4

طرق إحصائية لتحديد العيوب في عملية تكييف الاختبارات

ستيفن ج. سيرغي
جامعة مستشوست/ أمهرست

ليان باتسولا
خدمات الاختبارات التربوية

رونالد ك. هامبلتون
جامعة مستشوست/ أمهرست

عند القيام بأبحاث علمية عبر الثقافات فمن الواجب على الباحثين التدقيق في وسائل التقييم المستخدمة للتأكد من خلوها من أية ترجمات/ تكييفات مثيرة للجدل (ملحوظة مهمة: في مجال التقييم عبر اللغات. يعتبر المصطلح adaptation "تكييف" أجدر بالتفضيل من المصطلح "translation" ترجمة" كونه لا يدل ضمناً على ترجمة حرفية. وتعد عملية تكييف الاختبارات بصورة نموذجية أكثر مرونة حيث تسمح باستعاضات لفظية أكثر تعقيداً بحيث يكون المعنى المقصود مصوناً عبر اللغات. حتى وإن لم تكن الترجمة حرفية برمتها (غيسينغر 1994). في هذا الفصل. يتم استخدام المصطلحين بصورة متبادلة؛ لأن كثيراً من القراء حديثي العهد بهذا المجال يمكن أن يكونوا غير ملمين بالمعنى المراد من المصطلح "adaptation" تكييف.

وبصورة خاصة عندما تقتضي الضرورة مقارنة نتائج اختبار مأخوذة من ثقافات مختلفة. فمن الواجب على الباحثين التدقيق في وسائل التقييم المستخدمة للتأكد من خلوها من انحيازات في المفهوم والطريقة والسؤال (انظر مثلاً فان دي فيفر ولونغ 1997. فان دي فيفر وتانزر 1997). لأنه في حال وجود مثل هذه الانحيازات ولم يتم تحديدها فإن الاستنتاجات المقارنة عبر الثقافات لن تكون صحيحة؛ لهذا السبب فإن كلاً من مؤلفي "معايير في الاختبار التربوي والنفسي" (جمعية الأبحاث التربوية الأميركية. الجمعية النفسية الأميركية. والمجلس الوطني للقياسات في التربية 1999) و"إرشادات في تكييف الاختبارات التربوية والنفسية" (انظر الفصل الأول من هذا الكتاب) يطالبان الباحثين عبر الثقافات أن يقدموا برهاناً على قابلية المقارنة بين نصوص لغوية مختلفة لتقييم ما عندما يراد أن تكون درجات النصوص المختلفة منه قابلة للمقارنة فيما بينها.

على الرغم من وجود استراتيجيات حكمية (نوعية) وإحصائية لتحديد ومعالجة الانحياز. فإن هذا الفصل يركز على التقنيات الإحصائية في معالجة الانحيازات في المفهوم والطريقة والسؤال التي يمكن ظهورها في عمليات التقييم عبر اللغات. ينقسم هذا الفصل، الذي يتوسع في العديد من الأفكار المطروحة من قبل فان دي فيفر و بورتينغا في الفصل الثاني من هذا الكتاب، إلى ثلاثة أقسام.

يتضمن القسم الأول وصفاً للتقنيات الإحصائية في تقييم تكافؤ المفاهيم. ويعرض القسم الثاني إستراتيجيات لتقييم ومعالجة الانحياز في الطرائق. وفي القسم الثالث، ندرج ونناقش الطرائق التقليدية والحديثة في تحديد الانحياز في الأسئلة. يلقي الجدول 1-4 نظرة عامة على الطرائق والأمثلة المشروحة في هذا الفصل. يتبع هذا الجدول خطة التصنيف المطروحة في فان دي فيفر و تانزر (1997) واللذين قاما بتقسيم مصادر الانحياز الشائعة في عمليات التقييم عبر الثقافات إلى هذه الفئات الثلاثة.

تقنيات إحصائية لتقييم تكافؤ المفاهيم:

يمكن للباحثين عبر الثقافات استخدام التقنيات الإحصائية معاً قبل وبعد الاختبار الميداني لتقدير تكافؤ المفاهيم ووسائلهم التقييمية. وبافتراض عدم توفر اختبارات أو بيانات بدرجة السؤال قبل الاختبار الميداني، فإن الباحثين محصورون ضمن كمية المعلومات الممكن جمعها؛ لهذا السبب، فإن غالبية الأبحاث حول تكافؤ مفاهيم وسائل التقييم المترجمة قد أجريت على الاختبارات الميدانية والبيانات العملية.

قبل الاختبار الميداني:

في حال عدم توفر بيانات بالإجابات على الأسئلة، فإنه يمكن معاينة تكافؤ مفاهيم نصوص لغوية مختلفة لاختبار معين؛ وذلك بجمع البيانات من خبراء في موضوع البحث يمثلون اللغات والثقافات المختلفة المراد دراستها. وبصورة مماثلة لدراسات صحة المضامين التي تجري في الاختبارات التربوية، فإنه يمكن تصنيف الأسئلة في تقييم معين بالاعتماد على معيار أو أكثر بهدف إلقاء الضوء على المفهوم المقاس. من أحد الأمثلة المبتكرة لاستخدام خبراء في موضوع البحث لتقدير التكافؤ في المفهوم هي الدراسة التي أعدها هيو و ترياندیس في العام 1985 . في هذه الدراسة، قامت عينات صغيرة من الحكام المنتمين إلى ثقافات مختلفة بتقدير "التشابه في المعاني لأزواج من الأسئلة المستخدمة في اختبار معين" (صفحة 208). استخدم المؤلفان قياساً متعدد الأبعاد للاختلافات الفردية للكشف عن الخاصيات التي اعتمدها الحكام في تصنيفاتهم للتشابه. في حال كون الخاصيات المستخدمة في تقدير التشابه بين الأسئلة متساوقة لدى جميع الحكام، فهذه دلالة أولية على تكافؤ المفاهيم. وفي حال استخدم الحكام المنتمين إلى ثقافات مختلفة لخاصيات مختلفة، فيمكن الاستفادة من هذه المعلومة في تعديل نسخة أو أكثر من الاختبار.

تبين الدراسة التي أعدها هيو وترنياندیس في العام 1985 إحدى وسائل جمع البيانات عبر الثقافات لتقييم معين قبل البدء بإدارته. وعلى الرغم من توفر أبحاث قليلة في هذا المجال، فإن تصاميم أخرى تستخدم خبراء مضامين من بيئات لغوية

مختلفة، أو خبراء مضامين ثنائيي اللغة أيضاً ممكنة. فعلى سبيل المثال، يمكن الطلب من الخبراء ثنائيي اللغة تقدير التشابه في الصعوبة لأسئلة من اختبار إنجازات. ويمكن لهذا الإجراء تحديد أسئلة يمكن ترميزها لاحقاً لاحتوائها على انحياز إذا ما أجريت دراسات للوظائف التفاضلية للأسئلة (DIF) بعد القيام بإدارة الاختبار.

الجدول 1-4

مصادر الانحياز في تكييفات الاختبارات وبعض المراجع الأساسية

مصدر الانحياز	الوصف	المراجع
الانحياز في المفهوم	المفهوم غير وثيق الصلة في جميع الثقافات (تكافؤ مفاهيمي)، المفهوم غير محدد عملياً بصورة منسجمة عبر الثقافات، قياس المفهوم غير منسجم عبر الثقافات.	
الانحياز في الطريقة	انحيازات في ظروف إدارة الاختبار، عدم الإلمام بأشكال الاختبار في ثقافة أو أكثر، أساليب إجابة تفاضلية (مثلاً: الرغبة الاجتماعية)؛ عدم القدرة على مقارنة العينات (انحياز في الانتقاء)؛ تأثيرات مجري المقابلة (مثلاً: تأثيرات التعميم).	
الانحياز في السؤال	ترجمة خاطئة، وثاقة الصلة التفاضلية للأسئلة عبر الثقافات، عوامل الإزعاج.	

بعد الاختبار الميداني:

بعد الاختبار الميداني، عندما تتوفر بيانات إجابة للممتحنين، فإنه توجد أربعة أساليب إحصائية على الأقل لتقييم تكافؤ المفاهيم عبر وسائل التقييم: التحليل الاستطلاعي للعوامل، التحليل الإثباتي للعوامل، القياس المتعدد الأبعاد، ومقارنة الشبكات المنطقية. نقدم في هذا القسم شرحاً موجزاً لكل أسلوب.

