلى اختبارات اللغة المحوسبة بعين الشك والريبة، بل يظل الأمر مثار اهتمام وتفكير من المنظور العلمي الداعي إلى تفهم الأداء في الاختبارات المحوسبة.

تفسير تأثيرات طريقة الاختبار بالحاسوب

ثمة نهج آخر لبحث مستويات الأداء في الاختبارات المحوسبة، سواء أثبت تأثير تقديم الاختبارات اللغوية بالحاسوب على النتائج من الناحية الإحصائية أم لا، ويكون ذلك بفحص عملية الاختبار ذاتها لدى الطلاب الذين ينهون اختبار اللغة (Cohen, 1998)، ذلك أن هذا النوع من البحث يتناول – من منظور كيفي – الفكرة القائلة بأن الاختبار المحوسب قد يؤثر في "العمليات الذهنية المطلوبة للاستجابة للبند على نحو صحيح" حسبما يرى واينر وآخرون (Wainer *et al*, 2000). وفي مراجعة بحثية تناولت اختبارات القراءة المحوسبة، توصل ساواكي (Sawaki, 2001) إلى أن مباشرة بحوث كهذه تقتضي "استحداث الإجراءات البحثية المناسبة وتضمينها في دراسات عملية، وأن أساليب بعينها مثل: تحليل حركة العين، وتحليل البروتوكول الشفهي، إلى جانب مقابلات تالية على ذلك، وتقديم استبانات في الشأن نفسه، تعتبر كلها مفيدة لهذا الغرض" (ص ٥١). وتمضي الباحثة (Sawaki) موضحة أن هذا النوع من البحث يُباشَر فيما يتصل بالبحث في العوامل البشرية (كتصميم واجهات الإعلام البحثية) وبحوث النصوص التشعبية، لكن هذه الدراسات لا وجود لها في أدبيات البحوث التي تعنى باختبارات اللغة الثانية (Sawaki, 2001). هذا لا يعني أن هذا النوع من البحوث لا يجرى من الأصل، بل المقصد أنها لا تُصمّم ولا تُوحّد على نحو يساعد

ويؤدي إلى نشرها في الأدبيات البحثية العامة، وإنما استخدمت كطرق بحثية في بحوث أخرى، كما تستخدم في الدراسات الاستكشافية المبدئية ضيقة النطاق في أثناء تصميم الاختبار وإعداده، ولذا لا تُعد أو تعرف طريقها إلى النشر.

وربما لو كان المقصد مقتصرًا على الوقوف عند أهمية النتائج المفصلة للدراسات بغية توضيح الاختلافات بين اختبارات اللغة المحوسبة والصور الأخرى من الاختبارات، لكان من الضروري الوقوف على الاختلافات الإحصائية في نتائج الاختبارات. ووفقًا لما تمخض عن مراجعات ساواكي وشالوب – ديفيل وديفيلز (Chalhoub-Deville and Deville's, 1999) للأبحاث المعنية باختبارات اللغة المحوسبة – وعلى الرغم من كثرة المشاريع في هذه الاختبارات فإننا لم نقف على بحث منشور حاول دراسة الجوانب الخاصة بالمقارنات بين النتائج. وعلى ذلك انتهت بحوث الأخيرين إلى أن "البحث في مجال اللغة الثانية ما زال شحيحاً فيما يتعلق بالمقارنة بين نتائج الاختبارات الورقية والمحوسبة" (١٩٩٩م، ص ٢٨٢). وهذه النتيجة قد تنطوي على بشائر بقرب إجراء المقارنات المنشودة، بل إن بعض الدراسات التي نحت هذا المنحى قد خرجت للنور منذ ذلك الحين. لكن مع الأخذ في الحسبان أن اختبارات اللغة الثانية بالحاسوب، التي ما برحت تُجرى منذ عقدين على الأقل، فإنه لا بد لنا من التساؤل عن السبب في محدودية الدور الذي يؤديه هذا النوع من البحوث في جداول أعمال معدي الاختبارات.

يبدو – كما أسلفنا – أن برامج الاختبارات لا تقرر بالقبول أو الرفض استحداث اختبار محوسب على أساس نتائج البحوث المعنية باختلافات الأداء، وبناء على ذلك يمكن القول: ما المستفاد من تلك النتائج؟ تستهل ساواكي ملخص بحثها بمقدمة منطقية تفيد بأن "وجود أثر (كبير) لطريقة الاختبار على الأداء في اختبارات القراءة، من شأنه إبطال التفسير المقدم بشأن نتائج اختبارات القراءة المحوسبة على نحو يدعو للقلق" (Sawaki, 2001، ص ٣٨). ولا شك أن هذا المنظور متوافق مع البحث في آثار طريقة

الاختبار في اختبارات اللغة ؛ فهو قائم على الافتراض بأن المكون محل القياس (كالقدرة على القراءة) يمكن صياغته صياغة مفاهيمية باعتباره سمة قائمة بذاتها. وبمعنى آخر، يهتم مستخدمو الاختبارات بقدرة الطلاب على قراءة أي شيء في أي مكان، وينبغي استنتاج هذه القدرة على أساس الأداء في الاختبار، وهو ما يقتضي ألاّ يشتمل مكون القراءة على أمور أخرى كالقدرة على التنقل في النصوص التشعبية على سبيل المثال ؛ لأن هذا التنقل لا يدخل ضمن مقاييس القدرة على القراءة. ومن ثمّ سنبيّن في الفصل الأخير أن الفائدة التي تستقى من منظور السمة القائمة بذاتها، والمكون "النقي" إنما تتسم بالمحدودية البالغة في الكثير من الأغراض والسياقات الاختبارية محل الاهتمام لدى مستخدمي الاختبارات. وعلى ذلك نعلم أنه بغية فهم المستخدمين المذكورين للآفاق الواعدة وكذا المخاوف التي تنطوي عليها اختبارات اللغة المحوسبة، فإنه لا بد لهم من الإحاطة الكاملة بكيفية فهم مكونات اللغة وإدراكها من منظورات متعددة، علاوة على كيفية تقاطع التقنية مع تعريف المكون وتحديد جوانبه.

أنواع المهام الجديدة

ثمة مجموعة أخرى من المخاوف المحتملة لصدق الاستدلالات المستقاة من اختبارات اللغة المحوسبة واستخداماتها، وهي مخاوف ذات صلة بقضايا محتوى الاختبار، تنجم عن القيود التي تفرضها هذه الوسيلة على أنواع المهام التي تقدم في الاختبار أو بنوده. وليست هذه مشكلة مقتصرة على الاختبارات المعتمدة على الحاسوب، وإنما تتحدد بالتفاصيل التي يمكن لمصممي الاختبار عملها أو العجز عن عملها من خلال الوسيلة المستخدمة لبناء مهامه. فعلى سبيل المثال، يشتمل اختبار Purdue ITA لاختبار مهارة الكلام، على الكثير من المهام المختلفة التي يكملها الطالب، بما في ذلك التحدث وفقًا لأوامر نصية مكتوبة، وكذلك الاستماع ثم الاستجابة إلى متحدث يظهر

على شريط فيديو؛ أي أن الاختبار يقدم مجموعة هائلة مثيرة من المهام، إلا أنها مقيدة بفعل اختيار الحاسوب وسيلة لتقديمها. ومن المهام التي قد يود المرء مراعاتها – لكنها تبقى دون إسهام في وسط الاختبار – تقديم درس محدود أو الإسهام في جلسة محاكاة لساعات العمل المكتبي.

