5

استخدام ثنائيي اللغة لتقييم التشابه بين صيغ لغوية مختلفة لاختبار ما

ستيفن ج. سيرسي

جامعة ماساشوستس في آمهيرست

لطالما واجه كل من العلماء المعنيين بقياس سرعة ودقة العمليات العقلية والباحثين التربويين والأطباء السريرين مشكلة تقييم الأفراد الذين يستعملون لغات مختلفة. وفي إطار هذه السياقات، فإن حقيقة وجود عالم متعدد اللغات يعيق استخدام وسيلة تقييم واحدة.

لهذا السبب عُدلت الاختبارات إجمالاً للاستخدام بأكثر من لغة واحدة. إلا أن عملية التكييف لسوء الحظ لا تضمن أن الصيغ اللغوية المتعددة لاختبار ما متعادلة. (انظر مثلاً: فان دي فيجفر و لونغ 2000). وهكذا، تكمن المشكلة الأساسية في تقييم التقاطع اللغوي في تجريد تأثيرات الاختبارات من تأثيرات المجموعة عندما تتم المقارنة بين المجموعات والأفراد الذين خضعوا لصيغ لغوية مختلفة لاختبار ما (لمناقشة مفصلة انظر هاملتون ودي جونغ 2003).

تم النقاش طويلاً أنه عندما يترجم اختبار ما أو (يكيف) من لغة إلى أخرى فإنه لا يمكن اعتبار صيغ اللغتين المختلفتين متعادلة. ويُعزى ذلك إلى الزيادة الحديثة في تقييم التقاطع اللغوي مثلا: بيللر، 1994 ، فوستر، أولسن، فورد و سيريسي، 1997، الجمعية العالمية لتقييم الإنجازات التعليمية، 1994، سيريسي،

إكسينغ و فيتز جيرالد، 1999 قد أعيد تأكيد هذه النقطة من قبل اختصاصيي اختبارات معاصرين عديدين (مثلاً: آنفوف وكوك، 1988، جيسينجر، 1994، هامبلتون،1993، 2002، أوليدو، 1981، بريتو، 1992، سيريسي، 1997، فان دي فيفر و تانزر، (1997). ولكن هذا التحفظ يكاد يكون قديماً قدم ممارسة الاختبارات نفسها. أدرك تيرمان (1916) عدم القدرة على مقارنة درجات النسخة الإنكليزية لستانفورد - بينيه مع التقييم الفرنسي الأصلي لبينية. وبشكل مماثل، حذر ليكرت (1932) في مقاله الإبداعي حول مقاييس النسب الإجمالية من استخدام مقاييس الإتجاهات عبر جماعات "ثقافية" مختلفة. على الرغم من ذلك، فإن العالم اليوم يصبح مكاناً أصغر وفعاليات تقييم تقاطع اللغات تتزايد؛ لذا فإن اتباع طرق يعد شيئاً ضرورياً لتقييم مقارنة الاختبارات المستخدمة عبر لغات متنوعة.

إن البحث المعني بالتقاطع الثقافي (بين الثقافات) والتقييم الأدبي للتقاطعات اللغوية يحتوي على عدة أمثلة إبداعية لتصاميم بحثية وطرق إحصائية صالحة لتقييم مقارنة الاختبارات المترجمة أو المكيفة عبر واحدة أو أكثر من زمرة لغوية. مثلاً: آلالوف، هامبلتون و سيريسي، 1999، بادجيل، راجو وكارتيتي، 1995، إليس وكيميل، 1992، هولين، دراسغو، و كوموكار، 1982، سيريسي، فيتز جيرالد، وإكسينغ، (1998).

لقد استعمل كل من فان دي فيفر وبورتينغا (1997)، فان دي فيفر وتانزر (1997) الدراسات السابقة لتطوير تصنيف أشكال التحيز والتعادل المرتبطة بالتقييم اللغوي التقاطعي (بين اللغات). ويشمل التحيز في هذا التصنيف كلاً من: تحيز المفهوم، تحيز المنهج وتحيز البنود. أما التعادل فيتألف من: التعادل البنيوي، تعادل وحدة القياس وتعادل مدى المقياس كاملاً. وتحدد هذه الفئات أنواع الاستنتاجات المقارنة التي يمكن تطبيقها عبر تقييمات للغات مختلفة. وبشكل أساسي، ومن أجل استنتاجات مقارنة عبر الأفراد الذين اختبروا وفق صيغ لغوية مختلفة لاختبار ما فإنه يجب إثبات صدق هذا الاختبار في كل اللغات.

وعلى الرغم من الجهد الكبير الذي تم بذله لتقدير مصادر الانحياز في تقييم التقاطع اللغوي (انظر مثلاً إلى الدراسات الحديثة التي قام بها آلالوف، 2003، آلالوف و آل، 1999، إيريكان 2002) فقد تم بذل أقل من ذلك بكثير لتقدير التعادل القياسي. واستعملت التصاميم البحثية والتحليلات الإحصائية لتطوير مجموع نقاط متسق عبر الزمر اللغوية، إلا أنه لم يتم تقييم مدى مقدرتهم وقصورهم بشكل كامل. إن الهدف من هذا الفصل هو نقد بعض علوم المناهج في هذا الميدان مع التركيز على تلك التصاميم التي توظف الخاضعين للاختبارات المتمكنين من اللغتين (ثنائيي اللغة) والقادرين عن الإجابة على أسئلة الاختبار سواء "الأصلية" أو "المترجمة/ المكيفة". وقد تم نقاش نقاط القوة والقصور لخيارات هذه التصاميم البحثية المختلفة اعتماداً على الدراسات في هذا الميدان.

مشكلة إنجاز تعادل معياري شامل في تقييم التقاطع اللغوي

إن المشكلة الكبيرة في تقييم التقاطع اللغوي تُوضح بشكل أفضل عبر مثال. لنفرض أننا أردنا مقارنة الإنجاز الرياضي لمجموعة طلبة يتكلمون اللغة الفرنسية مع مجموعة طلبة يتكلمون اللغة الإنكليزية بالمرحلة الدراسية نفسها في كندا. لنفرض أيضاً أنه تم وضع إطار المنهاج التعليمي نفسه لكل الطلاب في هذه الدراسة الافتراضية. يتم إنشاء نسختين من الاختبار الإنجازي. النسخة الأولى هي الاختبار الأصلي الذي صيغ باللغة الفرنسية أما الثانية فهي النسخة المعدلة باللغة الإنكليزية. للوصول للهدف من هذا المثال سنفترض التعادل في التركيب (أنظر إلى غيرل، 2000، كمثال لكيفية تدقيق التعادل في التركيب). بعد تقديم الاختبارات نلاحظ الفرق في الأداء بين الزمرتين الإنكليزية و الفرنسية. هل تمثل هذه الفروقات التي تم رصدها اختلافاً حقيقياً للمجموعة في الإنجاز الرياضي، أم فرق في الصعوبة بين نماذج الاختبار أم أنها تمثل كليهما؟

لفهم هذه المشكلة بشكل أفضل نتظاهر أنه لدينا مقياس إنجازي "حقيقي" تم على أساسه معايرة كلتا المجموعتين، وأن الزمرة الفرنسية "تفوق الزمرة الإنكليزية أداءً بواسطة انحراف قياسي متوسطياً 0.25 (SD) درجة. لتبسيط الأشياء بقدر الإمكان فإن هذا المقياس الحقيقي للعلامات يشمل الصفر ويشمل انحرافاً قياسياً لرقم واحد. ويتألف هذا الاختبار الرياضي المتخيل من خمسة أسئلة فقط. ونعالج هذه المشكلة باستخدام كل من نظرية الاختبار الكلاسيكية ونظرية الإجابة للسؤال (IRT) انظر مثلاً: هامبلتون، سواميناثان و روجرز، 1991). من المميزات المتعلقة باستخدام نظرية الإجابة للسؤال هي أن درجات المجموعة وإحصائيات صعوبة السؤال يمكن التعبير عنها بالارتكاز على مقياس الدرجات نفسه.

