6

إرساء قواعد مقارنة الدرجات لاختبارات معطاة بلغات مختلفة

ليندا ل. كوك
خدمة الاختبارات التربوية

أليسا ب. شميت - كاسكالار
مجموعة التقييم العالمية، بروكسل، بلجيكا

إن الاختبارات المكيفة في الإدارة لمجموعات لغوية مختلفة وإعطاء الاختبارات المكيّفة للمتقدمين للاختبار من ثقافات مختلفة هو تدريب يملك تاريخاً طويلاً في مجال التقييم النفسي. يشير عمل "نيرمان" (1916) إلى كم مضى من الزمن على معرفة الباحثين بالمشكلات المتصلة باستخدام الأدوات التي كانت قد طوّرت من أجل سكان بلد ما لتقييم صفات سكان بلد ثان والذي ربما يختلف عن الأول في الخلفية والثقافة.

إن استعمال اختبارات مكيّفة تم تطويرها لسكان بلد معين، ومن ثم إعطاء هذه الاختبارات إلى سكان بلد ثان، والذي قد يختلف عن الأول في كل من اللغة والثقافة، هو تدريب ازداد بدرجة كبيرة على مدى العشر سنوات الماضية. وضع هامبلتون (1993) وسيرغي (1997) في قائمة عدداً من العوامل المساهمة في

الاهتمام المتزايد في تكييفات الاختبار. ومن بين تلك العوامل: تعزيز العدالة في مقارنات الأفراد والمجموعات من خلفيات ثقافية ولغوية مختلفة؛ تسهيل الدراسات المقارنة عبر المجموعات الثقافية والعرقية والوطنية؛ تسهيل مقارنة إنجاز الطلاب في بلدان مختلفة. إضافة إلى هذه القائمة تأتي العولمة المتزايدة لأعمال تجارية عديدة، مؤدية الحاجة لتطوير وتكييف اختبارات في اللغة الأصلية لموظفين من أجل استخدامها في المصادقة عليها رسمياً. إن كلاً من العوامل مستقل في القدرة على مقارنة الدرجات النهائية المحصّلة من الاختبارات المقررة على مجموعات تختلف في كل من اللغة والثقافة.

بغض النظر عن الأسباب لأجل تكييف اختبار جرى إعطاؤه في إحدى اللغات إلى الإدارة في لغة ثانية أو في لغات متعددة، فالموضوعات التي تحيط بعلم المنهج الملائم من أجل تكييف الاختبار لدعم مقارنات ذات مصداقية للدرجات النهائية تكون معقدة إلى حد كبير.

أشار بورتينغا (1989) إلى أنه ربما تكون مقارنات قدرات الأفراد والجماعات مضللة لسببين: يتصل السبب الأول بالصفة المقاسة، وقد أعطى على سبيل المثال عدم جدوى مقارنة طول شخص ما بوزن فرد ثان. ويتصل السبب الثاني بوحدات القياس المستخدمة في المقارنة؛ على سبيل المثال، لا يستطيع المرء أن يقوم بمقارنة مباشرة لطول شيئين إذا جرى قياس شيء واحد بالإنش والآخر بالسنتيمتر. تبدو تلك كنقاط واضحة عندما يعود المرء إلى الصفات الفيزيولوجية مثل الارتفاع، والوزن، والطول. على أية حال، يصبح الوضع للتو أكثر تعقيداً عندما تمتد المقارنات إلى درجات نهاية محصّلة من تقييمات تربوية ونفسية.

لنأخذ بعين الاعتبار، على سبيل المثال، اختباراً للجبر يحوي شيئاً من مشكلات الكلمة. لننظر أبعد بأن الاختبار قد تم تصميمه باللغة الإنكليزية وجرى وضع درجاته النهائية باستخدام معطيات من سكان ناطقين بالإنكليزية. ثم يترجم

الاختبار إلى الإسبانية ويقرر على مجموعة طلاب ناطقين بالإسبانية. فإذا لم يأخذ الطلاب الناطقون بالإسبانية درجات نهائية جيدة في الاختبار مثل تلك التي أخذها الطلاب الناطقون بالإنكليزية، كيف لنا أن نعرف فيما إذا كانت الفروق في الدرجات النهائية بسبب اختلاف المجموعات في مقدرتهم في الجبر، أو أن يعزى ذلك إلى حقيقة أن الترجمة لمشكلات اللفظ الجبرية إلى اللغة الإسبانية جعلت جوهرياً المشكلات أكثر صعوبة على المتقدمين للاختبار الناطقين بالإسبانية؟

إمكانية أخرى وهو أن الاختبار المقرر في اللغة الإسبانية يتطلب وقت قراءة أكثر منه في الاختبار المقرر في اللغة الإنكليزية، وهكذا يجعله أكثر توفيقاً للسكان الناطقين بالإسبانية. هل ينبغي أن يكون النجاح عاملاً في تقييم المقدرة بالجبر للمجموعة الناطقة بالإسبانية وليس للمجموعة الناطقة بالإنكليزية.