وعلى صعيد أدق مما سبق، يمكن للمرء التساؤل بشأن كيفية تأثير تفاصيل الشاشة وواجهة المستخدم في الأسلوب المتبع لدى معدي الاختبار في وضع بنوده من جانب، ولدى الطلاب عند تفسيرها من جانب آخر. ولنضرب مثالاً على ذلك ببنْد في القراءة في اختبار "التوفل"، حيث يطلب من الطلاب تحديد موقع لجملة تُضاف إلى الفقرة بما يتيح للطالب مطالعة فقرة مؤلفة من تسع جمل، وإذا كانت الفقرة أطول، لاضطر معدو الاختبار إلى اختصارها لتناسب حجم الشاشة في حال تقديمها بحجم الخط نفسه. أضف إلى ذلك أن هذه المشكلة ذاتها – حجم الخط – في النصوص المطولة، توجد في أي اختبار للقراءة، لكن حدود المساحة يُلتَزم بها عندما يعرض النص على شاشة الحاسوب ليس إلا. وبناء على ذلك، نجد أن اختبار القراءة المعد لقياس القدرة على فهم النصوص أو سبرها لغرض معين، يستلزم طرح نص مطول، الأمر الذي يُلزِم الطالب بالقفز السريع أو التنقل من صفحة إلى أخرى، وفي كلتا الحالتين لا يرى الطالب مطلقًا النص بالكامل على الشاشة. إن القيود المحددة عادة ما تقتضيها قرارات استخدام البرمجيات والتصاميم المُتخَذة على مستوى أعلى من أي بند بعينه في الاختبار؛ فحجم الخط، على سبيل المثال، عادة ما يُختار للخروج بمنظر معين يحقق التناسق بين جوانب البرمجيات المستخدمة. خلاصة القول: إنه لا بد من اتخاذ مجموعة من القرارات الأخرى خلال عملية إعداد الاختبار وكتابته، ويدخل في ذلك تفاصيل مثل: وضع الصور، والوصول إلى البنود الصوتية، ومقدار ما يعرض على الشاشة، إلى جانب الوصول إلى النصوص والبنود السابقة. ويكمن الحل الأمثل

لطرق مثل هذه المشكلات، في فحص التأثير الناجم عن أنواع بعينها من البنود على إستراتيجيات الطلاب وأدائهم، ويحدث كل هذا من خلال البحوث الكيفية.

القصور الناتج عن الاختيار التكيّفي للبنود

ظل الاختبار الحاسوبي المتكيف أبزر صور اختبارات اللغة المحوسبة على مدار العقدين الماضيين، ولذا فقد حظي ببعض البحث والدراسة. وقد أشرنا في الفصل الثاني أثناء حديثنا عن اختبار تحديد المستوى (ACT ESL) إلى أن الاختيار التكيفي لبنود الاختبار، يسفر عن اختبار أقصر مما يمكن تقديمه دون خاصية التكيّف، هذا مع افتراض نسبة واحدة من الثبات في الاختبار التكيفي وغيره من الاختبارات. وفي الوقت ذاته، طُرِحت أسئلة حول الأثر الناجم عن ترك مسألة الاختيار التكيفي لبرنامج حاسوبي يختار البنود على أساس مستوى صعوبتها. ومنذ شهد العالم باكورة الاختبارات الحاسوبية المتكيفة، أشار كانال (Canale, 1986) إلى احتمال أن يكون أنموذج التكيف بالحاسوب "مبسِّطًا خطرًا" وأنموذجا "اختزاليًا" (Canale, 1986، ص ٥-٣٤). وقد لجأ أول من أعد الاختبارات الحاسوبية المتكيفة ، إلى جمع حصيلة كبيرة من البنود من خلال أخذ عينات من جوانب عديدة للمكون محل الاهتمام، لكن اختيرت هذه البنود فيما بعد من هذه الحصيلة لأي اختبار آخر على أساس من الخصائص الإحصائية فحسب ؛ أي دون النظر إلى خضوع الطالب لاختبار جُمِعت بنوده على النحو المناسب من محتوى مناسب أم لا. وإذا تُرك محتوى اختبار كهذا للصدفة، فقد يلجأ الطالب في اختبار قواعد الإنجليزية كلغة ثانية —على سبيل المثال— المخصص للمستوى المتقدم، إلى اختيار كل البنود الهادفة إلى استبيان معرفة الطالب بالجمل الاسمية على غرار المثال الوارد في الشكل رقم (٣,١).

وعلى الرغم من أن ترتيب الكلمات في الجمل الاسمية المعقدة يعد بنداً مناسبًا لقياس المعرفة بقواعد اللغة في المستوى المتقدم، فإن اختبار قواعد الإنجليزية، لا بد أن

يشتمل على عينات تراعي تمثيل مجموعة من التراكيب المختلفة حتى تكون نتائج الاختبار بمثابة مؤشرات صادقة للمعرفة بالقواعد. وهنا نجد أن بند "...Male lion" الوارد في الشكل المذكور، مقدم على نحو يبيّن اتجاه اختبار ACT ESL إزاء مشكلة محتوى الاختبار في الاختبارات التكيفية، ألا وهو: بدلاً من ترك الاختيار إلى لوغاريثم انتقاء البنود ليختار كل بند على حدة، تُجمَع البنود بحيث تُقدَّم خيارات الجمل الاسمية في المستوى المتقدم في صورة نص مصحوب ببنود تخص الأفعال المبنية للمجهول في المستوى المتقدم أيضًا، وكذا علامات الترقيم، والجمل الاعتراضية، والروابط كما هو موضح في الفصل الثاني. وبتقديم البنود في صورة مجموعات (حزم) — أو "مجموعات النصوص والأسئلة" السابق ذكرها — (Wainer *et al.*, 2000)، يتسنى لمصمم الاختبار مزيد من التحكم في أسلوب تمثيل المحتوى.