السيناريو1: أولاً، نعتبر أن الحالة هي حيث تكون الأسئلة متعادلة عبر اللغتين الفرنسية والإنكليزية وهذا يتحقق حين تتم ترجمة الأسئلة من الفرنسية إلى الإنكليزية وعملية الترجمة/التكييف لم تغير "الجوهر" الأساسي للأسئلة (مثلاً: الأسئلة متعادلة إحصائياً ولغوياً في كلتا اللغتين). بهذه الفرضية، إذا قمنا بحساب إحصائيات بند نظرية الإجابة للسؤال (IRT) والنظرية الكلاسيكية قد تبدو النتائج مماثلة لنتائج جدول رقم 5-1 . في الجدول رقم 5-1 تظهر بوضوح نتيجتان، الأولى بالنظر إلى القيم (P) نسبة الطلاب الذين أجابوا عن الأسئلة بشكل صحيح، ومتوسط علامات الاختبار، نرى أن المجموعة الفرنسية فاقت المجموعة الإنكليزية أداءً. وهذا استنتاج صحيح. لنتذكر أننا افترضنا أن المجموعة الفرنسية فاقت المجموعة الإنكليزية 0.25 (SD) وحدة. ثانياً إن المتغيرات b (IRT) تقديرات صعوبة الأسئلة) هي نفسها في كلا المجموعتين.

جدول 5-2

إحصائيات توضيحية تعتمد على الطريقة الكلاسيكية ونظرية الإجابة عن السؤال (IRT) حين يبقى تكافؤ السؤال عبر اللغات

(المجموعة الفرنسية تتفوق "فعليا" بـ 25 وحدة على المجموعة الإنكليزية)

قيم p الكلاسيكية وحدات القياس b المستندة على IRT

	رقم p الكلاسيكية		متغيرات d	
Item	فرنسي - إنجليزي		فرنسي - إنجليزي	
1	.50	.52	1.25	1.25
2	.60	.62	0.00	0.00
3	.55	.57	.63	.63
4	.65	.67	-0.63	-0.63
5	.70	.72	-1.25	-1.25
Mean Score[a]:	3.0	3.1	0.0	0.25

ملاحظة:

استنتاج 1: تفوقت المجموعة الفرنسية على المجموعة الإنكليزية بالأداء.

استنتاج 2: معايير وحدة قياس سؤال حسب نظرية الإجابة عن السؤال ثابتة عبر المجموعتين.

a: متوسط العلامات مرتكز على معيار العلامة الأولي للإحصائيات الكلاسيكية وعلى معيار الحرف اليوناني الثامن الموحد (1.0) لإحصائيات IRT.

توضح هذه النتيجة ميزة ثبات نموذج وحدة قياس السؤال المعروف لمعايرة IRT هامبلتون وآل، (1991)، أي أن معايير وحدة القياس المستندة على نظرية IRT لا تعتمد على النموذج المستعمل لتحديدها. هذه الخاصية تفسر لماذا تستعمل الطرق المعيارية غالباً لتحديد الاختبارات المقدمة إلى مجموعات لغوية مختلفة (مثلاً

آنغوف وكوك، 1988، إيليس، 1989، هولين وماير، 1986، وودكوك ومونوز-ساندوفال، 1993).

يوضح السيناريو 1: أنه حين تكون الأسئلة متكافئة عبر اللغات فإن المشكلات المتعلقة بترجمة المجموعات المختلفة تنعدم. لا يوجد اختلافات تتعلق بالاختبار. وفي حال ملاحظة مثل هذه الاختلافات فإنه يمكن إسنادها إلى الفروقات بين المجموعتين اللغويتين. لسوء الحظ فإننا لا نعلم من خلال الممارسة إذا كانت الأسئلة متعادلة عبر اللغات أم لا. لنرى كيف يمكن أن تبدو هذه النتائج في حال كانت الأسئلة أكثر صعوبة في واحدة من الصيغتين اللغويتين للاختبار.

السيناريو 2: هنا نحن نختلق حالة حيث تكون الأسئلة وسطياً أسهل في الإنكليزية بحوالي 0.25 (SD) وحدة. مع بقاء نتيجة المجموعة (المجموعة الفرنسية تتفوق "فعلياً" بـ 0.25 (SD) وحدة على المجموعة الإنكليزية). يمثل هذا السيناريو الحالة حيث يجعل تكييف الاختبار إلى اللغة الإنكليزية الصيغة الإنكليزية للاختبار أسهل من الصيغة الفرنسية. قد تحدث هذه الحالة إذا، على سبيل المثال، أدخل المترجمون دون انتباه مفاتيح للأجوبة الصحيحة أثناء تعديل الأسئلة أو استعملوا لغة أبسط عبر الاختبار. يمثل الجدول 5-2 النتائج المتصورة لهذا السيناريو.

إن أكثر ملاحظة ملفتة للنظر في الجدول 5-2 هي أن أول استنتاج غير صحيح. عُرفت المجموعة الفرنسية كمتفوقة على المجموعة الإنكليزية. إلا أن معايير النسبة للسؤال"P" ومعدلات المجموعة تشير إلى أن المجموعة الإنكليزية هي المتفوقة. هذا الاكتشاف هو نتيجة لحقيقة أن الأسئلة أكثر صعوبة بالفرنسية منها بالإنكليزية. في هذا السيناريو درجة الاختلاف بين متوسط الصعوبات بالأسئلة عبر اللغتين هي أكبر من درجة الاختلاف بين الإنجاز الحقيقي للمجموعتين (0.50 مقابل 0.25 ، بالتتابع). نشأ استنتاجنا الخاطئ لأن النموذجين الكلاسيكي ونظرية (IRT) لا يفسران حقيقة أن الصيغة الإنكليزية للأسئلة أسهل.

الاستنتاج الثاني هو أيضاً غير صحيح. إن الظروف المتصورة لهذا السيناريو تحدد الأسئلة الإنكليزية على أنها أكثر سهولة، إلا أن معايرة وحدة قياس صعوبة (IRT) لوحدة قياس وحدة قياس (b) هي نفسها لكلا المجموعتين اللغويتين. فكيف يمكن ذلك؟

جدول 5-2

إحصائيات توضيحية تعتمد على الطريقة الكلاسيكية ونظرية الإجابة للسؤال(IRT)

حين تكون الأسئلة الفرنسية أكثر صعوبة

(المجموعة الفرنسية تتفوق "فعليا" ب .25 وحدة على المجموعة الإنكليزية)

	Classical p *Values*		*IRT* b *Parameters*	
Item	*English-French*		*English-French*	
1	.50	.48	1.25	1.25
2	.60	.58	0.00	0.00
3	.55	.53	.63	.63
4	.65	.63	–0.63	–0.63
5	.70	.68	–1.25	–1.25
Mean Score[a]:	3.0	2.9	0.0	–0.25

ملاحظة:

استنتاج 1: تفوقت المجموعة الإنكليزية على المجموعة الفرنسية بالأداء.

استنتاج 2: الأسئلة متكافئة عبر اللغتين (مثلا": لا يوجد اختلافات).

a: متوسط العلامات مرتكز على معيار العلامة الأولي للإحصائيات الكلاسيكية وعلى معيار الرف اليوناني الثامن الموحد (1.0) لإحصائيات IRT.

يتضح أن معايير وحدة قياس صعوبة ونظرية الإجابة للسؤال (IRT) للصيغتين الإنكليزية والفرنسية متعادلة حيث يظهر أنها ترتكز على مقياس عام (مشترك). بينما هي ليست كذلك. فلم يقم أي من الطلبة الإنكليز بالإجابة عن الأسئلة الفرنسية في حين لم يجب أي من الطلبة الفرنسيين بالإجابة عن الأسئلة الإنكليزية. لهذا فإن معايير وحدة القياس للأسئلة الفرنسية تُحسب بالاعتماد على معطيات الطلبة الفرنسيين فقط، بينما تُحسب معايير وحدة القياس للأسئلة الإنكليزية بالاعتماد على معطيات الطلبة الإنكليز فقط. على سبيل المثال فإن الصيغة الإنكليزية للسؤال (1) تمثل سؤالاً أعلى بحوالي 1.25 (SD) درجة من متوسط صعوبات السؤال لكل الأسئلة الإنكليزية. لا توجد طريقة لمعرفة مدى انحراف هذا السؤال عن متوسط صعوبات السؤال الفرنسي. بطريقة مماثلة، تمثل الصيغة الفرنسية للسؤال سؤالاً أعلى بحوالي 1.25 (SD) درجة من متوسط صعوبات السؤال لكل الأسئلة الفرنسية. على الرغم من أنه لكلا مجموعتي الأسئلة قيمة انحراف 1.25 فهي تمثل انحرافات من متوسط معايير الاختلاف (مثلاً المعيار الإنكليزي والمعيار الفرنسي). هذه القيم الانحرافية المحددة لغوياً غير قابلة للمقارنة عبر اللغات. خذ الآن بعين الاعتبار أن 48 ٪ فقط من طلاب اللغة الفرنسية أجابوا عن السؤال (1) بشكل صحيح، في حين أجاب 50 ٪ من طلاب اللغة الإنكليزية على هذا السؤال بشكل صحيح. ولأن وحدات القياس b كانت نفسها بالصيغتين الإنكليزية والفرنسية فإن النتيجة هي أن متوسط الدرجة للمجموعة الفرنسية على مقياس العلامات منخفض بالمقارنة مع المجموعة الإنكليزية. وقد حدث تعديل مماثل للأسئلة الأخرى. تُزودنا نتائج هذا التحليل باستنتاجات مغايرة لما نعتقده صحيحاً. تبدو هذه الأسئلة متعادلة عبر اللغات (بينما هي ليست كذلك). ويبدو أن الطلاب الإنكليز يؤدون بشكل أفضل من الطلاب الفرنسيين (بينما العكس هو الصحيح).