التحليل الاستطلاعي للعوامل:

يعد التحليل الاستطلاعي للعوامل أحد أقدم الطرائق و أكثرها شعبية لتقدير ما إذا كانت النصوص اللغوية المختلفة لاختبار معين تقيس المفهوم نفسه. وفي الحقيقة، إن كلاً من فان دي فيفر و بورتينغا (1991)، و بورتينغا (1991) قد وصفا تحليل العوامل بأنه التقنية الإحصائية الأكثر استخداماً لتقييم ما إذا كان مفهوم معين في ثقافة ما موجوداً بنفس الصورة والتكرار في ثقافة أخرى (مثلاً: بوتشر وغارسيا 1978). يوظف أسلوب التحليل الاستطلاعي للعوامل عنصر تحليل العوامل أو بيانات درجات الاختبار بصورة مستقلة لكل فئة ثقافية. وتعاين مصفوفات تحميل العوامل بعد ذلك نظرياً لمعرفة التساوق عبر الفئات. وبالرغم من كون هذا الأسلوب جذاباً من الناحية الفطرية، فإن مقارنة بنى عوامل مستقلة لا يخلو من صعوبة، ولا توجد أية قواعد متفق عليها بصورة عامة لتحديد متى يمكن اعتبار هذه البنى متكافئة. لهذا السبب، فإن الأساليب الإحصائية خاصة تلك التي بمقدورها استيعاب فئات متعددة معاً، هي أساليب جديرة بالتفضيل. ويعتبر التحليل الإثباتي للعوامل والقياس متعدد الأبعاد ذو الأرجحية الأكثر أسلوبين من هذا النمط.

التحليل الإثباتي للعوامل:

تكون بنية الاختبار في التحليل الإثباتي للعوامل (CFA) مفترضة استنتاجاً وتستخدم بيانات الممتحنين لتقدير قابلية النجاح عملياً للبنية المفترضة (انظر مثلاً:

بايرن 1998، 2001، 2003). ويتم دمج هذه البنية المفترضة ضمن نموذج معادلة بنيوية وتجبر أن تكون متساوية عبر جميع الفئات. من إحدى فرضيات تكافؤ المفاهيم النموذجية المختبرة باستخدام التحليل الإثباتي للعوامل (CFA) هي ما إذا كانت مصفوفة تحميل العوامل متكافئة عبر جميع الفئات. وتكون بنية مصفوفة تحميل العوامل عادة "بنية مجموعات مستقلة" (مكدونالد 1985) تنص على أنه: (أ) كل متغير مقاس لا يساوي الصفر في التحميل (Loading) على العامل الذي صمم لقياسه فقط، (ب) العلاقات المتبادلة فيما بين العوامل (أي: الخط القطري السفلي للمصفوفة فاي (Δ) مقدرة بصورة حرة، (ج) الأخطاء المرتبطة بتحميلات العوامل (أي: مصفوفة ثيتادلتا) غير متبادلة العلاقة فيما بينها (مارس 1994).

لقد استخدم الباحثون في ميدان التقييم عبر اللغات التحليل الإثباتي للعوامل (CFA) لتقدير ما إذا كانت بنية العوامل لنص أصلي من تقييم معين متسقة عبر نصوص لاحقة مترجمة إلى لغة أخرى (مثلاً : براون وماركوليديس 1996، سيرسي، باستاري وآلالوف 1998). يعتبر التحليل الإثباتي للعوامل خياراً ملفتاً للانتباه لتقدير تكافؤ المفاهيم عبر وسائل مكيفة نظراً لقدرته على معالجة فئات متعددة معاً، ولتوفر اختبارات إحصائية لملاءمة النموذج، وللتزويد بدلائل وصفية لملاءمة النموذج. عندما يتألف تقييم معين من أسئلة مسجلة نتائجها بصورة ثنائية التفرع، فيمكن أن يكون التحليل الإثباتي للعوامل (CFA) مثيراً للجدل لأن النماذج الأساسية تكون خطية في طبيعتها بينما تكون العلاقات فيما بين الأسئلة ثنائية التفرع غير خطية (مكدونالد 1982) . بالرغم من ذلك، فإنه يمكن التغلب على هذا القصور بنظم هذه الأسئلة مع بعضها ضمن فئات قبل البدء بالتحليل. سيرد مثال على استخدام التحليل الإثباتي للعوامل (CFA) لتقدير تكافؤ مفاهيم نسخ لغوية مختلفة من اختبار معين في قسم لاحق.

القياس متعدد الأبعاد:

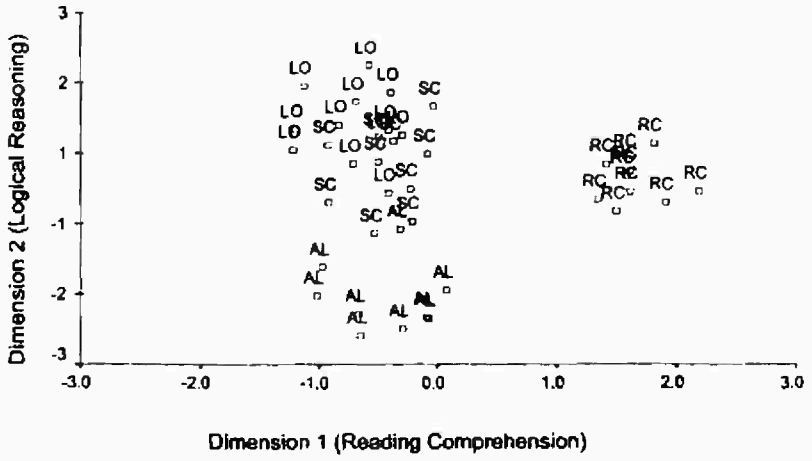
يعتبر القياس متعدد الأبعاد (MDS) أسلوباً آخر مسترعياً للانتباه لتقدير تكافؤ المفاهيم عبر نصوص لغوية مختلفة لاختبار معين.

وبصورة مماثلة للتحليل الاستطلاعي للعوامل، فإن تحليل القياس متعدد الأبعاد لا يتطلب تحديد بنية الاختبار استنتاجياً. بالرغم من ذلك، وبصورة مماثلة للتحليل الإثباتي للعوامل (CFA)، يمكن تحليل البيانات من فئات متعددة معاً . وباستخدام تحليل قياس متعدد الأبعاد للاختلافات الفردية، مثل نموذج (INDSCAL كارول و تشانغ 1970)، يمكن ملاءمة بنية مشتركة مع جميع الفئات معاً، ومن ثم يمكن تقييم الاختلافات البنيوية عبر الفئات بالنظر إلى "أوزان (العينات) الفئات"، والتي تستخدم لتعديل البنية المشتركة لكي تتلاءم بالصورة الأمثل مع بيانات كل فئة. يوفر القياس متعدد الأبعاد وسيلة لكشف الأبعاد التي ترتكز عليها بيانات إجابات الممتحنين، ولتقدير ما إذا كانت هذه الأبعاد متسقة عبر جميع الفئات (أو نصوص الاختبار) المراد دراستها. وهناك ميزة أخرى مثيرة للانتباه للقياس متعدد الأبعاد هي أنه لا يحتاج نموذجاً خطياً لاستنتاج البنية التي ترتكز عليها البيانات.

مثال على التحليل الإثباتي للعوامل (CFA) وتحليل القياس متعدد الأبعاد (MDS) لتكافؤ المفاهيم.

استخدم سيرسي، باستاري، آلالوف (1998) كلاً من التحليل الإثباتي للعوامل (CFA) والقياس متعدد الأبعاد (MDS) لتقدير تكافؤ مفاهيم أسئلة من قسم المحاكمات اللفظية لاختبار القبول بالاعتماد على قياس الذكاء (PET)، وهو عبارة عن اختبار تستخدمه الكليات والجامعات في إسرائيل لاتخاذ قرارات لقبول الطلاب (بيلر 1994؛ انظر أيضاً الفصل الثاني عشر من هذا الكتاب). يظهر الرسم التوضيحي 1.4 تمثيلاً ثنائي الأبعاد للأسئلة مستمداً من القياس متعدد الأبعاد (MDS).

تميل هذه الأسئلة إلى الانتظام مع بعضها في فئات في فضاء القياس متعدد الأبعاد (MDS) وفقاً لمواصفات المضامين (التشابهات، المنطق، الفهم عند القراءة، إكمال الجمل). يعد الرسم التوضيحي 4.2 أكثر أهمية، حيث يظهر أوزان الفئات على هذين البعدين ونفسيهما.