وثمة نهج آخر للتعاطي مع هذه المشكلة ممثلاً في ترميز هذه البنود كل على حدة في بنك مصحوباً بمعلومات حول المحتوى، مع توجيه لوغاريثم انتقاء البنود لاختيار بنود الاختبار على أساس المحتوى والخصائص الإحصائية. غير أنه في ضوء الرغبة في صياغة بنود اختبارات اللغة من نص مترابط — بدلا من الاعتماد على كلمات مفردة أو جمل غير مترابطة — يبدو أن التوجه نحو مجموعات النصوص والأسئلة، يمثل أنجع السبل المنشودة، ومن هنا كان هذا المجال مجالاً خصبًا للبحث والدراسة. وتوضح التقارير الأولى بشأن الاختبارات الحاسوبية المتكيفة لقياس القدرة على القراءة؛ أن سبب المشكلة يعود إلى استخدام بنود مستقلة في الاختبار، ومن ذلك ما أفاد به مادسن (Madsen, 1991)، بشأن اختبار الإنجليزية كلغة ثانية في صورة اختبارات تكيفية اشتملت على بنود تطلب من الطلاب قراءة فقرة تتكون تتراوح ما بين جملة إلى ثلاث جمل، لتعينهم على إكمال عدد مناسب من البنود في وقت قصير، وبما يتيح التعاطي مع افتراضات القياس النفسي الخاصة بلوغاريثم انتقاء البنود. وقد تمت تلبية متطلبات هذه

الافتراضات، وإن كان ذلك على حساب التقييم المفترض لبعض أوجه مكون القراءة والفهم، ومن تلك الأوجه: الكفاية البلاغية أو فهم الفكرة الرئيسة للنص، ويعد ذلك نوعاً واحداً للنزعة الاختزالية التي اهتم بها كانال (Canale, 1986).

Male lions are also guilty of what (4)————not very kingly behavior.

A. would we probably call

B. we would probably call

C. we would probably call it

D. would we probably called

الشكل رقم (١,٣). بند من اختبار ACT ESL في قواعد الإنجليزية.

(المصدر: http://www.act.org/esl/sample/grammar4.html)

شاع استخدام الاختبارات التكيفية شيوعًا يكفي لتناولها بالبحث والدراسة، وجاء ذلك بالقدر الذي أتاح تحديد المخاوف فيما يتعلق بعنصر الصدق في نتائج الاختبار بعيدًا عن المخاوف المتعلقة بمحتوى الاختبارات. وثمة ملاحظات أخرى بارزة لعل أظهرها قلق الطلاب عندما يبدو لهم — من وجهة نظرهم — قدر من الصعوبة المستمرة في انتقاء البنود في ظل ظروف الاختبار التي يتعذر معها مراجعة استجاباتهم وتغييرها إن أرادوا. وتبدو هذه الملاحظة على الطرف النقيض من الدعاوى التي يسوقها مناصرو الاختبارات الحاسوبية المتكيفة، ومفادها أنه ينبغي للطلاب الشعور بالرضا تجاه اختيار البنود المناسبة لمستوياتهم. وفي الدراسات المتعلقة باختبارات المفردات في اللغة الأولى (LI)، أقدم الباحثون على استكشاف آثار القلق وعنصر الوقت في الاختبار ومستويات الأداء المرتبطة بالتباينات في أحوال الطلاب كما تظهر على الصيغة القياسية للاختبارات الحاسوبية المتكيفة. وقد شملت التجارب على وجه الخصوص اختبارًا ذاتي التكيف (يتطلب من الطلاب أن يقرروا بأنفسهم ما إذا كان كل

بند مشابهاً لما اختاروه، أو أصعب من سابقه أو أسهل منه)، وكذلك اختبارات تقدم تعقيبات عن كل استجابة من استجابات الطالب. وفي كلا النوعين من الاختبارات، أسفر التعقيب عن كل بند عن استغراق أوقات أقل في الاختبارات الحاسوبية المتكيفة والاختبارات ذاتية التكيف مقارنة بالاختبارات الخالية من التعقيب، ويرجع السبب في ذلك — افتراضًا — إلى ما يتمخض عنه التعقيب من تحفيز باعثٍ على المزيد من التركيز (Vispoel, 1998). غير أن دراسة أخرى أفادت باتجاه الطلاب إلى تحسين إجاباتهم عندما تسنح الفرصة بذلك من خلال توفير إمكانية المراجعة، وأكدت الدراسة ارتفاع مستوى الرضا لدى الطلاب عند توافر هذه الخاصية (Vispoel, Hendrickson & Bleiler, 2000). وهكذا نرى أن هذه المحاولات الرامية إلى فهم أفضل وتعامل أدق مع المخاوف التي تكتنف عنصر الصدق في الاختبارات الحاسوبية المتكيفة إنما تسعى إلى استحداث سبل تتيح تعديل النماذج التقليدية السائدة لأسلوب القياس بغية الاستجابة لمتطلبات المحتوى الخاضع للاختبار من جانب، ومراعاة لمشاعر الطلاب من جانب آخر.

عدم الدقة في تصحيح الاستجابات إلكترونيًا

أحد المخاوف لعنصر الصدق لاختبارات اللغة المحوسبة يتجلى في الاختبارات التي تطلب من الطلاب الإجابة عن مهمة أكثر تعقيدًا. وقد أشير في الفصل الثاني إلى المهام المعقدة التي يمكن تضمينها في الاختبار عندما يمكن الاعتماد على تحليل الإجابات بالحاسوب من أجل تصحيح الاستجابات بدقة. ومن أمثلة ذلك مهمة إكمال الجدول المضمّنة في مكون القراءة باختبار "التوفل"، وهناك أمثلة أخرى كمهام كتابة المقال والتحدث في قسم الكتابة التحليلي في إطار اختبارات GRE سالفة الذكر وSET-10 على الترتيب (Ordinate Corporation, 2002b). وبالحديث عن عنصر القراءة في اختبار "التوفل"، يوضح هذا العنصر مهمة تتيح عددًا كبيرًا — ومحدودًا في الوقت ذاته — من

الترتيبات في عناصر الجدول، في حين يتطلب عنصر الكتابة في اختبارات GRE وSET-10 إجابات ذات تراكيب لغوية، الأمر الذي يتطلب – ولو نظريًا – عدداً غير محدود من الاستجابات. وفي كلتا الحالتين، تكون استجابة الطالب متسمة بالتعقيد، إذ يتمثل المخوف المحتمل لعنصر الصدق في إمكانية إخفاق البرنامج الحاسوبي لتسجيل الدرجات، في تقييم الخصائص المهمة ذات الصلة بالاستجابة، وقد يترتب على ذلك إعطاء درجة/علامة أعلى أو أقل مما ينبغي، أو تسجيل معلومات تشخيصية خاطئة عن الطالب. وهناك دراسات قليلة شرعت في توضيح التعقيد الذي يكتنف استحداث لوغاريثم يستند لأسس معقولة من أجل تصحيح استجابات كتابية مصوغة على نحو لا يخلو من التعقيد.