في السيناريو (2): نحدد كلا المجموعتين واختلافات الأسئلة عبر اللغات. إن السبب في أن تحليلاتنا أثمرت عن استنتاجات خاطئة هو أن نموذج المعيار لم يفسر

هذين العاملين. في الواقع حين تكون الفروقات بين الأسئلة والمجموعات غير معروفة فإنه يجب القيام ببعض الافتراضات. علينا أيضاً أن نفترض أن الأسئلة متعادلة عبر اللغات. ثم نقوم بالبحث عن فروقات بين المجموعات، أو نفترض أن المجموعات متعادلة ثم نبحث عن فروقات الأسئلة. يعكس السيناريو (2) الافتراضات التي حصلت حين تمت معايرة الأسئلة مستعملين الخيارات الافتراضية لبرنامج "IRT" النموذجي مثل بيلوك ميسليفي و بوك، (1990) هذا النوع من التحليل سيقوم بمعايرة وحدات القياس b بشكل مترابط (إلى مقياس شائع) بدون تفسير (صحيح) للفروقات بين المجموعات لتصل إلى نتائج كتلك المتمثلة في الجدول (5-2) وهي معقولة تماماً. إن استخدام طريقة قياس تحويلية كما في Stocking و(1983) Lord، التي تعدل وحدات القياس من معايرة ما لتكون على نفس المقياس كأسئلة من معايرة مختلفة لا يمكن أخذها بالاعتبار لأنه لا توجد أسئلة متاحة غير شائعة (مثلاً: غير مترجمة) لإجراء تعديل كهذا.

هل يعد السيناريو (2) واقعياً؟ على الأغلب أنه ليس كذلك. بإعطاء طرق تكييف وترجمة اختبار دقيقة (للتطورات في طرق تكييف الاختبار، انظر إلى هاملتون، المقطع (1)، هذا المجلد: هاملتون وباتسولا، 1999، ومولیس، كيللي، وهالي، 1996، فإنه من غير المحتمل أن كل الأسئلة في صيغة لغوية واحدة لاختبار ما ستكون أكثر صعوبة من نظيراتها في صيغ لغوية أخرى. إن سيناريو أكثر احتمالاً سيحوي بعض الأسئلة الأكثر صعوبة في الفرنسية وأخرى أكثر صعوبة بالإنكليزية. على أية حال فإن النقطة المهمة هي أن عدم تعادل السؤال و عدم تعادل المجموعة يمكن، وعلى الأغلب أنها كذلك، أن تظهر في الوقت نفسه في تقييم التقاطع اللغوي. حين يتوفر هذان العاملان، فإن مناهج القياس التقليدية غير كافية لاستخراج استنتاجات حول الفروقات بين الاختبارات والمجموعات عبر اللغات. إن الشيء الضروري للقيام بمثل هذه الاستنتاجات هو إما طريقة لتفسير اختلافات المجموعة ضمن المعايرة أو نماذج الدرجات، أو مجموعة أسئلة يمكن اعتبارها متعادلة عبر

اللغتين. كما سنشرح لاحقاً، فإن الطريقة لتحديد أسئلة يمكن اعتبارها متعادلة عبر اللغات هي تقديم الاختبارات إلى ممتحنين ثنائيي اللغة.

استخدام ثنائيي اللغة لتقييم صيغ لغوية مختلفة لاختبار ما:

واحدة من الطرق لمعالجة تعادل صيغتين لغويتين مختلفتين لاختبار ما هي تقديم الصيغتين اللغويتين المنفصلتين إلى مجموعة من الخاضعين للاختبار المتمكنين من كلا اللغتين (ثنائيي اللغة). المنطق الأساسي لهذا الاختبار هو أنه باستخدام مجموعة واحدة من الخاضعين للاختبار "مجموعة لغوية"، فإنه يتم إهمال النتائج، ويمكن تحقيق تعادل معياري كامل. وهكذا فإن رصد الفروقات لأداء السؤال أو الاختبار عبر اللغات يمكن أن يعزى إلى الاختلافات اللغوية بين الاختبارات أو الأسئلة. على الرغم من أنه باستخدام مجموعة واحدة عادة ما تهمل اختلافات المجموعة في أغلب التصميمات البحثية، فإن هناك بعض الخلل في هذا المنطق حين يطبق على مسألة تقييم التقاطع اللغوي. إن المشكلة الأكثر ظهوراً هي الافتراض المطلق أن ثنائيي اللغة متمكنين من كلتا اللغتين بالدرجة نفسها. على سبيل المثال في حال أدت مجموعة من ثنائيي اللغة بشكل مختلف في الصيغ اللغوية " A وB" لسؤال ما، فإن نسب هذا الاختلاف إلى التكييف الخاطئ يفرض أن ثنائيي اللغة سيؤدون بالطريقة نفسها في كلتا الصيغتين اللغويتين للسؤال إذا كان التكييف وافياً. على الرغم من ذلك، فمن المحتمل أن الممتحنين الثنائيي اللغة هم أكثر تمكناً في لغة من الأخرى. لذا، توجد فرضية منافسة معقولة وهي أن ثنائيي اللغة يؤدون بشكل أفضل في الأسئلة المقدمة بلغتهم الأقوى، حتى حين تكون نسختا السؤال متعادلتين بحق.

خلل آخر يتعلق بهذا المنطق هو أنه يصف ثنائيي اللغة كما لو أنهم أحاديو اللغة، مجموعة متجانسة من الخاضعين لاختبار، بينما في الحقيقة، إن مجموعة ثنائيي اللغة الخاضعين لاختبار تتألف على الأغلب من أفراد ذوي خلفيات ومقدرات ومهارات لغوية مختلفة بيكر، 1988، فالدي وفيجورو، (1994). في الولايات المتحدة.

على سبيل المثال، فإن مجموعة من ثنائيي اللغة الإنكليزية - الإسبانية يمكن أن تشمل أناساً لغتهم الأولى الإنكليزية وتعلموا الإنكليزية في المدرسة الثانوية و مهاجرين (من عدة بلدان متنوعة) ناطقين باللغة الإسبانية قد تعلموا حديثاً التكلم باللغة الإنكليزية و مهاجرين من الجيل الثاني تعلموا الإنكليزية كلغة ثانية في المدرسة الابتدائية. لهذا، فالفرضية بأن ثنائيي اللغة يمثلون "نمطاً" واحداً من الخاضعين لاختبار غير منطقية. كما سأناقش لاحقاً، إن الاختلافات اللغوية ضمن مجموعة من ثنائيي اللغة يجب أن تدرج في التصميم البحثي عند استخدام ثنائيي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة. إن المشكلة الأكثر دقة ولكن الجدية، هي أن استخدام ثنائيي اللغة من أجل تقييم الاختبارات هي المقارنة المثيرة للجدل لثنائيي اللغة و أحاديي اللغة (هاملتون وكانجي، 1995) تم تسمية هذه المشكلة سابقاً المشكلة التمثيلية (سيرسي، 1997). في الاختبارات التعليمية، مثلاً، يميل ثنائيو اللغة للاختلاف عن مجموعاتهم الأحاديي اللغة. قد يكون ثنائيو اللغة الذين يجيدون لغتين بكفاءة عالية ممثلين فقط للطلاب الأعلى إنجازاً في المجموعة الأحادية اللغة أيضاً. بشكل معاكس، فإن ثنائيي اللغة الذين يجيدون لغة أو لغتين بشكل هامشي يمثلون فقط الطلاب الأقل إنجازاً في واحدة من المجموعات الأحادية اللغة. على أية حال، فإن تصنيف الكفاءة في نموذج ثنائيي اللغة تميل إلى أن تكون مختلفة جداً عن التصنيفات المقابلة لجماعاتهم أحاديي اللغة.