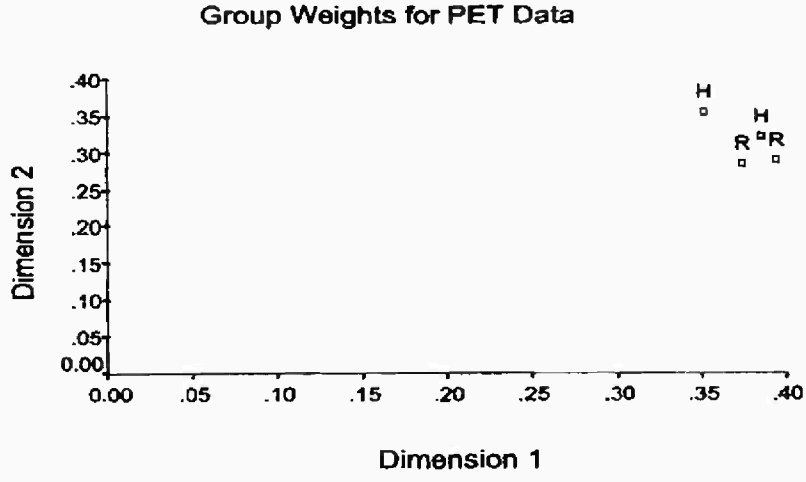


أوزان الفئات على بعدين

لقد تم في هذا التحليل استخدام بيانات فئتين من الممتحنين الذين خضعوا لاختبار النص العبري من الاختبار وفئتين أخريين ممن خضعوا لاختبار النسخة الروسية منه (كانت أحجام العينات حوالي 1300). وكما يتضح من الرسم التوضيحي 4.2، فإن أوزان الفئات كانت متشابهة كثيراً فيما بينها، الأمر الذي يوحي بتشابه البنى (تكافؤ المفاهيم عبر الفئات).

لقد قام سيرسي، باستاري، آلالوف (1998) أيضاً باستخدام التحليل الإثباتي للعوامل (CFA) لتقدير تكافؤ مفاهيم هذين النصين المختلفين من هذا الاختبار. وعملاً بمواصفات المضامين، فقد قاموا بملاءمة نموذج رباعي العوامل مع بيانات كلا الفئتين. لقد تم ملاءمة أربعة نماذج مختلفة من التحليل الإثباتي للعوامل (CFA) مع البيانات. قام النموذج الأول بإجبار العوامل الأساسية الأربعة على أن

تكون مشتركة عبر الفئات العبرية والروسية، وقام النموذج الثاني بإجبار مصفوفة تحميل العوامل على أن تكون ذاتها عبر الفئات وقام النموذج الثالث بإجبار الأخطاء المرتبطة بتحميلات العوامل هذه أن تكون ذاتها، أما النموذج الرابع فيحدد أن تكون العلاقات المتبادلة ضمن العوامل متكافئة. تتلخص نتائج تحليلهم في الجدول 2-4. في النماذج الأربعة جميعها، كانت جودة دلائل الملاءمة مرتفعة (96 . أو أكثر) بينما كانت بواقي جذور متوسطات المربعات منخفضة (076 . أو أقل). على الرغم من كون هذه النتائج متسقة مع تحليلات القياس متعدد الأبعاد، فقد تبين لدى استخدام بيانات من تقويم آخر أن الأمور لا تجري دوماً على المنوال نفسه. لذلك فقد أوصوا باستخدام كل من القياس متعدد الأبعاد (MDS) والتحليل الإثباتي للعوامل (CFA) معاً لتقدير تكافؤ المفاهيم لنصوص لغوية مختلفة من اختبار معين.



أوزان المجموعات حسب معطيات PET

مقارنة الشبكات المنطقية:

يعد تكافؤ المفاهيم مصطلحاً عاماً جداً يقول بأن المفهوم النفسي ذاته قابل للقياس عبر جميع الفئات المدروسة وبدقة متساوية في جميع الفئات. يمكن

لأساليب التحليل الاستطلاعي للعوامل، التحليل الإثباتي للعوامل، والقياس متعدد الأبعاد توفير برهان مهم على انسجام بنية الاختبار عبر نصوص لغوية مختلفة لتقييم معين. بالرغم من ذلك فإن البنية المتكافئة لا تقتضي بالضرورة مفاهيم متكافئة. فالتكافؤ البنيوي شرط أساسي ولكنه غير كاف لتكافؤ المفاهيم. لهذا السبب، فقد ارتأى الكثير من الباحثين تجاوز حدود دراسات التكافؤ البنيوي عند تقدير المفاهيم المقاسة عبر نصوص لغوية مختلفة لتقييم معين (مثلاً: فان دي فيفر و تانزر 1997). ويقترح هؤلاء الباحثون اعتماد أسلوب أكثر شمولية يوظف تحري العلاقات فيما بين درجات الاختبار ومتغيرات أخرى مفترضة العلاقة بالمفهوم المقاس.

في المقالة نفسها التي طرح من خلالها المصطلح "صحة المفهوم"، قام كرونباخ وميهل (1955) بطرح المفهوم "الشبكة المنطقية" أيضاً. لقد قاما باستخدام هذا المصطلح لإبراز الحقيقة القائلة بأنه لا يمكن إثبات صدق أي اختبار باستخدام معيار وحيد. بالأحرى، فقد برهنا على أن درجات الاختبار يجب أن تقدر ضمن نظام متعدد المتغيرات يأخذ بعين الاعتبار جميع مظاهر المفهوم المقاس. فيما يتعلق بالتقييم عبر الثقافات، فإن قابلية مقارنة العلاقات المتبادلة بين درجات الاختبار مع متغيرات أخرى يجب أن تكون منسجمة عبر الثقافات، بالإضافة إلى كونها متسقة عبر نصوص لغوية مختلفة بتقييم معين، حتى يكون تكافؤ المفاهيم متيناً. وبالتالي، فإن مقارنة الشبكات المنطقية عبر نصوص اختبارية هو تقييم صارم نظرياً لتكافؤ المفاهيم .

إن مقارنة العلاقات فيما بين درجات الاختبار والمعايير الخارجية المتعددة مهمة يصعب القيام بها في لغة واحدة، تزداد تعقيداً بوجود فئات ثقافية ونصوص اختبارية متعددة. إن تحديد وقياس المتغيرات الخارجية الصحيحة هما مجرد مشكلتين مهمتين علينا التغلب عليهما. لهذا السبب، فإنه من غير المفاجئ ندرة الدراسات الشاملة التي تقارن الشبكات المنطقية عبر الثقافات. بالرغم من ذلك،

فإنه يتوجب على الباحثين عبر الثقافات التدقيق في جدارة كل نص ثقافي لوسائلهم التقييمية والبحث معاً عن براهين صدق متقاربة ومميزة في كل مجموعة ثقافية.

خلاصة نتائج التحليل الإثباتي للعوامل

Model	*GFI*[a]	*RMSR*[b]
Four-factor model common for all groups	.97	.057
Equivalent factor loadings for all groups	.96	.060
Equivalent errors of factor loadings for all groups	.96	.066
Equivalent correlations among factors	.96	.076

[a]GFI = goodness of fit index.
[b]RMSR = root mean square residual

خلاصة:

لا يزال التحليل الاستطلاعي للعوامل أسلوباً شائعاً لتقييم تكافؤ المفاهيم عبر الثقافات. بالرغم من ذلك، يدرك مختصو الاختبارات الحاليون منافع كل من التحليل الإثباتي للعوامل (CFA) والقياس متعدد الأبعاد (MDS) لهذا الغرض. إن استخدام القياس متعدد الأبعاد (MDS) لتقويم تكافؤ المفاهيم عبر الثقافات آخذ بالرواج كونه ممكن الاستخدام قبل وبعد الاختبار الميداني، لا يكون أية افتراضات حول العلاقة فيما بين أسئلة الاختبار، لا يتطلب أن تكون البنية محددة استنتاجاً، وكونه يسمح بتقدير بنية الأبعاد لعدد من الاختبارات في وقت واحد. يعد التحليل الإثباتي للعوامل (CFA) ملفتاً للنظر كونه يمكن استخدامه لإثبات صحة بنية مفترضة معينة و لأنه يوفر نظاماً للاختبار الإحصائي للفرضيات المتنافسة فيما يتعلق ببنية الاختبارات. يوفر كل من القياس متعدد الأبعاد (MDS) والتحليل الإثباتي للعوامل (CFA) معلومات مهمة فيما يتعلق بتساوق بنية الاختبارات. عبر فئات ثقافية مختلفة ونصوص لغوية مختلفة لاختبار معين. بالرغم من ذلك، فإنه

يجب دراسة علاقات درجات الاختبارات بالمتغيرات الأخرى في جميع المجموعات الثقافية المراد دراستها كي نتمكن من تقييم تكافؤ المفاهيم عبر الثقافات على الوجه الأكمل.