الاستجابات غير اللغوية المعقدة

في دراسة لمهمة كان الهدف منها تقييم القدرة على التعرف على بنية النص – وهي التي عرفها كل من آلدرسون وبيرزيش وزابو (Alderson, Percsich, and Szabo, 2000) على أنها جانب من جوانب القدرة على القراءة – تبين أنه لا بد من اتخاذ قرارات بشأن أفضل طريقة لتصحيح مهمة ذات نصوص متتابعة. ومن المعلوم أن هذا النوع من المهام يقتضي من الطلاب ترتيب الجمل ترتيبًا صحيحًا في نص أُخرجت جمله مبعثرة وغير مرتبة. وهذه المهمة – التي تضيف كثيرًا لأنواع المهام المعقدة التي يمكن تضمينها في اختبارات اللغة – تنطوي على احتمال بإيجاد عدد كبير – ومحدود في الوقت ذاته – من الإجابات المختلفة. وحتى إذا كانت استجابة واحدة (تتابع إحدى الجمل) يعتبرها واضعو الاختبار أنموذجا للإجابة الصحيحة تمامًا، إلا إن بقية الاستجابات الأخرى لا تكون بالضرورة خاطئة تمامًا. ويجوز أن يرتب الطلاب مثلاً خمس جمل من أصل ست جمل ترتيبًا صحيحًا، أو ترتيب اثنتين منها فقط ترتيبًا صحيحًا بعد التعرف على ملمح ترابطي بينهما. لذا يرى الباحثون أن عملية تصحيح مهمة كهذه تصحيحاً ثنائي الخيار (أي: صفر للإجابة الخاطئة، وواحد للإجابة الصحيحة) تخفق في استجلاء جوانب

المكون بدقة، ومن هنا تبرز الحاجة إلى أسلوب صحيح أكثر دقة يأخذ بأسلوب متعدد الخيارات (أي يتضمن مجموعة من أساليب التصحيح).

وما كان للباحثين إلا أن يقرروا بشأن أفضل لوغاريثم حاسوبي يفضي إلى أساليب متعددة للتصحيح لتقييم الاستجابات في المهام المتتابعة، وقد تسنى لهم ذلك من خلال فحص أوجه الارتباط بين تصحيح المهام ذات النصوص المتتابعة (أي المحسوبة بأساليب عديدة ومختلفة) من جانب، والنتائج المسجلة في اختبارات لغوية أخرى من جانب آخر، وهو ما انتهى بهم إلى نتيجة عامة مفادها أن المهام التي صححت بأسلوب ثنائي الخيار، أسفرت عن نتائج ارتبطت بمقدار أكبر مع اختبارات اللغة الأخرى، وإن كان ارتباطاً محدودًا. ولكن هل يعني هذا أن العناصر المقيّمة بأسلوب ثنائي الخيار أفضل مما سواها بعد كل هذا؟ الحق أن المؤلفين لم ينتهوا إلى هذه النتيجة، غير أنهم أوضحوا أن "الارتباط وحده ليس كافياً لإصدار حكم عام وشافٍ" (ص ٤٤٣)، ويرجع السبب في ذلك – جزئيًا – إلى أن الاختبارات معيارية المحك غير قادرة لإصدار حكم وافٍ. بيد أنهم انتهوا في بحوثهم إلى نقطة متعلقة بالمكون محل الاختبار، وهي النقطة التي شكلت منطلقًا للبحث في هذا الجانب من الأساس: "المحصلات القليلة في النتائج ذات التوافق الكامل [أي: الأسلوب ثنائي الخيار] وحدها لا تعكس بالضرورة نقصًا في القدرة على رصد التماسك في النص، ولذا يؤخذ بالإجراءات ذات القبول الجزئي" (صفحة ٤٤٣). وبناء على ذلك، فقد أوضحت هذه الدراسة مدى التعقيد الذي يكتنف اتخاذ قرار بشأن كيفية تصحيح الاستجابات ومن ثم كيفية تقييم طرائق التصحيح هذه. ويصب الحل في هذه المسألة في صالح توقعاتنا بأن ما نتنبأ به سيكون محلا متناميًا لبحوث مماثلة، ومن ذلك أن تقييم التحليلات التفصيلية المعقدة متعذر التحقيق من خلال الربط بين النتائج المتحصل عليها بهذه الطرائق الحساسة، وبين تلك المتحصل عليها بإجراءات أكثر بساطة، ومن هنا تبرز الحاجة إلى وسائل أخرى للتقييم.

الاستجابات اللغوية

يتجلى مدى تعقيد هذه القضايا أكثر وأكثر في اختبارات اللغة التي تطلب من الطلاب الاستجابة لبنود الاختبار استجابات لغوية. وتجدر الإشارة في هذا المقام إلى أن المحاولات التي استهدفت إنتاج برنامج حاسوبي لتقييم مقالات مكتوبة بالإنجليزية بوصفها كلغة أولى، من خلال تقييم الخصائص اللغوية تقييمًا كميًا، لم تلق سوى نجاح محدود للغاية خلال القرن العشرين (Wresch, 1993). وبالمثل، عندما طُرِق هذا المسلك بغية تطبيقه على مكون الكتابة في اختبارات الإنجليزية بوصفها كلغة ثانية، جاءت النتائج مخيبة للآمال؛ شأنها في ذلك شأن المحاولات سالفة الذكر. ومن الدراسات التي تناولت هذا الجانب دراسةٌ بحثت استخدام برنامج لتحليل النصوص يحمل اسم "Writer's Workbench" (معمل الكاتب) لتقييم مقالات المتعلمين الذين يتعلمون الإنجليزية بوصفها كلغة ثانية، وانتهت الدراسة إلى أن الإجراءات الكمية (طول المقال، ومعدل طول الكلمات، ومقروئية النص بمقياس كنكيد (Kincaid readability)، ونسبة الجمل المعقدة إلى غيرها، ونسبة كلمات المحتوى ارتبطت ارتباطاً إيجابيًا مع النتائج العامة ذات الصلة بمكونات اختبارات الإنجليزية بوصفها كلغة ثانية (Reid, 1986). لكن الارتباط تراوح ما بين ٥٧, لعنصر طول المقال إلى ١٥, فيما يخص كلمات المحتوى، علمًا بأن هذه الأوجه من الارتباط لا تبرر استخدام هذه الطرائق الخاصة بالتقييم والتصحيح الحاسوبيين لهذا الغرض، وذلك بالقدر الذي تعتبر به النتيجة العامة دلالة جيدة على إتقان الكتابة.

من جانب آخر، طبقت البحوث الحديثة — في التصحيح الآلي للمقالات التي كتبت باللغة الأولى — فهمًا أكثر رحابة وعمقًا فيما يتصل بالأداء في مكون الكتابة، جنباً إلى جنب، مع أساليب حديثة في معالجة اللغات الطبيعية. كما تناولت البحوث المعنية بدراسة جودة التقييمات جوانبَ أوسع من أوجه الارتباط مع التقييمات الشاملة. وفي هذا الإطار نجد أن "المقيّم الإلكتروني" — الذي سبقت مناقشته في الفصل السابق — يستنتج

نتيجة المقال من التقييم الذي يجريه لثلاثة عناصر من خصائص المقال هي: تركيب الجملة، والتنظيم (يُقاس من خلال خصائص الخطاب مثل التعبيرات التي تؤدي إلى تماسك الجمل) والمحتوى (يُقاس من خلال المفردات ذات الصلة بالموضوع المحدد في المقال)، أما القيم الموجودة في هذه المتغيرات (القيم التي ينبغي ارتباطها بنتيجة مخصوصة وشاملة لموضوع مقال بعينه) فتقاس على أساس المقالات المأخوذة كعينات بعد تقييمها على يد العنصر البشري. وقد تضمن تقييم برنامج التصحيح على أوجه ارتباط مع النتائج التي صححت على يد العنصر البشري (علمًا بأنها جاءت قوية في دلالاتها)، ليس هذا فحسب، بل جاء التقييم أيضًا من خلال المقالات المكتوبة على وجه التحديد لإحداث إخفاق متعمد في برنامج التصحيح. وأعقب ذلك استخدام خصائص المقالات المسببة لإخفاق برنامج التصحيح في الارتقاء به (Powers, Burstein, Chodorow, Fowles & Kukich, 2001).