على الرغم من أن استخدام ثنائيي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة تتطلب البراعة حين تخصص للتصاميم البحثية وحين تستعمل البيانات التحليلية لمعطيات الحالة الفنية، فقد يزودنا ثنائيي اللغة بمعلومات قيّمة فيما يتعلق بتعادل الاختبار ومقارنة الدرجات. تم طرح خيارات التصاميم البحثية في إشعار هذا المقطع لاستخدام ثنائي اللغة من أجل تقييم الاختبارات المقدمة في لغات مختلفة، وتم ملاحظة متغيرات مربكة بحاجة إلى السيطرة عليها. بالإضافة إلى أنه تم التزويد باقتراحات لاستخدام ثنائيي اللغة للقيام بتقييمات أكثر شمولاً للاختبارات المقدمة بعدة لغات.

بدائل عن التصاميم البحثية باستخدام ثنائيي اللغة

تصاميم أحادية المجموعة:

يتطلب التصميم ثنائي اللغة للمجموعة الأحادية تقديم صيغتين لغويتين مختلفتين لاختبار ما لمجموعة واحدة من الخاضعين لاختبار ثنائي اللغة. في هذا التصميم لا يوجد تطابقات ضمن المجموعة ثنائية اللغة المُعَدة لتحديد "أنماط" مختلفة لثنائيي اللغة. إلا أن إدارة الاختبارات اللغوية المختلفة عادة ما تكون متعادلة. حيث إن نصف الممتحنين تقريباً يخضعون للاختبار في اللغة A أولاً بينما يخضع النصف الآخر للاختبار في اللغةB. يشابه هذا الاختبار تصميم المجموعة الأحادية المشروح في اختبار الدراسات السابقة المتعلقة بالتعادل. (مثال: كولن وبرونان، 1995).

من الممكن أن يستخدم التصميم ثنائي اللغة للمجموعة الأحادية لتعديل درجات اختبار بأكملها على صيغ لغوية مختلفة. على سبيل المثال، قام بولدت (1969) بحساب مقارنة العلامات للصيغتين اللغويتين الإنكليزية والإسبانية المتعلق باختبار الأهلية للمدارس الثانوية (SAT) عبر اختبار مجموعة صغيرة (عدد= 14) لثنائيي اللغة الإسبانية - الإنكليزية من طلاب المدرسة الثانوية بواسطة صيغتي الاختبار، استنتج أن طرح 200 درجة من درجة اختبار طلاب اللغة الإسبانية يزودنا بتقدير علامة الطلاب المتوقعة لاختبار اللغة الإنكليزية.

يمكن استخدام هذا التصميم أيضاً لتقييم أداء الأسئلة الفردية عبر اللغتين. على سبيل المثال، إن أداء ثنائيي اللغة في كل سؤال يمكن أن يستخدم لتقييم وظيفية الأسئلة التباينية (DIF) عبر اللغات. في حال أدت الأسئلة بشكل متباين في اللغتين (فيما يتعلق بإحصائيات السؤال كصعوبة السؤال أو التمييز)، فإن الأسئلة تصنف بوصفها تعمل بشكل مختلف عبر اللغات، وهي لا تستعمل لإرساء الاختبارات على مقياس عام. بدلاً من ذلك، فإن الأسئلة التي تظهر إحصائيات متشابهة عبر اللغتين ليست أسئلة "DIF" يمكن أن تستخدم في تصميم موازنة إرساء الاختبار لتثبيت أو ربط الاختبارات إلى مقياس عام. إن كلاً من طريقتي النظرية الكلاسيكية ونظرية (IRT) يمكن أن توازنا المعايير بهذا النمط.

هناك ثلاثة نقاط ضعف أساسية على الأقل في التصميم ثنائي اللغة للمجموعة الأحادية. من الواضح أن التصميم لا يفسر وجود أنماط مختلفة من ثنائيي اللغة من الخاضعين للاختبار. إذا تم إجراء الدراسة باستخدام ثنائيي اللغة الذين تهيمن عليهم اللغة A، فإن النتائج قد لا تعمم للحالة التي يستخدم فيها ثنائيي اللغة الذين تهيمن عليهم اللغة B. نقطة الضعف الثانية هي وجود أثر الممارسة. لأن الممتحنين يجيبون عن كل سؤال في كلا اللغتين، فإن الإحاطة بالسؤال في نموذج الاختبار الأول قد يؤثر على إجابات الممتحنين إلى السؤال المماثل في نموذج الاختبار الثاني. بالرغم من أن عامل الموازنة قد يضبط هذا على المعدل، إلا أن النتائج قد تختلف عما يمكن ملاحظته في حال أجاب الممتحنون على نموذج اختبار واحد. أما نقطة الضعف الثالثة فهي أن الطريقة غير اقتصادية نسبياً. إن اختبار مجموعة واحدة من الخاضعين لاختبار ذي نموذجين يضاعف وقت تقديم الاختبار اللازم لإكمال الدراسة. نقطة ضعف مرتبطة بذلك هي أن الممتحنين قد يفقدون الحافز أو يصبحون أكثر إرهاقاً من أن يخضعوا لنموذج ثان مشابه للأول.

مثال جدير بالذكر للتصميم الثنائي اللغة للمجموعة الأحادية هو الدراسة التي أجريت لربط التقييم الإسباني للتعليم الأساسي (SABE) بنظيراتها في اللغة الإنكليزية واختبار كاليفورنيا الإنجازي (CAT) ، والاختبار الشامل للمهارات الأساسية (CTBS) (1988 ، ماك - غراو هيل /CTB) هناك ميزات عدة لهذه الدراسة مثيرة للإعجاب. أولاً، من أجل الضمان أن الطلاب ثنائيي اللغة ماهرون في كلا اللغتين، تم استخدام تقييمات المدرس واختبارات المهارة اللغوية لتصفية الطلاب الذين لم يكونوا متمكنين في اللغتين الإنكليزية والإسبانية. ثانياً، بدلاً من جعل الطلاب ثنائيي اللغة يخضعون لاختبارين منفصلين، قُدم للطلاب مجموعات أقصر من الأسئلة الإنكليزية و الإسبانية المعتمدة. وقد استُخدم أداء ثنائيي اللغة في هذه الأسئلة المعتمدة لاشتقاق جداول تحويلية لمقارنة أداء الطلاب في الـ SABE مع أداء الطلاب في CAT و CTBS.

على الرغم من أنه تم توظيف التصميمات البحثية المحددة في دراسة الـ SABE التي عالجت بعض النقائص لتصميم المجموعة الأحادية، إلا أنه باستخدام أكثر من مجموعة واحدة من ثنائيي اللغة يمكن تحسين الصدق الداخلي والخارجي لهذا النوع من الدراسة.

تصميمات متعددة - المجموعات

تصميم ثنائي - المجموعة. إن التصميم الأكثر بساطة للمجموعة المتعددة ثنائية اللغة يستخدم بشكل عشوائي مجموعتين متعادلتين ثنائيتي اللغة. يمكن إنشاء هذه المجموعات من خلال تمرير صيغتي اختبار أو تعيين ممتحنين لهذه الصيغ بشكل عشوائي. في هذا التصميم، تخضع كل مجموعة لواحد أو اثنين من صيغ الاختبار، مع إلغاء أي احتمال لأثر الممارسة. بالإضافة إلى هذا، فإن كون المجموعتين متعادلتين بشكل عشوائي، يحتم انعدام أي تأثير على المجموعة. هذا التصميم أيضاً اقتصادي أكثر من تصميم المجموعة الأحادية. يمكن جمع بيانات صيغتي الاختبار خلال مقدار الوقت الذي يستغرقه تقديم صيغة اختبار واحدة.

إنشاء صيغتي اختبار. إن نمط الاختبار لكل مجموعة يمكن أن يكون أكثر تعقيداً عند استخدام ثنائيي اللغة من تنفيذ تصميم مجموعتين متعادلتين. إن أكثر الخيارات المباشرة هي جعل مجموعة واحدة تخضع لصيغة الاختبار A، بينما تخضع الأخرى لصيغة الاختبار B بالرغم من أن هذا الخيار يوازي حالة المجموعتين المتعادلتين (تخضع كل مجموعة لاختبار لم يتم التطرق إليه أو صيغة معتمدة)، فإنه ليس الأمثل عند اختبار ثنائيي اللغة. عند استخدام هذا التصميم فإنه لا يمكن تقييم أداء المجموعة الأولى في اللغة B، ولا أداء المجموعة الثانية في اللغة A يكون الخيار الأفضل هو جعل كل مجموعة تخضع لصيغة مختلطة تحوي على أسئلة من كلا اللغتين، اللغة A واللغة B.