استراتيجيات إحصائية لمعالجة وتقييم الانحياز في الطرائق:

بالإضافة إلى تقييم الانحياز في المفاهيم، فإنه يجب على الباحثين أيضاً تقييم الانحياز في وسائل التقييم عبر الثقافات. يشرح فان دي فيفر وتانزر (1997) أن الانحياز في الطرائق نابع من مصادر موجودة في قسم الطرائق للدراسات التجريبية ووفقاً لفان دي فيفر وتانزر فإن هناك ثلاثة أنواع للانحيازات في الطرائق: الانحياز في العينات، في الوسائل، وفي الإدارة. يشير الانحياز في العينات إلى الاختلافات الأساسية عبر الفئات الثقافية أو اللغوية والتي لا علاقة لها بالمفهوم المقاس (مثلاً: الاختلافات في الحوافز على الأداء الجيد، أو الوضع الاجتماعي الاقتصادي). ويشير الانحياز في الوسائل إلى عدم الاتساق في وظائف وسائل القياس عبر الفئات (مثلاً: الإلمام التفاضلي بأشكال الاختبار). أما الانحياز في الإدارة فيشير إلى المشكلات في الإدارة، كإجراءات إدارية غير قياسية (مثلاً: الخطأ في فهم التعليمات الإمتحانية من قبل مديري الاختبار في إحدى الفئات).

نقدم في هذا القسم شرحاً لبعض الإجراءات المتبعة في تقييم الانحياز في الطرائق.

معالجة الانحياز في العينات:

في حال اعتبار المجموعات الثقافية تختلف فيما بينها عبر متغيرات مهمة لا علاقة لها بالمفهوم المقاس، فيمكن استخدام تصاميم أبحاث شاملة وتحليلات إحصائية للتحكم بهذه المتغيرات "المزعجة". حيث يمكن استخدام تحليل مقدار التغاير، تصاميم قوالب عشوائية، وتقنيات إحصائية أخرى (تحليل التراجع، تبادل العلاقة الجزئي إلخ) لعزل تأثيرات مصادر التغيير غير المرغوب بها فيما بين

المجموعات. بالرغم من ذلك، فإن تحليلات كهذه تتطلب جمع بيانات حول هذه المتغيرات الخارجية والتأكد من أن افتراضات الإجراءات الإحصائية قد تم استيفاؤها (مثلاً: تجانس التراجع).

تقييم الانحياز في الوسائل والإدارة:

توجد على الأقل ثلاث استراتيجيات إحصائية لتقييم ما إذا كان الانحياز في الوسائل و/ أو الإدارة موجوداً بين الثقافات قيد الدراسة: الدراسات أحادية الميزة متعددة الطرائق، استخدام معلومات إضافية، واختبار التغيرات. ففي الدراسة أحادية الميزة متعددة الطرائق (فان دي فيفر وتانزر 1997)، يتم استخدام إجراءات تقييم متعددة لقياس الميزة ذاتها عبر المجموعات. وفي حال كون الاختلافات بين الفئات غير متسقة عبر طرائق التقييم المختلفة، فإن تقييماً أو أكثر يمكن أن يكون منحازاً.

يمكن استخدام معلومات إضافية أيضاً لتقدير الانحياز في الوسائل أو الإدارة. توظف هذه الاستراتيجية تحليل متغير ذي علاقة بالمفهوم قيد الدراسة. وفي حال كون الاختلافات الملحوظة عبر الفئات فيما يتعلق بالمعلومات الإضافية غير المتسقة مع الاختلافات الملحوظة فيما يتعلق بدرجات الاختبار، فإن الانحياز في الوسيلة أو الإدارة يمكن أن يوجد. من أحد الأمثلة على استخدام معلومات إضافية للكشف عن الانحياز في الوسيلة و/ أو الإدارة هو استخدام معلومات حول زمن الإجابات، حيث تتم مقارنة مقدار الزمن الذي يستغرقه الممتحنون من فئات مختلفة للإجابة عن سؤال معين (سيرسي، فوستر، أولسن، وروبن 1997). تجعل التقييمات الحديثة والتي تجري بواسطة الحاسب تلك المقارنات أكثر سهولة من أي وقت مضى. فباستخدام اختبارات إحصائية قياسية أو صياغة نماذج معادلات بنيوية أكثر تعقيداً (بايرون 2001)، فإنه يمكن تحديد ما إذا كانت أزمنة الإجابات مختلفة على نحو مهم عبر الثقافات. وفي حال وجود اختلافات عبر الثقافات في أزمنة الإجابات، فيمكن عندئذ السماح بإطالة الحدود القصوى للزمن المعطى لبعض الفئات أو جميعها.

استراتيجية ثالثة لكشف الانحياز في الوسائل و/أو الإدارة تعتمد على إعادة اختبار الممتحنين ضمن كل ثقافة (فان دي فيفر وتانزر 1997). حيث يمكن للاختلافات غير المتوقعة الناتجة عن التغيرات بين الاختبار وإعادته عبر الثقافات أن تعكس الانحياز في الوسائل والإدارة (مثلاً: فورمان، يوشيدا، سوانك، وغارسون 1989، فان دي فيفر، دال وفان زونيفيلد 1986). فعلى سبيل المثال، في حال وجود أرباح أكبر في الثقافة (أ) من الثقافة (ب)، فيمكن أن يكون ذلك إشارة إلى أن الثقافة (أ) لم تكن على قدر الإلمام بشكل الاختبار الذي كانت عليه الثقافة (ب) وبالتالي لم يكن أداؤها بنفس الجودة في الاختبار الأولي وحازت على درجات منخفضة فيه. يجب إجراء مثل هذه الدراسات إذا كان هناك أي شك بوجود إلمام تفاضلي بصيغة الاختبار. وفي حال وجود إلمامات تفاضلية كهذه، فإنه يجب أن تتلقى كل ثقافة تعريفاً كافياً بظروف إدارة الاختبار وأشكال الأسئلة قبل أن تتم مقارنة درجاتها.

خلاصة:

يعد تقييم وجود أي انحياز في الطريقة بين ثقافات مختلفة خطوة غالباً ما يتم إهمالها في الدراسات عبر الثقافات. على الرغم من ذلك، فإنها خطوة مهمة. في حال وجود انحياز في الطريقة ولم تتم معالجته، فإن نتائج الدراسة ستكون مضللة. من ناحية أخرى، في حال إمكانية الكشف عن الانحياز في الطريقة ومعالجته باستخدام تحليلات إحصائية أو من خلال تعريف الممتحنين بوضع التقييم، فيمكننا عندئذ الانتقال إلى الخطوة التالية في تقدير قابلية مقارنة وسائل القياس عبر الثقافات؛ أي تقييم تكافؤ الأسئلة.

تقنيات إحصائية لتقييم الانحياز في الأسئلة

قبل البدء بمناقشة التقنيات المتبعة في تقييم الانحيازات في الأسئلة، يجب علينا أولاً التمييز بين ثلاثة مصطلحات مهمة ولكنها مستقلة: تأثير السؤال، الوظيفة التفاضلية للسؤال (DIF)، والانحياز في السؤال. يشير تأثير السؤال إلى اختلاف مهم بين الفئات على سؤال معين. على سبيل المثال، عندما تملك إحدى الفئات نسبة أعلى من الممتحنين الذين أجابوا على سؤال معين بصورة صحيحة من فئة أخرى. ويمكن لتأثير السؤال أن يكون ناتجاً عن اختلافات حقيقية في الكفاءة بين الفئات أو نتيجة لانحياز فيه. تسعى تحليلات الوظائف التفاضلية للأسئلة (DIF) إلى معرفة ما إذا كان تأثير سؤال معين ناتج عن الاختلافات الكلية بين الفئات في الكفاءة أو نتيجة للانحراف فيه. للقيام بذلك، تتم المطابقة بين ممتحنين من فئتين مراد دراستهما بالنسبة للكفاءة المراد قياسها. يجب على الممتحنين ذوي الكفاءة المتساوية والمنتمين إلى فئات مختلفة الإجابة بنفس الصورة على سؤال الاختبار المعطى. في حال عدم الإجابة بصورة مماثلة، فيمكن القول بأن السؤال يؤدي وظيفة مختلفة عبر الفئات.