تظهر المشكلة ذاتها — ولكن بمقياس آخر — في اختبارات اللغة التي تطلب إلى الطلاب تقديم استجابات بنيوية قصيرة، كجزء من كلمة مثلاً أو عبارة أو جملة واحدة. وقد درس الباحثون في الستينيات والسبعينيات من القرن المنصرم آثار طرائق التصحيح المختلفة في استجابات اختبارات التتمة المنتظمة في القراءة، عندما تكون هذه الاستجابات مؤلفة من كلمات مفردة يكتبها الطلاب في فراغات تتخلل النصوص. ويذكر أن هذه الأبحاث أُجريت قبل التوسع في استخدام الحواسيب في اختبارات اللغة، إذ كان الغرض الرئيس منها تبرير استخدام طريقة ثنائية الخيار في تصحيح الاستجابات المؤلفة من كلمات بعينها، وذلك من خلال إثبات مكافأتها لطرائق أكثر تعقيدًا من شأنها إلزام المقيّمين بإصدار أحكام عن مدى اقتراب استجابة الطالب من الكلمة الدقيقة المحذوفة. وفي هذا السياق، استهل أولر (Oller, 1979) ملخصه لنتائج هذا البحث في الأطروحة ذات الصلة بقوله: "كل طرائق التصحيح التي خضعت للدراسة تسفر عن قياسات ذات ارتباط عالٍ إلى حد بعيد" (ص ٣٦٧)، وينتهي إلى أنه "باستثناء أغراض البحوث

الخاصة، يبدو أن هناك القليل المنتظر من استخدام مقياس معقد لدرجات الملاءمة في تسجيل نتائج اختبارات التتمة المنتظمة في القراءة " (صفحة ٣٧٣). غير أن هذه النتيجة لم تحظ بقبول عام في أوساط الباحثين في اختبارات اللغة (انظر مثلاً Alderson, 1980). كما جاءت هذه النتيجة (استنادًا إلى دليل الارتباط) في وقت كان تحليل الاستجابات فيه على يد العنصر البشري، في حين كان التقييم على أساس دليل الارتباط بين نقاط البحوث. ولم يكن التحليل الآلي للاستجابات حينها خيارًا عمليًا، وربما كان هذا هو السبب في الذكر الهامشي لاحتمال التوسع في التفسيرات المحتملة المستقاة من نتائج الاختبارات من خلال التحليل التفصيلي. و يبدو أن هذه القضايا تستلزم التطرق إليها مرة أخرى لاستكشاف الاحتمالات التي تنطوي عليها تلك الاستجابات اللغوية القصيرة لكل من المتعلمين ومستخدمي الاختبارات.

شهدت الفترة الأخيرة أنواعًا مماثلة من البنود مضمّنة في مهام إكمال المكونات في اختبارات القراءة والاستماع في نظام WebLAS الذي سبق ذكره في الفصل الثاني، إلى جانب اختبار ملء الفراغات في اختبارات النظام ذاته. ومن المعلوم أن البحوث المعنية بتصحيح الاستجابات في الاختبارات المذكورة لم تُنشر بعد، لكن قليلاً من الدراسات تناولت تصحيح تلك البنود بأنواعها في مختلف الاختبارات، ومن هذه البحوث ما درس اختبارًا في قراءة الإنجليزية بوصفها لغة ثانية حيث طُلِب من الطلاب الإتيان بعبارات وجمل كاستجابات عن أسئلة مفتوحة من نصوص مخصصة للقراءة (Henning, Anbar, Helm & D'Arcy, 1993). وقد استخدم الباحثون برنامجًا حاسوبيًا لتعيين النتائج للأسئلة، مع تقييم جزئي للاستجابات التي خلطت بين الخطأ والصواب لسبب مفاده —كما يرى الدرسون وآخرون (٢٠٠٤م)— القلق من دقة طريقة التسجيل في الوقوف على مقدار المعرفة لدى الطلاب. كما حاول الباحثون تقييم ما إذا كانت الجهود المبذولة في تحليل الاستجابات، وتعيين النتائج والدرجات بصورة جزئية، قد

أتاحت لهم تقييم مكون الفهم بالقراءة على نحو مغاير مقارنة بحال أخرى هي تسجيل درجات الاستجابات بأسلوب ثنائي الخيار. وبناء على ذلك، درس الباحثون الاختلاف بين طريقتي التسجيل ثنائية الخيار ومتعددة الخيارات من خلال حساب معامل الارتباط في كل طريقة مع النتائج المحصلة من مجموعة من أسئلة الاختيار من متعدد تتعلق بنص واحد. وقد أسفرت النتائج المستمدة من الطريقة ثنائية الخيار والعناصر المفتوحة عن أوجه أوفر للارتباط مع نتائج اختبارات الاختيار من متعدد (٩٩,) مقارنة بما أسفرت عنه النتائج المسجلة بطريقة الخيارات المتعددة مع البنود المفتوحة (٨٩,). وفي ذلك تأكيد على أنه كلما سُجلت نتائج البنود المفتوحة بطريقة الخيارات المتعددة، فإن نتيجة الاختبار المترتبة على ذلك تكشف عن قدرة مختلفة بصورة ما عن النتيجة المسجلة لبنود الاختيار من متعدد، علمًا بأن هذه النتائج لا تتناول التساؤل بشأن أي الطريقتين للتسجيل تتمخض عن مستوى أفضل في اختبار القراءة قياسًا على الغرض المنشود منه، وإنما تؤكد على أن طريقة التسجيل قد تمخضت عن اختلاف يسير. ومن هذا المنطلق نجد أن هناك حاجة إلى طرق أخرى لتقييم الاستجابات أو العمليات التي ينتهجها الطلاب في الاستجابة بغية تسليط الضوء على فحوى هذه الاختلافات.

وفي السياق ذاته، ثمة دراسة في إدراك الاستجابة الآلية في مهمة اختبارية لمكون الإملاء، حيث تناولت المشكلة من منظور كيفي أكثر دقة من غيرها. وقدم كونيام (Coniam, 1998) مثالاً على كيفية قيام لوغاريثم التسجيل بتقييم عبارة كتلك التي كتبها الطالب في اختبار إملاء جاء فيها "which needed to be typed in"، مضيفًا أن الطالب الذي كتب "which are needing to be typing" يستحق درجة جزئية على هذه الاستجابة. وإذا كان برنامج التقييم يقوم بذلك، فالسؤال إذًا: ما مقدار الدرجة المستحقة للاستجابة على وجه التحديد؟ ولماذا؟ هل ينبغي أن تكون درجة الاستجابة مطابقة لمن كتب " that needs to be typed" أو "which needs typing" أو "which needle to eyed"؟ أدرك كونيام غياب

تبرير واضح لتخصيص درجات جزئية، مناديًا باستحداث لوغاريثم لتسجيل النتائج "بلا تحديد إلى حد ما: فعبارة "which are needing to be typing" تحصل على ٤٢٪ ، في حين تحصل عبارة "which are needly to be typest" على ٣٣٪ على الرغم من أن الأخيرة أكثر افتقارًا إلى الترابط اللغوي مقارنة بالأولى" (Coniam, 1998، ص ٤٤).