قام سيرسي وبيربيرولو (2000) بإعطاء مثال عن هذا النوع من التصميم المقدم للغة المختلطة. وقد قيموا دقة الترجمة لكلا المجموعتين من الأسئلة من الصيغتين من نموذج تقييم المدرس. إن النسخة الأصلية لهذا الاختبار كانت باللغة التركية بينما كانت الصيغة المكيفة بالإنكليزية. للسيطرة على أثر الممارسة، أجاب الممتحنون على صيغة لغوية واحدة فقط من كل سؤال. ولكن ظهرت الأسئلة الإنكليزية و التركية على كل من نموذجي الاختبار. وقد تم ذلك بالتبديل بين اللغتين في كل نموذج. في النموذج الأول، كانت كل الأسئلة الفردية الرقم بالإنكليزية وكل الأسئلة الزوجية الرقم بالتركية. ظهر المخطط المعكوس على الصيغة الثانية. مثلاً، إن السؤال رقم واحد على النموذج الأول ظهر بالإنكليزية بينما ظهر نظيره التركي كالسؤال الأول على النموذج الثاني. السؤال الثاني على النموذج الأول كان بالتركية ونظيره الإنكليزي ظهر كالسؤال الثاني على النموذج الثاني، وهكذا دواليك. علاوة على ذلك، تم إدراج سؤالين إنكليزيين على كل نموذج. قدمت هذه الأسئلة معياراً تم استعماله ضمن تحليل IRT للبرهان على أن فرضية المجموعات المتعادلة بشكل عشوائي صحيحة. تم توضيح دراسة هذا التصميم في الشكل 1.5 . تم استعمال الأسئلة المقدمة بالإنكليزية في كلا النموذجين (معيار 1 و 2) لتقدير إذا ما كان الممتحنون ثنائيو اللغة الذين يخضعون لكل نموذج اختباري متعادلين بشكل عشوائي. تم استعمال تحليلات IRT المرتكزة على DIF لاختبار إذا ما كانت الصيغتان الإنكليزية والتركية لكل سؤال يمكن أن تثبتا باستعمال وحدات القياس نفسها.

الصيغة الثنائية اللغة ١

المعتمد ١ (إنكليزي)	المعتمد ٢ (إنكليزي)	السؤال ١ (إنكليزي)	السؤال ٢ (تركي)	السؤال ٣ (إنكليزي)	السؤال ٤ (تركي)

الصيغة الثنائية اللغة ٢

المعتمد ١ (إنكليزي)	المعتمد ٢ (إنكليزي)	السؤال ١ (تركي)	السؤال ٢ (إنكليزي)	السؤال ٣ (تركي)	السؤال ٤ (إنكليزي)

الشكل رقم 1.5 مثال عن تصميم إدارة اللغة المختلطة للممتحنين ثنائيي اللغة.

يعد أثر التبديل بين اللغات على أداء المتحنين غير معروف. توجد استراتيجية بديلة وهي الحصول على قسمين منفصلين من الاختبار لكل لغة. نصح سيرسي وبيربيرولو (2000) بإجراء المقابلات مع المتحنين ثنائيي اللغة لاكتشاف إذا كان تغيير لغة الأسئلة ضمن الاختبار شيئاً مربكاً أو معيقاً لأدائهم بطريقة ما.

وقد استنتج سيرسي وبيربيرولو (2000) أن ثنائيي اللغة مفيدون لتحديد الأسئلة التي لم تكن متعادلة عبر اللغات. من ناحية ثانية، لقد صرحوا أن هذا الإجراء لا يمكن أن "يثبت" أن الأسئلة غير المصنفة DIF كانت متعادلة عبر اللغات. لكنهم أفادوا أن الأسئلة التي لم تظهر DIF هي المرشحة بشكل أقوى لتثبيت المعايير اللغوية المنفصلة من الأسئلة المصنفة أو التي لم يتم تقييمها.

كان هناك قصور في دراسة سيرسي وبيربيرولو (2000) وهو أنه تم استعمال نمط واحد فقط من ثنائيي اللغة الخاضعين للاختبار. تضمنت عينة ثنائيي اللغة طلاباً في الجامعة التركية حيث كانت الإنكليزية هي اللغة الأساسية للتعليم. على الرغم من أنهم لم يقوموا بتصفية الطلاب الذين ذكروا بتقاريرهم أنهم "قليلو البراعة" في قراءة أو فهم الإنكليزية، فإن تصميمهم لم يشمل على أي من ثنائيي اللغة الذين كانت الإنكليزية لغتهم الأولى. إن تقييماً أكثر شمولاً لأمانة الترجمة سيحوي كل من ثنائيي اللغة الإنكليزية - التركية و ثنائي اللغة التركية - الإنكليزية. وهكذا، في حال أمكن ذلك، فإنه يمكن تطوير تصميم مجموعتي ثنائيي اللغة بإدراج أكثر من نمط واحد من المتحنين ثنائيي اللغة.

تصميم رباعي - المجموعات: هناك إضافة واضحة على تصميم المجموعتين ثنائية اللغة وهي جعل مجموعتين ثنائية اللغة تختلف من حيث اللغة الأصلية، للخضوع إلى كلاً من صيغتي الاختبار. تشمل المجموعة الأولى ثنائيي اللغة المتمكنين من اللغة الأولى، بينما تشمل المجموعة الثانية ثنائيي اللغة المتمكنين من اللغة الثانية. يتم تعيين الأفراد في كل مجموعة لواحدة من صيغتي الاختبار (قد تكون

لغة مختلطة). بالإضافة إلى تأمين مجموعات ممثلة أكثر لثنائيي اللغة، فإن هذا التصميم يسمح بتحليل الفروقات الأداء بين النمطين من ثنائيي اللغة. إن تحليلات DIF يمكن إجراؤها بشكل منفصل لكل مجموعة. مثلاً، إذا بدا أن سؤالاً ما متعادل إحصائياً لكل من المجموعتين التركية - الإنكليزية و الإنكليزية - التركية ثنائيتي اللغة، فإنه يتم جمع حقائق أخرى تفيد بأن الأسئلة هي "نفسها" في كلتا اللغتين. إذا أظهر أن سؤال ما DIF في واحدة من المجموعات ثنائية اللغة دون أن يظهر في المجموعة الثانية، تجمع معلومات متعلقة بتفسيرات الاختلافات اللغوية للسؤال.

إذا تم استخدام صيغتي اختبار أحادية اللغة في تصميم رباعي - المجموعات، فإن التحليل التقليدي للإجراءات المتباينة يمكن أن يفيد في تقييم تأثيرات الاختبار "الترجمة" وتأثيرات المجموعة. تم تصوير هذه الحالة في الشكل 2.5، الذي يحوي تحليلاً لتقييم صيغتين افتراضيتين إنكليزية وإسبانية.

نموذج اختباري

	إنكليزي	إسباني
اللغة المهيمنة		
إنكليزي	؟	؟
إسباني	؟	؟

تفسيرات ونتائج محتملة:

-لا توجد تأثيرات: تدعم تكافؤ صيغ الاختبار عبر اللغات

-التأثير الرئيس لصيغة الاختبار: مشكلة ترجمة

-التأثير الرئيس للغة المهيمنة: اختلاف المجموعة

-التأثير التفاعلي: المجموعة و/ أو صيغة الاختبار غير متكافئة، لا يوجددعم لتعادل الترجمة

الشكل 2.5 تصميم افتراضي رباعي - المجموعة

إذا وُجد تأثير تفاعلي، قد يشير إلى أن اللغة متى تم اختبارها فإنها تشكل فرقاً بحسب المجموعة التي جرى عليها الاختبار. إذا وُجد التأثير الرئيس لصيغة الاختبار اللغوية، فسيشير إلى: مشكلة تتعلق بالترجمة. إن التأثير الرئيس للغة المهيمنة للمجموعة سيشير إلى أن النمطين من ثنائيي اللغة غير متساويين. يمكن إجراء عدة تحليلات باستخدام تغيرات تابعة مختلفة (مثلاً: علامات السؤال، أو علامات الاختبار الإجمالية، أو الدرجات الثانوية على مجالات ضمنية محددة). وهكذا، فإن الإضافات على التصميم الثنائي - المجموعة سيقدم معلومات متزايدة تتعلق بالتفاعل بين اتجاه اللغة الأصلية لثنائيي اللغة واللغة الأصلية للسؤال.