تعد تحليلات تأثير السؤال والوظيفة التفاضلية له إحصائية في طبيعتها. بينما تكون تحليلات الانحياز في السؤال، من ناحية أخرى، نوعية بصورة أساسية. ويمكن القول بأن سؤالاً معيناً يعتبر منحازاً ضد فئة معينة عندما يكون أداء الممتحنين من تلك الفئة أكثر رداءة في الإجابة على السؤال المتصل بالممتحنين في الفئة المرجع والذين هم من الكفاءة نفسها، ويكون سبب الأداء المتدني لا علاقة له بالمفهوم الذي ينوي الاختبار قياسه. لهذا السبب، يشترط لوجود انحياز في سؤال معين تحديد ميزة خاصة بالسؤال تكون غير منصفة لفئة أو أكثر (مثلاً: مفهوم مألوف بصورة أكثر للممتحنين من إحدى الفئات دون غيرها عندما يكون المفهوم ذاته غير ذي أهمية بالنسبة للمهارة المراد تقييمها). وبالتالي، فإن التقنيات الإحصائية لتحديد

الانحيازات في الأسئلة تفتش عن الأسئلة التي تؤدي وظائف مختلفة عبر المتحنين الذين ينتمون إلى فئات مختلفة ولكنهم من كفاءة متساوية. وحالما يتم تحديد هذه الأسئلة، فإنها تخضع إلى اختبار نوعي لتفسير الاختلافات الملحوظة. وعندما يتضح أن لا علاقة لتفسير هذا الاختلاف بالهدف من الاختبار، فإن السؤال يصنف على أنه "منحاز".

نقدم في هذا القسم شرحاً للعديد من أكثر الطرائق شيوعاً والتي تم استخدامها لكشف الانحياز في أسئلة الاختبارات المسجلة درجاتها بصورة ثنائية. يمكن العثور على دراسات أكثر شمولية لاستخدام طرائق الوظائف التفاضلية للأسئلة (DIF) في تحليل البيانات الثنائية في كاميلي وشيبرد (1994)، كلاوسر ومازر (1998)، هولاند و واينر (1993)، ميلسان و إيفرسون (1993)، بوتينزا و دورانس (1995)، وسيرسي وآلالوف (2003). يظهر الجدول 4-3 قائمة بطرائق الوظائف التفاضلية للأسئلة (DIF) في تحليل بيانات الاختبارات. ويوفر هذا الجدول استشهادات بكل طريقة ويشير إلى أنواع البيانات المناسبة لكل طريقة. يحال القراء إلى بينفيلد و لام (2000) بغية الحصول على مراجعة شاملة لطرائق إجراء دراسات الوظائف التفاضلية للأسئلة (DIF) لبيانات إجابات متعددة التفرع.

هناك تطبيقات عديدة لمنهجية الوظائف التفاضلية للأسئلة (DIF) في معالجة مشكلة تقدير الأسئلة المترجمة/المكيفة (مثلاً: ألالوف، هامبلتون، وسيرسي 1999، أنغوف و كوك 1988، بادجيل، رادجو، وكوارتيتي 1995، سيرسي و بيربيروغلو 2000). الطرائق المختارة للمناقشة في هذا الفصل تمثل الطرائق الأكثر استخداماً في المؤلفات العلمية حول تكييفات الاختبارات. الطرائق المبحوثة هي: الرسم البياني للدلتا، التقييس (التوحد القياسي)، مانتل-هاينزل، والطرائق المبنية على نظرية الإجابة على سؤال .(IRT)

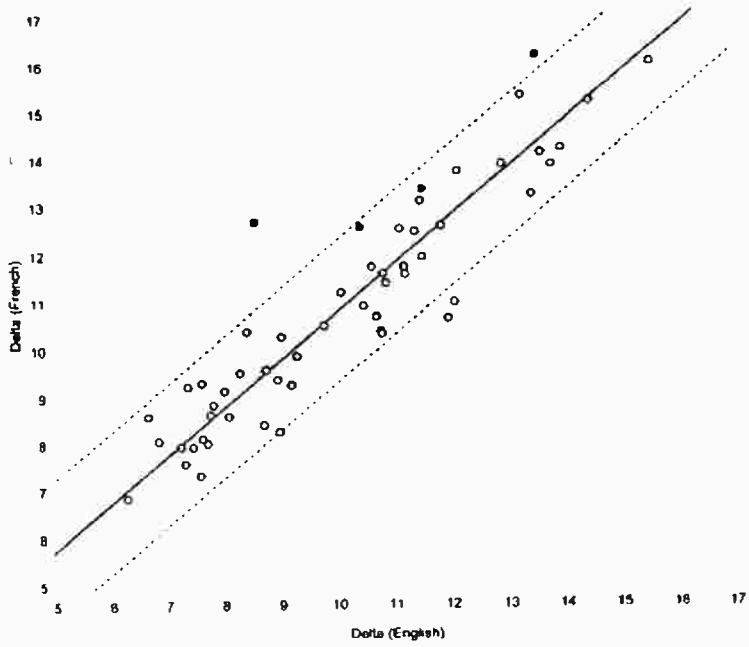
Method	*Sources*	*Appropriate for*	*Applications to Cross-Lingual Assessment*
Delta Plot	Angoff (1972, 1993)	Dichotomous data	Angoff & Modu (1973) Cook (1996) Muniz et al. (2001) Robin, Sireci, & Hambleton (2003)
Standardization	Dorans & Kulick (1986); Dorans & Holland (1993)	Dichotomous data	Sireci, Fitzgerald, & Xing (1998)
Mantel–Haenszel	Holland & Thayer (1988); Dorans & Holland (1993)	Dichotomous data	Allalouf et al. (1999) Budgell et al. (1995) Muniz et al. (2001)
Logistic Regression	Swaminathan & Rogers (1990)	Dichotomous data Polytomous data Multivariate matching	Allalouf et al. (1999) Gierl et al. (1999)
Lord's Chi-Square	Lord (1980)	Dichotomous data	Angoff & Cook (1988)
IRT Area	Raju (1988,1990)	Dichotomous data Polytomous data	Budgell et al. (1995)
IRT Likelihood Ratio	Thissen et al. (1988) Thissen et al. (1993)	Dichotomous data Polytomous data	Sireci & Berberoglu (2000)
SIBTEST	Shealy & Stout (1993)	Dichotomous data	

طريقة الرسم البياني للدلتا:

بالنسبة لأسئلة الاختبارات المسجلة درجاتها بصورة ثنائية (مثلاً: صواب/خطأ)، فإن رسماً بيانياً بسيطاً للتشتت) للإحصائيات الصحيحة للنسبة (قيم النسبة) لكل سؤال غالباً ما يعطي معلومات كافية كاختبار أولي لوظائف الأسئلة عبر اللغات أو الثقافات. (في حالة الإجابات الثنائية على أسئلة نفسية، فإن قيم النسبة تعبر عن نسبة الأشخاص المتفقين مع السؤال). ولإنشاء رسم بياني كهذا، يتم تمثيل قيم النسبة لفئة ثقافية على أحد المحاور، بينما تمثل قيم النسبة لفئة ثقافية أخرى على المحور الآخر. وباستخدام هذين المحورين، فإنه يتم تمثيل كل

سؤال كنقطة في هذا الحيز ثنائي الأبعاد. في حال كون الصعوبات في الأسئلة متسقة عبر الثقافات، فإنها ستقع على طول خط مستقيم بزاوية انحراف قدرها 45 درجة. وبرغم وجود اتساق في صعوبة الأسئلة عبر الفئات الثقافية، فإن بعض التشتت حول الخط المستقيم شيء متوقع نتيجة للأخطاء في أخذ العينات. في حال كون أحد الأسئلة أكثر صعوبة إلى حد كبير (أو أن قلة من الأشخاص يتفقون مع العبارة، في حال كون السؤال مأخوذاً من اختبار نفسي) في ثقافة ما منه في الأخرى، فإنه سيقع بعيداً عن هذا الخط المستقيم، وستتم دراسة هذا السؤال ومثيلاته بصورة إضافية لمعرفة الانحياز المحتمل.

أحد الانتقادات التي توجه إلى طريقة الرسم البياني للتشتت لقيم النسبة بغية تقدير الوظائف التفاضلية للأسئلة (DIF) هو انعدام التحكم بالتأثير. ولأن قيم النسبة هي عبارة عن قيم تابعة للفئة، فمن الصعوبة بمكان إجراء مقارنات واضحة الدلالة لقيم النسبة عبر الفئات. فعلى سبيل المثال، فإن قيم النسبة التي نحصل عليها من فئة ذات أشخاص ذوي كفاءة عالية قد تختلف عن قيم النسبة التي نحصل عليها من فئة ذات أشخاص أقل كفاءة. ويمكن لهذا الاختلاف أن لا تكون له أي علاقة بالانحياز. في حالة كهذه، فإن الاختلافات في قيم النسبة قد لا تكون دالة بالضرورة على عدم تكافؤ الأسئلة عبر الثقافات. وفي الغالب، فإنها ستكون نتيجة للاختلاف في الكفاءة بين الفئات، أو نتيجة للتفاعل بين عدم تكافؤ الأسئلة واختلافات الفئات في الكفاءة. ولمعالجة هذه المشكلة، فقد اقترح آنفوف (1972 و1973) الرسم البياني لـ " قيم الدلتا للأسئلة" لكل فئة عوضاً عن قيم النسبة للأسئلة.