ولا شك أن توجهًا محدد التفاصيل لتقييم الاستجابات اللغوية يجب أن يعتمد على نظرية قائمة على المكون الذي أُعد الاختبار لقياسه. وهناك دراسة أخرى تناولت تفاصيل لوغاريثم تقييمي للاستجابات المقدمة على مكونات محددة، وقد كشفت هذه الدراسة عن الكيفية التي استخدمت بها النظرية الموضوعة لما طُلِب قياسه في تعيين القيم المحددة لاستجابات بعينها. وإضافة لما تقدم، أقدم كل من جيميسون وكامبل ونورفليت وبربيسادا (Jamieson, Campbell, Norfleet and Berbisada, 1993) على تصميم برنامج لتسجيل النتائج، بغية تقييم استجابات الطلاب بطريقة مطابقة لطريقة العنصر البشري، وعلى نحو يعكس القيمة النسبية للاستجابات المحتملة في ضوء نظرية المكون بصورة منتظمة. وقد تألفت الاستجابات اللغوية من ملاحظات الطلاب المأخوذة في أثناء قراءة نص ما، إلى جانب بروتوكولات الاستذكار الخاصة بهم، في حين تقرر إعطاء الدرجات العالية عند وجود معلومات كاملة عما كان في نص القراءة وحضور هذه المعلومات في ملاحظات الطلاب وما استذكروه. أما الدرجات المنخفضة، فمُنحت في حالات عدم اكتمال المعلومات، وعندما تكون ملاحظات الطلاب وما استذكروه مشتملاً على معلومات أقل في أهميتها من النص. وقد أكدت النتائج وجود معامل ترابط قوية بين الدرجات الممنوحة من جانب المقرِّر البشري وبين الدرجات الممنوحة من جانب طريقة التصحيح وتسجيل الدرجات والنتائج بالحاسوب. كما كان من المهم أيضًا إثبات الباحثين لعنصر الاتساق في مخرجات طريقة تسجيل الدرجات مع تعريفهم بعملية كتابة الملاحظات والاستدعاء في اختبار مكون القراءة.

تلكم الأمثلة القليلة من اختبارات اللغة الثانية، بدأت تشير إلى ما أوضحه الباحثون في مجال القياس التربوي: وهو أن "البحوث التي أجريت على عنصر الصدق في تقييم الاستجابات على الاختبارات بالحاسوب والاستجابات على عناصر في اختبارات المكونات اللغوية بالحاسوب تنطوي على عدد من الاعتبارات والتعقيدات التي يقل حضورها في التحقق من عنصر الصدق في الاختبارات المحوسبة التي تعتمد على أسلوب الاختيار من متعدد" (Williamson, Bejar and Hone, 1999). وتتجلى هذه القضية في أبرز صورها عندما يُقدِم الباحثون على تفصيل الأساس النظري (أو انعدامه) لتعيين النتائج والدرجات بصورة جزئية. وهذا التفصيل يبيّن ضرورة توفير تعريف أكثر دقة مما هو مطلوب من أجل تقييم الاستجابات بطريقة ثنائية الخيار والحكم عليها بالصواب أو الخطأ. علمًا بأن إصدار قرار بالصواب أو الخطأ، لا يتطلب سوى مطابقة الاستجابة بالأنموذج اللغوي المستهدف، وبذلك يطوِّق الأسئلة المفيدة بما يجعل الاستجابة صحيحة، وأي الاستجابات أكثر صحة من غيرها، والأسس التي يستند إليها واضع الاختبار عند اتخاذ ذلك القرار. وإذا كانت أكثر القضايا أهمية واعتبارًا تبدأ في الظهور حتى يتخذ واضع الاختبار قرارات بشأن صحة الإجابات من عدمها (كتسجيل الدرجات لحالات الخطأ الهجائي مثلا)، فإن هذه القضايا تتعاظم في أهميتها لدى التعاطي مع المهام التي سُجلت درجات إجاباتها اعتمادًا على طريقة الخيارات المتعددة.

إن المدخل إلى الاستفادة من تقنيات التعرف على اللغة في اختبارات اللغة الثانية (بدلاً من إحاطة هذه الاختبارات بالمخاوف منبثقة عن تلك التقنيات) يقتضي وضع مهام اختبارية تنبثق عنها درجات تتسم بالدقة وانضباط المعلومة في المكون محل القياس والاختبار. وهذا الهدف يتطلب فهمًا جليًا لطبيعة المهام الاختبارية، وكذا الوقوف على روابط واضحة بين نظرية المكون وطريقة تسجيل الدرجات والنتائج. غير أن المجموعة سالفة الذكر (Alderson, Percsich, and Szabo, 2000) أوضحت أن طرق التقييم المناسبة

– غير المقتصرة على معامل الارتباط – لا بد من توظيفها إذا توافرت النية للوقوف على قيمة طرق التسجيل. ويمكن تحديد الاتجاهات المطلوبة للطرق الإضافية للتحقق من عنصر الصدق من خلال التعريج على الأعمال القياسية في مسألة التحقق من عنصر الصدق (انظر مثلاً Messick, 1989) مع إيلاء اهتمام خاص بالطرق الكمية لتقييم إستراتيجيات الخضوع للاختبار وتقييمه.

الإخلال بعنصر تأمين وسرية الاختبار

من الآفاق الواعدة لاختبارات اللغة المحوسبة، ما أشير إليه في الفصل الثاني من اليسر وسهولة الوصول إليها عبر شبكة الويب. وفي هذا الصدد أوضح روفر (Roever) أن ميزة الويب لا تقتصر على تيسير تقديم الاختبارات للمتعلمين عند طلبها فحسب، بل إن "استخدام مقومات تسجيل الدرجات في البنود التي سُجلت درجاتها بطريقة الخيارات الثنائية، من الممكن أن يجعل النص مستقلاً عن الطالب استقلالاً كاملاً وأن يزيد من المرونة واليسر للطالب أكثر وأكثر" (Roever, 2001، ص ٨٨). بيد أن ما يزيد من المرونة واليسر للطلاب، ينطوي على مخاوف تتعلق بعنصر الصدق في نتائج الاختبارات، ويبدو أن هذا الأمر – حتى الآن – تعتريه إشكالية يتعذر تخطيها في الاختبارات المصيرية؛ لأن مستخدمي الدرجات بحاجة إلى طمأنتهم إلى أن الطالب هو الشخص نفسه الذي وُضعت له الدرجات. ومن المعلوم أن الاختبارات المصيرية – كاختبار "التوفل" – لا يمكن لها الاستفادة من هذه "المزية المنطقية الوحيدة على عظمها"، ولذلك فهي محتومة بتقديم الاختبارات في مراكز اختبار يوجد بها مراقبون قادرون على التحقق من هوية الطلاب. وعلى الرغم من أن الاختبارات المستخدمة لتحديد المستوى والتصنيف، لا تنطوي على هذا القدر من المصيرية، فإن ترك أجواء هذه الاختبارات مفتوحة للطالب لدخولها (وقتما شاء) ما يزال ينطوي على احتمالات