تصميمات متعددة - المجموعات: يمكن للتصميمات رباعية - المجموعات، أن تُوسع بشكل طبيعي لمجموعات أكبر. على سبيل المثال، قد يشمل تصميماً ما لمجموعة من الممتحنين الذين يعتبرون" متمكنين بشكل متساو" من كلا اللغتين. ويمكن استخدام التصاميم أيضاً للتعامل مع ثنائيي اللغة الذين ينتمون إلى خلفيات متعددة بشكل منفصل. مثلاً، قد ترغب دراسة تقارن بين الصيغتين الإنكليزية والإسبانية بالاطلاع على الفروقات بين الكاريبيين، والأمريكيين المتوسطين (أمريكا الوسطى) والمكسيكيين و ثنائيي اللغة الإسبانية - الإنكليزية في جنوب أمريكا. إن الخيار المحدد لعدد المجموعات في التحليل يجب أن يُحرك بواسطة اهتمامات تتعلق بتصاميم بحثية تقليدية كتحديد المتغيرات الخارجية وحجم العينة وقياس الفرضيات المنافسة المعقولة.

باعتبار أن عدد المجموعات المحتملة يتزايد، يظهر سؤال بديهي، وهو هل يمكن للمقاييس المتواصلة لهيمنة اللغة أن تندمج في التصميم البحثي بدلاً من استخدام مجموعات متعددة غير مترابطة. على سبيل المثال، مقاييس كفاءة اللغة A واللغة B يمكن أن ترتد عبر السؤال وبيانات أداء الاختبار لاكتشاف إذا ما كانوا مرتبطين بفهارس DIF فروقات علامات الاختبار الإجمالي. مثلاً، بينوك - رومان (1995) استخدم التحليل الارتدادي لتحديد آثار العوامل اللغوية على أداء

اختبار (GRE اختبار تقرير التخرج) لثنائيي اللغة الإسبانية - الإنكليزية بيرتو ريكان. كان التركيز الرئيس لتحليلها حول إذا ما كانت الصيغة اللغوية لآثار اختبار تؤثر على الاستنتاجات المكتسبة حول المتحنين ثنائيي اللغة. وقد وجدت باستخدام هذه التصميمات أن الكفاءة باللغة الإنكليزية فسرت تباين علامات الاختبار الشفوي GRE حتى 34 ٪. على الرغم من أن تحليلها لم يتحقق من اختبارات المقارنة المقدم بلغات مختلفة، فهي موضحة لأنواع المعلومات التي يمكن جمعها باستخدام التصاميم المتطورة للمجموعات الثنائية اللغة.

تصميمات تستخدم ثنائي اللغة وأحادي اللغة:

تم مناقشة قصور التصميمات التي تستخدم مجموعات منفصلة من المتحنين أحاديي اللغة سابقاً باعتباره قصوراً متعلقاً بالتصاميم الثنائية اللغة. إن التصميمات الأحادية اللغة محدودة لأن هذه النماذج غير قادرة على تحقيق تعادل قياسي كامل. أما التصميمات الثنائية اللغة فهي محدودة بسبب المشكلة التمثيلية. يقترح هذا القسم تصميماً أكثر شمولاً يستخدم كلا النمطين من المتحنين.

استخدام ثنائيي اللغة لتحديد أسئلة معتمدة (معيارية) للتحليلات أحادية اللغة.

تحتاج التصاميم أحادية اللغة إلى بعض الآلية لتفسير الاختلافات في المهارة بين المجموعتين اللغويتين. لو توفر معيار خارجي مرتبط بقوة بالمهارة التي يتم قياسها، فمن الممكن أن يستخدم لتعديل اختلافات المجموعة في الاختبارات. ولكن المعايير الخارجية الصالحة نادرة، هذا إذا توفرت على الإطلاق. خيار ثان لتفسير الاختلافات اللغوية للمجموعة، هو استخدام مجموعة أسئلة متعادلة بما يتعلق بقياس سرعة ودقة العمليات العقلية في اللغتين (أسئلة معتمَدة). بتقديم أسئلة متعادلة، يمكن استخدام الفروقات في أداء المجموعة في هذه الأسئلة، وذلك لتعديل الدرجات في واحد أو أكثر من صيغتي الاختبار (كما تم في تحليلات التعادل التقليدية). أو يمكن استخدام هذه الأسئلة لمعايرة الأسئلة الأخرى على مقياس شائع

كما سنشرح لاحقاً (انظر إلى ووددوكومونوز- ساندوفال، 1993، لتوضيح هذه العملية).

إن احتمال كون ثنائيي اللغة مفيدون بشكل خاص يكمن في تحديد الأسئلة المتعادلة. إذا تم تقييم مجموعة من الأسئلة لـ DIF عبر اللغات باستخدام تصميم شامل لمجموعة ثنائية اللغة (مثل تصميم رباعي - المجموعات المشروح سابقاً)، فإنه يمكن استخدام الأسئلة التي لا تظهر DIF وذلك لتطوير مجموعة من الأسئلة (المعيارية) المعتمدة التي يمكن استخدامها لربط صيغ لغوية متعددة لأسئلة أخرى بمقياس عام. مثلاً، استخدام طرق قياس (IRT) حيث يمكن قياس صيغتين لغويتين بشكل منفصل في الوقت نفسه (باستخدام ممتحنين أُحاديي اللغة)، ووحدات القياس للأسئلة المعتمَدة (معرّفة من حيث استخدام ثنائيي اللغة) يمكن أن يتم إجباره ليصبح متساوياً عبر المجموعتين اللغويتين. إن نتيجة هذه القيود هي إحداث مقياس عام لكل أسئلة الاختبار الأخرى (مفترضين بالطبع أن الافتراض اللا بعدي لـ IRT يتعلق بالبيانات، وأن مجموعة الأسئلة المعتمدة تمثل على نحو كاف المنشأ الذي تم قياسه). هناك بديل للطريقة المرتكزة على IRT وهي معايرة النموذجين اللغويين بشكل منفصل ثم استخدام الأسئلة المعتمَدة (المعيارية) لتعديل وحدات القياس من صيغة ما إلى مقياس الصيغة الأخرى. (مثلاً، انظر إلى الستوكينغ ولورد، 1983، الطريقة التحويلية). إن تعادل هذه الأسئلة، بالطبع، سيحتاج أيضاً أن يُحفظ على أساس التحليلات النوعية التي أجريت من قبل اختصاصي الاختبارات الثنائية اللغة. إن مجموعة الأسئلة المعتمَدة ستستلزم أن تكون ممثلة للاختبار بأكمله من حيث المميزات الإحصائية والضمنية.

هل تؤكد التحليلات الثنائية اللغة أن الأسئلة المنتقاة على أنها معتمدة هي فعلاً متعادلة عبر اللغتين؟ لسوء الحظ، هي ليست كذلك. إلا أنه باقترانها مع تطوير اختبار "الحالة- الفنية" وطرق تكييف الاختبار (هامبلتون، 1994، موليس وكيلي

وهالي، 1996، انظر أيضاً هامبلتون، المقطع 1، هذا المجلد)، يمكن إعطاء الكثير من الإثباتات لدعم استخدام هذه الأسئلة كأسئلة معتمدة. مثلاً، إن تصميماً ثنائي اللغة لمجموعة رباعية مقترناً بتطور اختبار صائب وطرق تكييفية يمكن أن تزودنا بالأنماط التالية من الدلائل المتعلقة للتعادل المختص بقياس سرعة ودقة العمليات العقلية للأسئلة المنتقاة على أنها معتمَدة:

- يعتبر معدو الاختبار أنه على الأسئلة أن تقيس التراكيب نفسها في كلا اللغتين.
- يُعتقد أن الأسئلة متعادلة من قِبل خبراء موضوعات البحث (مثلاً، علماء النفس أو خبراء المقررات التعليمية).
- يُعتقد أن الأسئلة متعادلة من قِبل الخبراء اللغويين.
- لا تظهر الأسئلة DIF لثنائيي اللغة الذين كانت لغتهم الأم لغة المصدر.
- لا تظهر الأسئلة DIF لثنائيي اللغة الذين كانت لغتهم الأم اللغة الهدف.