4.3 قيم النسبة للأسئلة

بما أن قيم النسبة للأسئلة عبارة عن قياسات ترتيبية، فمن المألوف اعتبار أن قيم النسبة للأسئلة قد تم الحصول عليها من مجيبين من توزعات مقدرة طبيعية، وإيراد قيم النسبة كانحرافات طبيعية على مقياس ذي متوسط يساوي 13 وانحراف معياري يساوي 4 (يعرف باسم "قيم ETS للدلتا" تيمناً بالمنظمة التي كانت لها الصدارة في استخدامها في مجال تطوير الاختبارات). على هذا النحو، فإن قيمة الدلتا الموافقة لقيمة نسبة 50 . (مثلا) ستكون 13. وإذا كانت قيمة النسبة لسؤال .84 . فإن قيمة الدلتا ستكون 9.0 لقيمة نسبة سؤال .16 فإن قيمة الدلتا ستكون 17.0 . وبكل وضوح، فإن قيمة الدلتا للأسئلة الصعبة تكون قيماً مرتفعة، بينما تكون القيم منخفضة في الأسئلة السهلة.لقد جرت العادة على اعتبار الاختلاف في قيمة الدلتا المساوي 1.5 بين فئتين جديراً بالمراجعة الجدية، بعد أخذ أية اختلاف فئوي كلي بعين الاعتبار (هولاند وواينر 1993). عند الرسم البياني التشتتي لقيم الدلتا،

فإن الجزء المحصور من الخط (المساوي طوليا) المار عبر الرسم البياني يمثل الاختلاف الكلي في المقدرة بين الفئات. وتمثل النقاط الواقعة على الرسم البياني ضمن دوائر صغيرة الأسئلة التي لها صعوبة نسبية متساوية تقريباً في كلا الثقافتين. يظهر الرسم التوضيحي 4.3 مثالاً للرسم البياني للدلتا، حيث يبين قيم الدلتا المحسوبة بواسطة مجيبين أجروا اختبار النسخة الفرنسية (المحور العمودي) أو الإنكليزية (المحور الأفقي) لشهادة دولية (من ميونز، هامبلتون، وكروسينغ 2001). وتعبر حقيقة عدم مرور الخط المساوي طولياً من الأصل (نقطة تقاطع محاور الإحداثيات) عن الاختلاف الكلي في الكفاءة بين الفئتين، وهو 77. (لمصلحة الممتحنين باللغة الإنكليزية والذين كان أداؤهم أفضل قليلاً) في هذه الحالة، لقد قمنا برسم نطاق ثقة حول هذا الخط المساوي طولياً وتتم الإشارة إلى الأسئلة الواقعة خارج هذا النطاق على أنها أسئلة ذات وظائف تفاضلية. إن تلك الأسئلة المشار إلى احتوائها وظائف تفاضلية باستخدام إجراءات إحصائية أكثر تعقيداً موضحة أيضاً في الرسم (لمزيد من التفاصيل، انظر ميونز وآخرون 2001) في هذه الحالة، فإن عملية الرسم البياني قد أشارت إلى الأسئلة ذاتها.

تعد طريقة الرسم البياني للدلتا لتقدير الوظائف التفاضلية للأسئلة عبر الثقافات سهلة التطبيق نسبياً ونتائجها سهلة التفسير. بالرغم من ذلك، فقد أظهرنا أن الرسوم البيانية للدلتا تهمل الأسئلة محتملة الانحياز عندما تختلف في قدراتها التمييزية (دورانس وهولاند 1993). لهذه الأسباب مجتمعة، فإن طريقة الرسم البياني للدلتا تقترح كاختبار أولي فقط، أو في تلك الحالات التي تمنع فيها أحجام العينات من إجراء تحليلات إحصائية أكثر تعقيداً. لقد برهن ميونز وآخرون (2001) أن طريقة الرسم البياني للدلتا كانت فعالة في تحديد الأسئلة التي كانت تؤدي وظائف مختلفة جداً عن بعضها عبر الفئات اللغوية، حتى عندما كانت أحجام العينات صغيرة إلى حد 50 شخصاً لكل فئة. لقد تم استخدام الرسوم البيانية للدلتا بصورة فعالية أيضاً مع أحجام عينات كبيرة (آنغوف ومودو 1973، كوك 1996).

دليل التقييس:

لقد تم اقتراح دليل التقييس للكشف عن الوظائف التفاضلية للأسئلة من قبل دورانس وكوليك (1986). ويمكن أن تعرف هذه الطريقة باسم طريقة "قيمة النسبة المشروطة"، حيث تحصى قيم النسبة المستقلة لكل سؤال متوقفة عليه درجة الاختبار الكلية. فعلى سبيل المثال، يمكن مقارنة ممتحنين أجابوا على نصوص لغوية مختلفة من سؤال بالنسبة لدرجة الاختبار الكلية. الفكرة المراد الإشارة إليها هنا هي إمكانية وجود بعض الأسئلة المثيرة للجدل، ولكن إجمالاً، فإن مطابقة ممتحنين من الفئتين اللغويتين طريقة معقولة للعثور على فئات متكافئة من الممتحنين. بعد ذلك، وبالنسبة لممتحنين ذوي درجة اختبار معطاة، يتم حساب نسبة الممتحنين الذين أجابوا على السؤال بصورة صحيحة لكل فئة ومقارنتها. في حال خلو السؤال من أية مشكلات، فإنه يجب على الفئتين ذوات الأداء الكلي المتكافئ أو القريب من المتكافئ أن يكون أداؤهما متساو تقريباً في الإجابة عليه. وتعاد هذه العملية بالنسبة لجميع المستويات الأخرى من درجات الاختبار. من الناحية العملية، تحسب فواصل درجات الاختبار بصورة نموذجية لمطابقة الممتحنين بحيث لا تكون أحجام العينات لكل فاصل درجة اختبار صغيرة جداً (أي: مطابقة كثيفة). ولجعل مهمة ترميز الأسئلة ذات الوظائف التفاضلية أكثر سهولة، فقد اقترح دورانس وكوليك دليل التقييس، والذي يمثل المتوسط عبر الدرجات أو الفواصل الزمنية على مقياس درجات الاختبار لقيم النسبة المشروطة للفئتين. بالنسبة للعينات الصغيرة، يتم أحياناً اختيار خمسة أو ستة فواصل زمنية بين درجات الاختبار. يحسب هذا الدليل (STD-P) بالعلاقة

$$STD-P=\frac{\sum_{m} w_m (E_{fm} - E_{rm})}{\sum_{m} w_m}$$

حيث تعبر Wm عن التكرار النسبي للفئة الهدف عند مستوى الدرجة) m أو نسبة الفئة المرجع والفئة الهدف عند مستوى الدرجة، الخيار للباحث). وتكونErm

و Efm نسبة الممتحنين عند مستوى الدرجة m والذين أجابوا عن السؤال بصورة صحيحة في الفئة المرجع و الفئة الهدف على التوالي. يمكن للفئة المرجع أن تمثل الممتحنين الذين أجابوا على النص الأصلي من سؤال معين، بينما يمكن للفئة الهدف أن تمثل الممتحنين الذين أجابوا على النص المكيف منه. أحياناً أيضاً يمكن اختيار الأوزان لتتوافق مع الفئة الهدف، وأحياناً أخرى، يمكن اختيارها لتمثل نسبة الفئتين المرجع والهدف معاً عند مستوى درجة معينة، ويعتمد اختيار الأوزان على اهتمام الباحث الأساسي.

يتراوح دليل التقييس بين -1.0 و1.0 على الرغم من عدم توفر أي اختبار إحصائي مرتبط بالمفردة الإحصائية، فإنه يمكن حساب حجم التأثير. فعلى سبيل المثال، يشير دليل انحراف معياري بقيمة 0.10 وسطياً إلى أن الممتحنين في الفئة المرجع الذين تتم مطابقتهم مع ممتحنين في الفئة الهدف يتجاوزون أداء الفئة الهدف عند كل مسافة منتظمة للدرجة بقيمة 0.10 على مقياس تصحيح النسبة.