أكيدة بعدم صدقية مدلول النتيجة، ومن ثمّ لا ينبغي استخدام الدرجات للتصنيف، ولا لتحديد المستويات كذلك. أما عنصر الصدق في الاختبارات غير المصيرية، مثل: اختبار DIALANG سالف الذكر، فلا تكتنفه المخاوف جراء الأخذ بالمزية الواردة آنفًا؛ نظرًا لأن الطلاب يفتقرون إلى الحافز الذي يدعوهم إلى الغش عندما تكون الدرجات مصروفة لتقييم معلوماتهم فحسب. لكن الاختبارات غير المصيرية لا تشكل إلا جزءًا يسيرًا من الاختبارات قيد الدراسة لدى واضعي الاختبارات والمعنيين بنتائجها.

وثمة مشكلة أخرى تتعلق بعنصر الأمان والسرية وتختص بالاختبارات الحاسوبية المتكيفة، وهي المشكلة التي وصفها كل من واينر وإيجنور (انظر Wainer and Eignor, 2000، ص ٢٧٤) في الفصل الذي يحمل عنوان " Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing" (التحذيرات والمثالب والنتائج غير المتوقعة لتطبيق الاختبارات المحوسبة على نطاق واسع). وقد عُرضت إحدى هذه المثالب من خلال خبر يروي الإخلال بنسخة من الاختبارات التكيفية في إطار برنامج GRE الذي يعد أول برنامج واسع النطاق للاختبار في هيئة الاختبارات التعليمية، يُقدِم على تغيير شكل الاختبار الى اختبار حاسوبي متكيف. وفي عام ١٩٩٤م، كشفت منظمة أخرى عن مدى حساسية الأمر نظرًا لسهولة حفظ العناصر المأخوذة من بنك البنود ثم تمريرها على الطلاب اللاحقين. وعلى ذلك، أظهرت هذه التجربة ضرورة توافر بنك ذي بناء مخزون وفير من البنود، كما سلطت الضوء على أهمية الحاجة إلى النظر في قضايا الأمان والسرية ذات الصلة باستخدام التقنية، لاسيما لدى التعاطي مع الاختبارات المصيرية. لكن علماء المستقبليات – بدلا من التوسع في طرق تقنية للتحقق من الهوية – (انظر مثلاً Bennett, 2001) يتجهون للإشارة إلى استخدامات أخرى للاختبارات الأقل مصيرية وأهمية، وكذلك الاختبارات المنطوية على فائدة للمتعلمين، كما هي الحال في برنامج الاختبار الأوروبي DIALANG. ويبقى النظر في مدى تحقق الاحتمالات ذات

الصلة بالاختبارات غير المصيرية عبر شبكة الويب من خلال مجموعة من الخيارات في إطار أجواء مفيدة لمتعلمين من خلفيات متباينة، أو عدم تحققها.

التبعات السلبية

المخوف السادس المحتمل لاختبارات اللغة المحوسبة، يتأتّى مما قد يشكل تبعات أو آثاراً سلبية على المتعلمين وبرامج اللغة والمجتمع. ومن أنواع هذه التبعات السلبية، ما هو ماثل في إشارة المنتقدين إلى أن تكاليف هذه الاختبارات سيحول الأموال عن احتياجات أخرى مطلوبة في إطار البرنامج. ومن التبعات السلبية الأخرى، ما يُرى في مجموع آثار العملية التعليمية، إذا ركز المدرسون على أنواع من المهام التي تظهر في الاختبار من أجل إعداد الطلاب لأداء الاختبار المحوسب. وإذا كانت أعداد الطلاب محدودة نتيجة للقيود التقنية، كما سبق وأسلفنا، فإن المواد التعليمية قد يطالها الأمر ذاته، غير أن هذه التبعات لم توثق توثيقًا فعليًا، كما أن البحث في مجموع الاختبارات حول عنصر التدريس، يفيد بوجود علاقات أكثر تعقيدًا بين نوعي الأحداث التربوية (Alderson and Hamp-Lyons, 1996)، لكن الفكرة الرئيسة تتمثل في أن واضعي اختبارات اللغة المحوسبة والآخذين بنتائجها، ينبغي أن يكونوا على دراية باحتمال وقوع تبعات سلبية كهذه، وأن هذه التبعات جديرة بالبحث والدراسة.

وينبغي أن يسعى البحث في تبعات اختبارات اللغة المحوسبة، إلى توثيق بعض هذه المخاوف المطروحة، وإذا كانت الاختبارات الحاسوبية المتكيفة مسؤولة عن مقدار أوفر من القلق مقارنة بغيرها من الاختبارات، فإنه ينبغي توفير السبل المناسب لتوثيق مثل هذه الآثار السلبية. وإذا أبدى الطلاب — الذين يعرفون أن مقالاتهم ستُقيّم بلوغاريثم آلي مثل المقيّم الإلكتروني — اهتمامًا قليلاً بتعلم كيفية الكتابة، واهتمامًا كبيرًا بتعلم كيفية تسجيل البرامج الحاسوبية لدرجات مقالاتهم، فإنه ينبغي كذلك توفير السبيل الممكنة لتسجيل هذا الاهتمام. وبالمثل، إذا كان الطلاب المقبلون على تعلم اللغة (الإنجليزية مثلاً) غير راغبين

في الانتظام في الفصول، ويفضلون الجلوس في معامل الحاسوب لإكمال مهام اختبارية مماثلة لمهام اختبار "التوفل"، فإنه يمكن أيضًا توثيق ذلك كله من خلال دراسات كيفية وبحوث استطلاعية. ولا شك في أن هذا النوع من البحوث، سيشكل دليلاً على وجود تبعات سلبية لاختبارات اللغة المحوسبة.