علاوة على ذلك، في حال توفر البيانات المعيارية المستقلة، فإن العلاقات الإحصائية بين الأسئلة والمعايير الخارجية يمكن أن تُدرس للتيقن من أن هذه العلاقات متشابهة عبر اللغات.

على الرغم من أن هذه الأنماط المتغايرة من الدلائل لا تؤكد أن الأسئلة هي نفسها في كلتا اللغتين، إلا أنها بالتأكيد تقدم برهاناً قوياً أن الأسئلة التي تستوفي هذه الشروط مناسبة لتكون أسئلة معتمَدة. من الواضح أن استخدام أسئلة كهذه كأسئلة معتمَدة يبرهن من خلال التصاميم السابقة أن معيار الأسئلة اللغوية المختلفة بشكل متوافق بدون استخدام سؤال معتمَد أو جعل التحولات المعيارية مرتكزة على أسئلة تستوفي عدداً قليلاً فقط من هذه المعايير.

معالجات تحليل البيانات

في القسم السابق أُشير إلى أن الأسئلة التي حُددت على أنها متعادلة بما يتعلق بقياس سرعة ودقة العمليات العقلية يمكن أن تستخدم للمساعدة في تشكيل معيار عام أو مشترك عبر نموذجي اختبار بلغتين مختلفتين. وفي هذا القسم، يتم شرح عدة خيارات لإنجاز هذه المعايرة.

افترض أن هذه الحالة اعترضت مختصاً معنياً بقياس سرعة ودقة العمليات العقلية الذي أكمل سلسلة من الدراسات الشاملة المتعلقة بالـ DIF باستخدام ثنائيي اللغة. بناء على المقياس الإحصائي والحاسم المنصوص سابقاً، فإن هذا المختص قد حدد مجموعة من الأسئلة ليتم استخدامها كأسئلة معتمَدة. وقد قام هذا المختص بتحصيل بيانات عن كلا الصيغتين اللغويتين للاختبار من المجموعتين المعنيتين أحاديتي اللغة. يتوفر لدينا هنا خياران. الأول، يمكن للمختص المعني بقياس سرعة ودقة العمليات العقلية أن يعاير في الوقت نفسه الأسئلة اللغوية المختلفة على مقياس عام؛ وذلك بإجبار وحدات القياس للأسئلة المعتمدة أن تكون متعادلة عبر المجموعتين اللغويتين. سيتم تقدير القياس للأسئلة الأخرى بشكل منفصل لكل صيغة لغوية للسؤال. السؤال الثاني. هو أن يتم القياس بالمزيد من تحليلات DIF، وذلك باستعمال الأسئلة المعتمدة لتشكيل مقياس عام عبر المجموعتين اللغويتين. بعد أن يتم تحديد الأسئلة التي تعمل بشكل مختلف عبر اللغتين، يمكن أن تتم معايرة الاختبار وذلك بتوجيه كل الأسئلة الأخرى (مثلاً، الأسئلة التي ليست DIF) إلى أن تكون متعادلة عبر اللغتين. إن المعالجة التفاعلية تستخدم كلاً من المتحنين ثنائيي اللغة وأحاديي اللغة وذلك لربط الفحوص اللغوية المختلفة إلى مقياس عام.

بافتراض هذين البديلين، وبسبب نقص أساس بحثي قوي للاختيار بينهما، فإن الخيار الثاني يبدو أفضل كونه يقوم بالمزيد بالتحليلات DIF باستخدام طريقة IRT، فإن هذه الطريقة ستقوم بتقدير وحدات القياس بشكل منفصل للمجموعتين

اللغويتين المختلفتين لكل الأسئلة ما عدا الأسئلة المعتمدة (التي توجه وحدات قياسها إلى أن تكون متساوية).

عند ذلك يمكن أن يتم تقدير هذه الأسئلة لـ DIF باستعمال طريقة نسبية راجحة (مثالاً، سيرسي بيربيرولو، 2000، سيسن، ستنبيرغ، ووينر، 1988، 1993) إلا أن الطرق التي لا ترتكز على IRT، DIF يمكن أيضاً أن تطبق باستعمال أداء الممتحنين في الأسئلة المعتمدة كالمتغيرات المتطابقة (مثلاً: آلالوف وآل، 1999، بادجل وآل، 1995، سيرسي وآلالوف، 2003) إن المعايرة النهائية المتواصلة سوف تقدر وحدات القياس لأي أسئلة لا تظهر DIF في التحليلات السابقة بشكل منفصل لكل مجموعة لغوية، وسوف توجه وحدات القياس لتكون متساوية لتلك الأسئلة التي لم تظهر DIF آنغوف وكوك، 1988على الرغم من أن طرق معالجة نظرية الاختبار الكلاسيكية يمكن أن تستخدم للمعايرة النهائية، فإن معايرة IRT هي المرجحة (افتراض المدركات اللا بعدية)، بتقديم مميزاتها الإحصائية هامبلتون وآل، (1991).

على الرغم من أن هذه الفكرة مغرية نظرياً، فإن تطبيقات هذه الطريقة سوف تساعد بتحديد فائدتها. إنه لمن المهم أن نلاحظ أنه بغض النظر عن الاستراتيجية التحليلية البيانية المنتقاة، فإن صدق مجموعة الأسئلة المعتمدة هو شيء حاسم. يفترض التحليل الموجز هنا أن مجموعة الأسئلة المستخدمة لاعتماد المقياس عبر اللغات مناسب. إن صدق هذا المعتمد يجب أن يتم دعمه باستخدام معايير خارجية كأحكام خبراء موضوعات البحث وتحليلات تعادل التركيب سيرسي وآل، (1999) إن الأسئلة المعتمدة يجب أن يتم اعتبارها كممثلة لأشكال الاختبار الكامل من حيث المميزات الإحصائية والضمنية.

هناك خيار آخر لم يتم تطبيقه بعد على مجال تقييم التقاطعات اللغوية، ألا وهو استعمال الأسئلة المعتمدة كواحدة من المقاييس العدة لمطابقة الخاضعين للاختبار بلغات مختلفة. مثلاً، إن الممتحنين في مجموعات لغوية مختلفة يمكن أن

تتم مطابقتهم على معيار المطابقة المتعددة التغيرات التي تشكل الأداء في الأسئلة المعتمدة، درجات المناهج المعنية، الحالة الاجتماعية الاقتصادية، والمعتقدات المتغيرة الأخرى المرتبطة بالتركيب. أشار سيرسي (1997) أن ميل الدرجات يمكن أن يستخدم لمزاوجة الممتحنين في هذا السياق. بالإضافة إلى ذلك تم إجراء دراسات DIF عبر المجموعات التي تجيد اللغة نفسها (مثلاً. الإناث والذكور) باستخدام المنطق الرمزي للتراجع في سلوك الفرد؛ وذلك لتكييف التحليلات باستخدام متغيرات متعددة (كلوزر، نانجستر، مازور وريبكي، 1996، مازر، كانجي وكلورزر 1995). تحوي هذه الإستراتيجية وعداً لربط مقاييس الدرجات عبر الصيغ اللغوية المختلفة لاختبار ما.

الاستنتاجات

في هذا المقطع، تم تنقيح استعمال التصاميم البحثية المتضمنة ممتحنين ثنائيي اللغة وذلك لتقدير الاختبارات المقدمة بعدة لغات. إن التقنيات المتعلقة بقياس وسرعة ودقة العمليات العقلية في هذا المجال هي فقط بطور التطور؛ لذلك هناك حاجة للمزيد من البحوث التجريبية. بالطبع، قد لا يتوفر ممتحنون ثنائيو اللغة في كل حالات تقييم التقاطع اللغوي. ولكن في تلك الحالات حيث يمكن إدراج ثنائيي اللغة في التصميم البحثي يمكن جمع دلائل أكبر تتعلق بالاختبار ومقارنة الأسئلة. بافتراض الدروس التي تم تعلمها في هذا التنقيح، تقدم بعض الاقتراحات باستخدام ثنائيي اللغة لتحسين تعادل الاختبار عبر اللغات.