إضافة إلى ذلك، ربما من المحتمل ألا تكون التعليمات للاختبار مترجمة بوضوح ويكون المتقدمون للاختبار الناطقون بالإسبانية مشوشين نظراً للخطط الاستراتيجية لمفتاح أخذ الاختبار، مثلاً فيما إذا كانوا سيعاقبون أم لا من أجل إجابات تخمينية على الأسئلة.

إن قائمة الأسباب للفروق بين الدرجات النهائية المحصّلة في اختبار للجبر معطى لتوه لمجموعات ناطقة بالإنكليزية ومجموعات ناطقة بالإسبانية هي بالتأكيد القاطع ليست كاملة؛ المقصود بها فقط هو كم يكون صعباً تجنب بناء مصادر تنوع في الدرجات النهائية للاختبار عند مقارنة الدرجات النهائية في الاختبارات المكيّفة.

لقد قام عدد كبير من الباحثين بوصف الإجراءات التي تناولت قضايا التباين في درجات الاختبار غير المتصل بالموضوع وبالتالي ترويج مستوى متزايد لقابلية مقارنة الدرجات في الاختبارات المكيفة.

(انظر كيسنجر، 1994، وسيرغي 1997، وهامبلتون 1993، من أجل نقاشات عميقة لهذه الإجراءات). تشمل الإجراءات الترجمة، الترجمة الارتجاعية للأداة التي ستكيّف، اختبار الدليل وغربلة مفردات الاختبار لتوظيف مفردة مميزة، اختبار ميداني وموزون، تطوير إجراءات الإدارة، وبحث علمي ذو مصداقية.

هذه النقطة الأخيرة مهمة إلى حد كبير لأنه، على الرغم من درجة الانتباه القصوى المبذولة للمسائل المنهجية، قد لا يكون من الممكن الحصول ببساطة على بناء مواز لاختبار تم إعطاؤه إلى سكان بلدان متعددة تختلف في اللغة والثقافة. بالتالي، من المهم للبحث العلمي ذي المصداقية أن يؤخذ به في أي اختبار مكيّف لتأكيد أن المقارنات والتفسيرات الصادقة للدرجات النهائية يجري تدعيمها بالدرجات النهائية للاختبار.

يجري التركيز في هذا الفصل على عامل واحد فقط مؤثر على قابلية مقارنة الدرجات النهائية المحصلة في الاختبارات المكيفة. السبب الثاني لتضليل المقارنات وفقاً لـ "بورتينغا" (1989) هو وحدات قياس غير متكافئة. في الأقسام التالية من هذا الفصل، نزود قاعدة لفهم الطرائق الإحصائية المتوفرة حالياً من أجل تعادل وموازنة الاختبارات النفسية والتربوية، نصف وننقد إجراءات ربط المقاييس المحددة التي تستخدم في دراسات تكييف الاختبار، ونوضح إجراءات ربط منتقاة وقضايا بوصف ونقد دراسات ثلاث جرى القيام بها عبر العشرين سنة الماضية وذلك لربط الدرجات النهائية في اختبار التقييم للمدارس الثانوية (SAT) Scholastic Assess-ment Test) بالدرجات النهائية في الـ (PAA) (Prueba de Aptitud Academica).

نظرة شاملة على طرق الربط:

ناقش لين (1993) حقيقة أن العديد من التقنيات المختلفة متوفر من أجل ربط نتائج الاختبار وأن المصطلح المستخدم لوصف التقنيات لم يستخدم دائماً بصورة واضحة. استمر لين في وصف خمس طرق مختلفة لربط نتائج الاختبار وكيف يؤثر

نمط الربط على المقارنات والتفسيرات. بيّن القصد من أن التداخلات التي تدّعي قابلية تبادل الدرجات النهائية تتطلب طرائق قوية لربط مقاييس الاختبارات. ربما تكتفي أنماط أخرى للتداخلات بأشكال للربط أضعف، لكن أشكال الربط الأضعف تلك هي بطبيعتها تابعة للسّياق، المجموعة والزمن.

في مقالته 1993، وصف لين خمس طرائق لربط الاختبارات النفسية والتربوية. هنا يتم فقط وصف أربع منها. هذه الطرائق الأربع هي: التساوي، المعايرة، التعديل الإحصائي، والتنبؤ.