كما ينبغي للآخذين بنتائج الاختبارات، تقليب الأمر أيضًا على وجهيه؛ أي بالنظر في الثمرات الإيجابية لاختبار اللغة المحوسبة. إن الباحثين في القياس التربوي مفعمون بالحماسة بشأن الآفاق الإيجابية للتقنية وما يُنتظر منها في النهوض بعملية التقييم. ومن هنا يشير بيكر (Baker, 1998) إلى التقنية على أنها "وافد قدير" وصل لإنقاذ النظام العاجز والمتخبط للاختبارات في المنظومة التعليمية بالولايات المتحدة. ونحن نعلم أن التقنية غير قادرة وحدها على إتمام ذلك، لكن الحل يكمن في الأسلوب الذي تقدم به التقنية الدعم لأغراض التوثيق والدراسة التي يُرجى منها الخروج بنظريات أفضل وأدق فيما يتعلق بالمعرفة والتعلم والنتائج التربوية في مختلف المناهج. كما يؤكد بينيت (Bennett, 2001) على الحاجة إلى تغيير منظومة التقييم التربوي تغييرًا جزئيًا؛ بغية الارتقاء بأثره في التعليم. وفي بحث تناول بالتحليل أوجه الترابط بين التقييم والتقنية والأعمال وغيرها من جوانب المجتمع على المستوى العالمي، يوضح الباحث نفسه بأسلوب واقعي، أن الأساس العلمي لهذا التغيير سيعقب التقنية من حيث ترتيب الأدوار؛ نظرًا لأن الاستثمار المالي والفكري الذي يوجهه المجتمع إلى التقنية. يفوق بكثير ما هو موجه إلى العلم والفن المأخوذ بهما في التقييم التربوي. "إذا كانت إسهامات العلوم الذهنية والقياسية أهم باعتبارات كثيرة من التقنية الحديثة، إلا أن التقنية الحديثة تعم مظاهر الحياة في مجتمعنا" (ص ١٩). وتعيد نظرة بينيت صياغة مفهوم التبعات (التي تتركها الاختبارات على الآخذين بنتائجها) في صورة أكثر دينامية ذات علاقة تبادلية بين كيفية تقديم الاختبارات من جانب، وكيفية صياغة المجتمع لأدوات التطور المستخدمة في الاختبارات من جانب آخر.

خاتمة

نخلص مما سبق إلى أن كل مخوف من المخاوف التي نوقشت وجاء ملخصاً في هذا الفصل كما هو موضح في الجدول رقم (٣,٢) يستحق الاهتمام والبحث. لكن استنادًا إلى البحوث التي بين أيدينا، فإننا لم نتمكن من العثور على أي دليل دامغ يفيد عدم القدرة على تجاوز تلك التحديات بالدرجة التي يُنظر بها إلى اختبارات اللغة المحوسبة بعين الشك والريبة دون غيرها من أنواع الاختبارات.

الجدول رقم (٣,٢). ملخص بالمخاوف المحتملة للصدق والحلول المقترحة لها.

المخوف المحتمل لعنصر الصدق	الحلول المقترحة
اختلاف معدلات الأداء في الاختيار	ينبغي للآخذين بنتائج الاختبارات النظر ما إذا كان هذا مهدداً حقيقيًا أم لا. وإذا كان ذلك مهدداً حقيقيًا، فلا بد من تناوله بالبحث لمقارنة الأداء في نماذج مستقبلية موازية.
أنواع المهام الجديدة	يحتاج تفسير الأداء في أنواع الاختبارات الجديدة إلى فحص كيفي وكمي بما يتيح استخدامها بصورة مناسبة.
القصور الناتج عن الاختيار التكيّفي للبنود	لا بد من استكشاف مجموعة متنوعة من أنواع التكيف وتناولها بالبحث والتقصي.
عدم الدقة في تصحيح الاستجابات إلكترونيًا	لا بد من توجيه جهود البحث والتطوير المنسقة إلى استحداث معايير مناسبة ومتعددة الجوانب لتقييم عملية التصحيح الآلي لمختلف الاستجابات على البنود الاختبارية للمكون محل القياس.
الإخلال بعنصر تأمين وسرية الاختبار	لا بد من أخذ عنصر الأمان في الاعتبار في ضوء مستوى التأمين المطلوب قياسًا على الغرض المطلوب من الاختبار قيد النظر.
التبعات السلبية	لا بد من فهم التبعات الإيجابية والسلبية المحتملة لاختبارات اللغة المحوسبة والتخطيط للتعامل معها، ثم توثيقها.

بل على النقيض من ذلك، فقد لاحظنا عددًا من المجالات التي كانت اختبارات اللغة المحوسبة سببًا في تناولها بالبحث والدراسة، وهو ما سيعود بالنفع على اختبارات اللغة بصفة عامة. كما أن فكرة استخدام التقنية في التقييم، ستدفع الباحثين فعلاً إلى دراسة عمليات الاختبار على نحو أكثر دقة وعناية، الأمر الذي يمكن أن يشكل فائدة تقدمها التقنية للإسهام في التعاطي مع التبعات المعقدة ذات الصلة بالتقييم اللغوي. ومن ثمّ، ينبغي النظر إلى أنواع البحوث المقترحة بوصفها حلولاً للمخاوف الوارد ذكرها في الجدول رقم (٣,٢) على أنها منطلق لعملية نشطة وضرورية للوقوف على التوجهات الرامية إلى دراسة القضايا ذات الصلة كما تتجلى في احتياجات الطلاب ومخاوف العامة.

وبصفة عامة، فإن قضايا الصدق المرتبطة باختبارات اللغة المحوسبة، تماثل نظيراتها في أي اختبار لغوي آخر، وقِوام ذلك أن الاستدلالات والاستخدامات لنتائج الاختبارات، بحاجة إلى ما يدعمها على مستوى النظرية والبحث، وذلك بما يتيح للآخذين بالنتائج معرفة مدلولات الدرجات ومجالات استخدامها. وفي الوقت ذاته، نجد أن المخاوف الخاصة لعنصر الصدق (وما يترتب عليها من قضايا، قد تتطلب التعامل معها من خلال البحوث في هذا الصدد) ترتبط ارتباطًا مباشرًا باستخدام التقنية، ولذا كان الرجوع إلى البحوث المعنية بعنصر تحقيق الصدق في اختبارات اللغة المحوسبة التي أثبتت بقاء الكثير من الأسئلة بلا إجابات، وطرحت أسئلة حول قلة الدراسات التي عنيت بهذا الشأن، وكلها بحوث حاولت الوصول إلى فهم أدق وأرحب لمدلولات نتائج اختبارات اللغة المحوسبة ، وكذا التبعات المترتبة على استخدامها. فهل سيتناول الباحثون في اختبارات اللغة الثانية مستقبلاً القضايا التي أوجزناها في هذا الفصل بالبحث والدراسة بما يتيح التعويل على أساس بّين ومتين لدى التعامل مع قضايا الصدق في اختبارات اللغة المحوسبة؟ أم أن البيئة المفعمة بشواهد التقنية – التي يحيا فيها الجيل التالي

من معدي الاختبارات والطلاب — ستغير من النظرة الحالية للكثير من المخاوف القائمة حاليًا حول عنصر الصدق؟ أيًا كان الخيار في المستقبل، فإن معدي هذه الاختبارات، هم بحاجة الى التمكن من تقييم مدى صدق أي اختبار قياسًا على الغرض الموضوع من أجله. ومن هنا سنوضح في الفصل الخامس القضايا المرتبطة بتقييم اختبارات اللغة المحوسبة، وهي قضايا تتجاوز ميدان النظر في المخاوف المتعلقة بعنصر الصدق. لكن المعرفة الدقيقة بهذه القضايا، تستلزم منا أولاً أن الانتقال — في الفصل التالي — إلى وصف المصادر والقيود التي تقدمها التقنية وتفرضها على معدي الاختبارات.