لقد تم استخدام قيمة دليل تقييس تساوي 0.10 كمعيار لترميز الأسئلة لاحتوائها على وظائف تفاضلية (مثلاً: سيرسي، فيتزجيرالد، كروسينغ 1998). باستخدام هذا المعيار، فإنه إذا تم ترميز 10 أسئلة في اختبار معين لاحتوائها وظائفاً تفاضلية، وكانت كلها تصب في مصلحة واحدة من الفئتين فقط، فإن المستوى الإجمالي للوظائف التفاضلية للأسئلة في الاختبار سيكون حوالي 1 نقطة على مقياس درجة الاختبار الأولية الكلية لمصلحة الفئة المرجع. وباستخدام بيانات حقيقية أو زائفة، فقد خلص ميونز وآخرون (2001) إلى أن دليل التقييس كان فعالاً في ترميز النصوص المكيفة من الأسئلة لاحتوائها على وظائف تفاضلية عندما تكون أحجام العينات صغيرة.

طريقة مانتل - هاينزل

تعد طريقة مانتل-هاينزل (MH) لتحديد الوظائف التفاضلية للأسئلة شبيهة بدليل التقييس في أن ممتحنين من فئتين مختلفتين يطابقون بالنسبة للكفاءة المراد

قياسها وأن احتمالية النجاح في السؤال تتم مقارنتها عبر الفئات. وتعد طريقة مانتل- وهاينزل (MH) توسعاً في اختبار كاي تربيع للاستقلال (مانتل وهاينزل 1959) إلى الوضع الذي يكون فيه ثلاثة مستويات للتطبق. في محيط الوظائف التفاضلية للأسئلة، تكون هذه المستويات هي: فئة الممتحنين (مثلاً: فئتين لغويتين/ثقافيتين)، فاصل متغير المطابقة (الدرجات التي تتم بالاعتماد عليها مطابقة ممتحنين في فئات مختلفة)، والإجابة عن السؤال (صحيحة أو غير صحيحة). ولكل مستوى من متغيرات المطابقة (بصورة نموذجية، درجة الاختبار الكلية)، يتم تنظيم جدول بأبعاد اثنين في اثنين يصنف فئات الممتحنين بحسب الأداء في الأسئلة. من إحدى المميزات الملفتة للنظر في طريقة مانتل- هاينزل توفر اختباراً إحصائياً للوظائف التفاضلية للأسئلة. بالإضافة إلى توفير اختبار للأهمية الإحصائية، فإنه يمكن أيضاً حساب حجم التأثير وتوجد قياسات تقريبية لتصنيف أحجام التأثير هذه إلى وظائف تفاضلية صغيرة، متوسطة، وكبيرة للأسئلة (دورانس وهولاند 1993). يمكن العثور على تفاصيل حول حساب وتفسير إحصائيات مانتل- هاينزل في هولاند وباير (1988) أو دورانس وهولاند (1993).

تعد طرائق الرسم البياني للدلتا، التقييس، ومانتل-هاينزل شائعة؛ لأنها تحتاج أحجام عينات بسيطة فقط ولا تتطلب برامجيات إحصائية متخصصة لإجراء التحليل. إضافة إلى ذلك، فقد أظهرنا أ ن طريقة مانتل-هاينزل فعالة على الأخص في الكشف عن الوظائف التفاضلية للأسئلة. لهذا السبب، غالباً ما يتم استخدامها كمعيار المقارنة في الدراسات التي تقارن طرائق الكشف عن الوظائف التفاضلية للأسئلة. أحد عيوب هذه الطرائق أنها غير فعالة في تحديد الوظائف التفاضلية "غير المنتظمة" للأسئلة. تصور الوظائف التفاضلية غير المنتظمة للأسئلة الوضع الذي تتغير فيه احتمالية النجاح في سؤال معين عبر الفئات عند نقاط مختلفة على طول مسلسلة الكفاءة. إن الطرائق المبنية على نظرية الإجابة عن سؤال والتراجع النسبي لا تحتوي مواطن الضعف هذه. عيب ثان في هذه الطرائق هو أنها وغيرها

من الطرائق المطروحة في هذا الفصل مقتصرة على البيانات الثنائية. ولحسن الحظ فإن معظم الطرائق في الوقت الحاضر قد تم تعميمها لتعالج بيانات إجابة متعددة الفروع، ولكن تلك الطرائق لن تتم مناقشتها هنا (أنظر مثلاً بينفيلد ولام 2000).

طرق نظرية الإجابة عن سؤال

هناك طرق عديدة للكشف عن الوظائف التفاضلية للأسئلة (DIF) ذات البيانات الثنائية تعتمد «نظرية الإجابة عن سؤال» (انظر هامبلتون، سواميناثان، وروجرز 1991). وبصورة أساسية، تقدر جميع هذه الطرائق إمكانية استخدام مجموعة مشتركة من المقادير متغيرة القيمة لسؤال لشرح وظيفة سؤال معين في كل فئة لغوية/ ثقافية. وفي حال الحاجة إلى مقادير متغيرة مختلفة لشرح وظيفة السؤال في كل فئة، عندها يتم ترميز السؤال لاحتوائه على وظيفة تفاضلية. إحدى طرائق نظرية الإجابة على سؤال للكشف عن الوظائف التفاضلية للأسئلة هي طريقة كاي تربيع التي استخدمها لورد، والتي تختبر المقادير المتغيرة للقدرات التمييزية للأسئلة ومقادير صعوبتها عبر الفئات (لورد 1980). لقد استخدم آنغوف وكوك (1988) هذه الطريقة لتحديد الأسئلة المشتركة المستخدمة في الموازنة بين اختبار أهلية التعليم (SAT) والنسخة الإسبانية منه.

طريقة أخرى مبنية على نظرية الإجابة عن سؤال لكشف الوظائف التفاضلية للأسئلة هي اختبار رادجو للمنطقة بين منحنيين مميزين لسؤال (رادجو 1988،1990). في هذا التحليل، يحسب المنحنى المميز (ICC) لسؤال معين بصورة مستقلة لكل فئة. بعد ذلك، يتم اختبار المنطقة بين المنحنيين المميزين (ICCS) لمعرفة الأهمية الإحصائية. بالنسبة للبيانات المسجلة بصورة ثنائية التفرع، فقد أدخلت هذه الطريقة تحسينات على طريقة كاي تربيع التي استخدمها لورد في أن الاختلافات في الأداء في الأسئلة نتيجة للحدس (أي: المقدار المتغير C) يمكن

تقديرها أيضاً. على الرغم من كون هذه الطريقة قد تم استخدامها غالباً مع فقرات مسجلة بصورة ثنائية التفرع، فإنه يمكن توسيعها إلى الحالة متعددة الفروع.

قام بادجل وآخرون (1995) بمقارنة نتائج الكشف عن الوظائف التفاضلية للأسئلة لطريقة كاي تربيع التي استخدمها لورد، مناطق رادجو المعلمة وغير المعلمة، وإجراءات مانتل-هاينزل عبر الاختبارات العددية والمنطقية التي تم تطويرها باستخدام اللغة الإنكليزية أولاً ثم تكييفها إلى اللغة الفرنسية بعد ذلك. لقد وجدوا درجة كبيرة من الاتساق عبر هذه الطرائق في تحديد الأسئلة ذات الوظائف التفاضلية المهمة.

طريقة شائعة ثالثة مبنية على نظرية الإجابة عن سؤال للكشف عن الوظائف التفاضلية للأسئلة هي طريقة نسبة الاحتمالات (ثيسن، شتاينبيرغ، ووايز 1988، 1993). باستخدام هذه الطريقة، تتم ملاءمة نموذجين مبنيين على نظرية الإجابة على سؤال (IRT) مع بيانات إجابات الممتحنين ويتم تقدير الاختلاف بين ملاءمة هذين النموذجين للبيانات لبيان الأهمية الإحصائية. يكون النموذج الأول الذي تمت ملاءمته مع البيانات نموذجاً لا يحتوي أية وظائف تفاضلية للأسئلة "NO-DIF" حيث يتم استخدام ذات المقادير المتغيرة للسؤال في معايرته في كل فئة أو مجموعة. ويكون النموذج الثاني الذي تمت ملاءمته مع البيانات نموذجاً يحتوي وظائف تفاضلية للأسئلة (DIF)، حيث يتم استخدام مقادير متغيرة مستقلة لمعايرة السؤال في كل فئة. أي أن النموذج الذي لا يحتوي أية وظائف تفاضلية للأسئلة (NO-DIF) يعامل الفقرة على أنها متكافئة عبر الفئات، بينما يعامل النموذج الذي يحتوي وظائف تفاضلية للأسئلة (DIF) السؤال على أنه مستقل في كل فئة. وبصورة واضحة، يكون النموذج المحتوي وظائف تفاضلية للأسئلة أقل محدودية لأنه يدخل في اعتباراته مقادير متغيرة أكثر لملاءمتها مع البيانات. ولتحديد ما إذا كانت هذه المقادير المتغيرة الإضافية ضرورية (أي: هل نحتاج مقادير متغيرة مستقلة لمعايرة

السؤال في كل فئة؟)، فإنه تتم المقارنة بين الاحتمالات المرتبطة بكل نموذج (أي: احتمال الحصول على البيانات في حال كون النموذج صحيحاً). ويتم توزيع الاختلاف بين احتمالات كل نموذج ككاي تربيع (في الحقيقة، إن سجل الاحتمالات هو الذي يوزع ككاي تربيع، ويستخدم عملياً قيمة أقل بمرتين من سجل الاحتمالات لمقارنة الملاءمة عبر النماذج)، وتكون درجات الحرية المرتبطة باختبار كاي تربيع هذا بكل بساطة عبارة عن الاختلاف في عدد المقادير المتغيرة التي تم أخذها بعين الاعتبار في كل نموذج.