أولاً، إن دراسات تعادل الاختبار عبر اللغات يجب أن يتضمن تحليلات لكل من الممتحنين الثنائيي اللغة والأحاديي اللغة. تعد التصميمات التي تستخدم ثنائيي اللغة مفيدة بشكل استثنائي لتقييم تعادل الأسئلة عبر مجموعة عادية للخاضعين للاختبار. يجب أن تقدم نتائج هذه التحليلات دلائل قيمة لاختيار الأسئلة المعتمدة ليتم استخدامها في التحليلات اللاحقة. ولكن يجب ألا تتخذ القرارات المتعلقة

بالإدارة والمجموعة وتطور الاختبار فقط على أساس التحليلات التي تستخدم ثنائيي اللغة. بل إن هذه التحليلات يجب أن تكون جزءاً من دراسة أكثر شمولاً والتي تقيّم بدورها أداء المجموعات الأحادية اللغة في كل صيغة لغوية من الاختبار، ويجب أن تتحرى عن العلاقات بين نقاط الاختبار ونقاط الأسئلة ومتغيرات أخرى ضمن شبكة علم القوانين الطبيعية والمنطقية المتعلقة بالمنشأ المقاس.

ثانياً، حين يتم استخدام ثنائيي اللغة لتقدير الاختبار وتعادل الأسئلة عبر اللغات، فإنه يجب ألا تعامل المجموعات الثنائية اللغة كمجموعة واحدة. كحد أدنى، إن أداء المجموعتين ثنائيتي اللغة الذي يمثل الهيمنة في كل من اللغتين يجب أن يتم التحري عنها. لهذا، هناك ميزة رئيسة لتصاميم البحث الثنائية اللغة ألا وهو الآلية لتصنيف ثنائيي اللغة على مجموعتين أو أكثر، بالإضافة إلى التصديق على أنهما متمكنتين من كلتا اللغتين (بيكر 1988). لا يوصى بتصاميم الثنائية اللغة أحادية المجموعة حيث يخضع ثنائيو اللغة إلى كلتا الصيغتين اللغويتين للاختبار أو الأسئلة، ويعود ذلك إلى مشكلات تتعلق بالتعب والحافز وأثر الممارسة.

ثالثاً، إن فوائد استخدام صيغ الاختبار المختلطة اللغة بدلاً من صيغ الاختبار الأحادية اللغة يجب أن تؤخذ بعين الاعتبار في التصاميم الثنائية اللغة. تقوم الصيغ اللغوية المختلطة بجمع بيانات عن اللغتين من كلتا المجموعتين ثنائيتي اللغة. تعتبر هذه الطريقة نافعة لأنها تمنح لكل الممتحنين الفرصة لإثبات مهارتهم في مجال الموضوع الذي يتم اختباره في كلتا اللغتين. ولكن إذا أولينا الأهمية الكبرى للآثار التفاعلية للّغة الأصلية في لغة الاختبار، فإن التصميم الذي يستخدم صيغاً لغوية منفصلة يمكن أن يكون هو الراجح. في كلتا الحالتين، يجب أن تكون المجموعات التي تخضع لصيغ الاختبار متعادلة بشكل عشوائي (عبر الفروض العشوائية أو أشكال الاختبار الحلزوني). علاوة على ذلك، فإن افتراض التعادل العشوائي يمكن أن يتم قياسه باستخدام عدة أسئلة شائعة (يفضل في كلتا اللغتين) في كلتا الصيغتين.

رابعاً، إن أثر الـ DIF عبر اللغات يجب أن يتم تقديره مع الأخذ بعين الاعتبار مجالات المضمون المتعددة المدرجة في التقدير. إذا كانت أسئلة الـ DIF مترافقة بشكل سائد مع بعض مجالات المضمون، فإن المقارنات عبر المجموعات اللغوية فيما يتعلق بمجالات المضمون قد لا يمكن تبريرها. بالنظر إلى نماذج الـ DIF ضمن مجالات المضمون سيظهر قصور على أنماط استنتاجات التقاطعات اللغوية التي يمكن تطبيقها.

كانت مقاييس الاختبارات النفسية والتربوية (جمعية البحوث التربوية الأمريكية [AERA]، الجمعية النفسية الأمريكية [ABA]، والمجلس الوطني للقياس في التعليم [NCME]، 1985) واضحة في طلب حقائق عن مقارنة الاختبارات المقدمة بعدة لغات: "حين يكون القصد أن صيغتي اختبارات اللغة المزدوجة قابلة للمقارنة، فإنه يجب أن يورد إثبات مقارنة الاختبار ص.75". تم تأكيد هذا النموذج بشكل خاص في التنقيح الحديث لهذه النماذج:" حين يقصد من الصيغ اللغوية المتعددة لاختبار ما أن يكون قابلاً للمقارنة فيجب على مطوري الاختبار أن يقوموا بالتبليغ عن إثبات مقارنة الاختبار" (AERA، ABA، NCME، 1999، ص99). إن التغيير من "صيغتين" إلى صيغ "متعددة" يجب أن يسلم بالتزايد الكبير في تقييم التعدد اللغوي خلال الـ 15 السنة الماضية منذ أن نشرت الطبعة الأخيرة لمقاييس الاختبار. إن الاقتراحات الموجزة في هذا المقطع يجب أن تساعد الباحثين مطوري الاختبارات ببذل أفضل ما يمكن في تقييم الصيغ اللغوية المختلفة لاختبار ما وفي تقديم شاهد لمقارنة الدرجات.

إن بعض الاختبارات، مثل المسابقات الرياضية البارزة في عالم الأولمبياد، تتجاوز الحواجز اللغوية. لسوء الحظ، إن مقارنات المعرفة والمهارات النفسية الأخرى هي إجمالاً لا تقاس باستخدام التقييمات "المستقلة لغوياً". هناك على الأغلب عوامل كثيرة تتعلق بتقييم التقاطع اللغوي للاستنتاج المطلق. إن اختلافات

الاختبار يمكن أن تكون منفصلة بشكل كامل عن اختلافات المجموعات اللغوية. لذا، يجب أن نبذل كل ما بوسعنا لتفسير التأثيرات اللغوية حين نقوم بالمقارنات للأفراد الذين يتكلمون لغات أخرى. إن دراسة أداء الاختبار لثنائيي اللغة في اختبارات اللغة المزدوجة تقدم إطاراً واعداً لتقييم هذه التأثيرات.

*****

المراجع

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55–73.

Allalouf, A., Hambleton, R K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36,* 185–198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and*
psychological testing. Washington, DC: American Educational Research Association.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Report No. 88-2). New York: College Entrance Examination Board.

Baker, C. (1988). Normative testing and bilingual populations. *Journal of Multilingual and Multicultural Development, 9,* 399–409.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli Universities. *Educational Measurement: Issues and Practice, 13,* 12–20.

Boldt, R. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers* (College Entrance Examination Board Research and Development Report 68-69, No. 3). Princeton, NJ: Educational Testing Service.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202–214.

CTB/McGraw-Hill. (1988). *Spanish assessment of basic education: Technical report.* Monterey, CA: McGraw-Hill.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74,* 912–920.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177–184.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3&4), 199–215.

Foster, D., Olsen, J. B., Ford, J., & Sireci, S. G. (1997, March). *Administering computerized certification exams in multiple languages: Lessons learned from the international marketplace.* Paper presented at the meeting of the American Educational Research Association, Chicago.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6.* 304–312.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57–68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross -national ssurveys of educational achievement* (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests [Special Issue]. *Language Testing, 20*(2), 127–240.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–157.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guideleines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1–16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818–825.

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology, 71,* 83–94.

International Association for the Evaluation of Educational Achievement. (1994). *TIMSS main study manuals: Population 1 and 2.* Hamburg, Germany: Author.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.

Mazor, K. M., Kanjee, A., & Clauser, B. (1995). Using logistic regression and the Mantel–Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131–144.

Mislevy, R. J., & Bock, R. D. (1990). *PC-BILOG 3: Item analysis and test scoring with binary logistic items.* Mooresville, IN: Scientific Software.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation verification procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection* (pp. 1–14). Chestnut Hill, MA: Boston College.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Pennock-Román, M. (1995). *Measuring developed academic abilities using Spanish- versus English-language tests: PAEG/GRE relationships for Puerto Ricans who are more proficient in Spanish than in English* (GRE Report No. 89-01). Princeton, NJ: Educational Testing Service.

Prieto, A. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*(1), 12–19, 29.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 147–165.

Sireci, S. G., & Berberolu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education, 13*(3), 229–248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Sireci, S.G., Xing, D., & Fitzgerald, C. (1999, April). *Evaluating translation DIF across multiple groups: Lessons learned from the Information Technology industry.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

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

Terman, L. M. (1916). *The measurement of intelligence.* Boston: Houghton-Mifflin.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 147–169). Mahwah, NJ: Lawrence Erlbaum Associates.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Norwood, NJ: Ablex.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1977). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29–37.

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47*(4), 263–279.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment, 3,* 1–16.