لقد تم استخدام اختبار نسبة الاحتمالات المعتمد على نظرية الإجابة عن سؤال بصورة واسعة لتقصي الوظائف التفاضلية للأسئلة (DIF) عبر فئات فرعية أجرت الاختبار بلغة واحدة باستخدام نماذج ثنائية ومتعددة الفروع معاً لنظرية الإجابة عن سؤال (ثيسن، شتاينبيرغ، وجيرارد 1986، ثيسن وآخرون 1988، واينر 1995، واينر، سيرسي، وثيسن1991). في الوقت الحالي، توجد بعض التطبيقات لهذه التقنية في معالجة مشكلة الكشف عن العيوب في تكييفات الأسئلة (سيرسي وبيربيروغلو 2000). من محاسن هذه الطريقة قوتها الإحصائية، مرونتها في معالجة بيانات ثنائية ومتعددة الفروع معاً، وقدرتها على تقدير الأسئلة في أكثر من فئتين في وقت واحد. بالرغم من ذلك، فإن لهذه الطريقة عيباً مهماً يتلخص في أن استخدامها يستغرق وقتاً طويلاً جداً. وتجب لكل فقرة ملاءمة نماذج متعددة معتمدة على «نظرية الإجابة عن سؤال» (IRT) مع البيانات. وعندما يتألف التقييم من عدد كبير من الأسئلة ويتم استبعاد النموذج الذي لا يحتوي أية وظائف تفاضلية للأسئلة (NO-DIF)، فإن عزل الأسئلة ذات الوظائف التفاضلية (DIF) المحددة يصبح عملية شاقة (سيرسي، وبيربيروغلو 2000).

خلاصة:

تتنوع الطرائق الإحصائية في استقصاء الأسئلة المثيرة للجدل نتيجة لعيب في التكييف اللغوي/ الثقافي بين طرائق بسيطة تعتمد التحليل النظري، وطرائق معقدة

تعتمد نظرية القياس الحديثة. يعتمد اختيار طريقة بعينها على عدة عوامل من ضمنها أحجام العينات، عدد الأسئلة الداخلة في التقييم، تحديد درجات الأسئلة، وتوفر برامجيات إحصائية. وفي تلك الحالات المتضمنة أحجام عينات صغيرة نسبياً و أسئلة مسجلة بصورة ثنائية التفرع، فإن طريقتي الرسم البياني للدلتا والتقييس (Standardization) هي الطرائق الموصى بها. وبازدياد أحجام العينات، إن طريقتي مانتل-هاينزل (MH) ونظرية الإجابة عن سؤال (IRT) يمكن أن تكونا جديرتين بالتفضيل. وفي كل الحالات يجب التذكير بأنه قبل إجراء أية تحريات حول العيوب على مستوى الأسئلة فإن الانحيازات في المفاهيم والطرائق يجب أن يتم استبعادها. إن جميع طرائق الكشف عن الوظائف التفاضلية للأسئلة تفترض متغير التطابق المستخدم لمطابقة المتحنين، ولتكن درجة الاختبار الكلية مثلاً، درجة متغير كامنة (أي: درجة معتمدة على نظرية الإجابة على سؤال)، أو متغيراً خارجاً عن نطاق التقييم، وأنه صالح لغرض المطابقة. إن أي انحياز نظامي في هذا المتغير لن يتم الكشف عنه عند مستويات الأسئلة وسيضعف صحة نتائج الوظائف التفاضلية للأسئلة (DIF) ولتلافي هذا الوضع، ننصح باتباع إجراءات تكييف حذرة (مثلاً: إرشادات تكييف الاختبارات التي تصدرها لجنة الاختبارات الدولية، أنظر الفصل الأول من هذا الكتاب، أو هامبلتون وباتسولا 1999) وإجراء اختبارات إحصائية على الانحيازات في المفاهيم والطرائق.

استنتاجات:

يعتبر تقييم ومقارنة الأفراد الذين يعملون في إطار لغات وثقافات مختلفة تحدياً كبيراً. ولقد أظهرت المقالات النقدية في هذا الفرع العلمي العديد من الأخطار التي تهدد الصدق الداخلي للدراسات عبر الثقافات كالانحيازات في المفاهيم والطرائق والأسئلة. لقد قمنا في هذا الفصل بتلخيص وشرح الطرائق الإحصائية التي يمكن للباحثين استخدامها لتقدير أثر هذه الأخطار على صدق

وسائلهم في التقييم عبر الثقافات. ضمن نطاق خبرتنا العلمية، فإن نتائج تحليلات كهذه يمكن أن تستخدم في إدخال تعديلات على التطور اللاحق للوسائل والذي ستتمخض عنه تقييمات أكثر صحة، ومقارنات أكثر شرعية عبر الأفراد المختلفين في اللغة والثقافة.

*****

المراجع

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Detecting the causes of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36*(3), 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. (1972). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 96–116). Baltimore: Johns Hopkins University Press.

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

Angoff, W. H., & Modu, C. C. (1973). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Rep. No. 3). New York: College Entrance Examination Board.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13,* 12–21.

Brown, R., & Marcoulides, G. A. (1996). A cross-cultural comparison of the Brown Locus of Control Scale. *Educational and Psychological Measurement, 56,* 858–863.

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

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal, 56*(8), 472–475.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1,* 55–86.

Byrne, B. (2003). Confirmatory factor analysis. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (Vol. 1). Thousand Oaks, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart–Young decomposition. *Psychometrika, 35,* 283–319.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Cook, L. L. (1996, August). *Establishing score comparability for tests given in different languages.* Paper presented at the meeting of the American Psychological Association, Toronto, Canada.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests *Psychological Bulletin, 52,* 281–302.

Dorans, N. J., & Holland. P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23,* 355–368.

Ellis, B. B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment, 11,* 184–193.

Foorman, B., Yoshida, H., Swank, P., & Garson, J. (1989). Japanese and American children's styles of processing figural matrices. *Journal of Cross-Cultural Psychology, 20,* 263–295.

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., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

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

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

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16,* 131–152.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 19–48.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5–34.

McDonald, R. P. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Millsap, R. J., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in test translation. *International Journal of Testing, 1,* 115–135.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5–15.

Poortinga, Y. H. (1991). Conceptual implications of item bias. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.), *Advances in computer-based human assessment* (pp. 279–290). Dordrecht, Netherlands: Kluwer Academic.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23–37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* U552–U566.

Robie, C., & Ryan, A. M. (1996). Structural equivalence of a measure of cross-cultural adjustment. *Educational and Psychological Measurement, 56,* 514–521.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing, 3*(1), 1–20.

Shealy, R. & Stout, W. (1993). A model-based standardization differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–167.

Sireci, S.G., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests.* Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Sireci, S. G., & Berberoglu, G. (2000). Evaluating translation DIF using bilinguals. *Applied Measurement in Education, 13*(3), 229–248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses* (Laboratory of Psychometric and Evaluative Research Rep. No. 329). Amherst: University of Massachusetts, School of Education.

Sireci, S. G., Foster, D., Olsen, J. B., & Robin, F. (1997, March). *Comparing dual-language versions of international computerized certification exams*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

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). Hillsdale, 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. 67–114). Hillsdale, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J., Daal, M., & van Zonneveld, R. (1986). The trainability of abstract reasoning: A cross-cultural comparison. *International Journal of Psychology, 21,* 589–615.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F. J., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht, Netherlands: Kluwer Academic.

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

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–186.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.