التساوي (Equating)

احتفظ لين (1993) بمصطلح "متساوي" من أجل الروابط التي تعطي درجات نهائية يمكن استخدامها بالتبادل. حدد الغاية بأن أقوى صورة لمقياس ربط الدرجة النهائية هو التساوي. يرجع لين إلى لورد (1980) وتعريفه "للتساوي" الذي يتطلب أن يستخدم مصطلح "درجات نهائية متساوية" فقط عندما يكون اختيار أي ترجمة أو شكل للاختبار ينبغي أخذه مسألة لا تقدم ولا تؤخر بالنسبة للمتقدم للاختبار ومستخدم درجاته النهائية. من الواضح أنه إذا كانت المقارنات للدرجات النهائية للاختبار تتطلب اختبارات ينبغي اعتبارها قابلة للتبادل (مثل: درجات نهائية على ورق اختبار لـ SAT مقررة في تشرين الأول ودرجات نهائية على ورق اختبار لـ SAT في حزيران) عندئذ يجب أن تستخدم إجراءات التساوي. أشار لين إلى أن المتطلبات لأجل التساوي هي أن أشكال الاختبار يجب أن تقيس التركيب نفسه بدرجات متساوية من المصداقية، وهذا يعني أن الأشكال يجب أن تكون قابلة للتبادل. أشار آخرون (فان دي فيفر وبورتينغا 1991) إلى أنه ليس فقط يجب أن تكون الأشكال قابلة للتبادل من أجل إجراءات دراسة متساوية مناسبة بل يجب أيضاً أن تكون الشروط الفيزيائية لإدارة الاختبار قابلة للمقارنة.

المعايرة (Calibration)

وصف لين (1993) "معايرة" كوسيلة لمقارنة الدرجات النهائية على أوراق اختبار ترضي المتطلبات الأقل تشدداً من المتطلبات لأجل اختبارات متساوية، أعطى التالي

كأمثلة على "المعايرة": اختبارات الربط التي تختلف بدرجة معتبرة في الطول، وبالتالي في المصداقية؛ اختبارات الربط التي يمكن أن تستخدم لمقارنة طلاب في مستويات متطورة مختلفة (جرت الإشارة إليها في الأدب كدراسات متدرجة عمودية).

وضع لين (1993) في قائمة المتطلبات من أجل المعايرة (مثل: يجب أن تقيس "الاختبارات" التركيب نفسه. لكن يمكن أن تختلف في المصداقية. يمكن أن تختلف أيضاً في المستوى الذي تكون فيه المقاييس أكثر فائدة) (صفحة 90). أشار لين إلى أن المعايرة تعطي وسيلة لمقارنة الدرجات النهائية على أوراق الاختبار التي ترضي متطلبات أقل تشدداً من تلك التي على أوراق اختبار متوازية. على أي حال، يوجد ثمن ينبغي دفعه. يوجد هنالك الإمكانية بأن يكون بالاستطاعة إجراء نماذج مختلفة متعددة من دراسات المعايرة وستعطي كل معايرة الإجابة عن سؤال مختلف.

استشهد لين باتصال شخصي من مسليفي وستوكنغ كما أشار إلى أنه عندما لا تكون الاختبارات X و Y غير ذات مصداقية بدرجة متساوية تستطيع المعايرة التي تحول الدرجات النهائية Y إلى المقياس X أن تجيب عن السؤال "ما تكون قيمة X لأجله تكون الدرجة النهائية X للشخص الأكثر ميلاً؟"، أشار المؤلفون إلى أنه من الأرجح أن تعطي المعايرة نفسها إجابات غير صحيحة لأسئلة حول خواص توزيعات الدرجات النهائية للمجموعات المتقدمة لاختبارات X و Y.

التعديل الإحصائي (Statistical Moderation)

يرجع إجراء الربط الثالث الموصوف من قبل لين (1993)، "التعديل الإحصائي"، إلى إجراءات تشمل عادة استخدام الدرجات النهائية في اختبار خارجي لدرجات نهائية وسطية تم الحصول عليها في الاختبارات التي ينبغي أن تقارن. إن التعديل الإحصائي هو وسيلة تقنية يمكن استخدامها لمقارنة الدرجات النهائية في اختبارات التحصيل التي تقيس مناطق مادة مختلفة. على سبيل المثال، ترغب الكليات أحياناً في مقارنة درجات نهائية اكتسبها طلاب خضعوا لاختبارات مختلفة لمادة SAT II.

يجري تطوير نظام متري عام للدرجات النهائية لاختبار مادة SAT II باستخدام الدرجات النهائية للرياضيات والكلمات لـ SATI وذلك كاختبار خارجي في دراسة التعديل الإحصائي. على العكس من طريقتي الربط التي جرت مناقشتهما سابقاً، لم يتطلب التعديل الإحصائي تقييمين سيجري ربطهما لقياس التركيب نفسه. على أية حال، يتطلب هذا الإجراء عند استخدامه في عمل ربط اختبار خارجي عام، ويعتمد نجاح هذا الإجراء بدرجة كبيرة على متانة العلاقة بين الاختبار الخارجي والمقاييس التي ينبغي ربطها.

إن أحد المساوئ الرئيسة لتقنيات التعديل الإحصائي هو أن مكوناته تابعة للسياق والمجموعة، والزمن. بالتالي ربما تتنوع العلاقة التي جرى تأسيسها بين الدرجات النهائية على تقييمين تم تطويرهما باستخدام تقنيات التعديل الإحصائي وذلك وفقاً لمجموعة المتقدمين للاختبار الذي جرى انتقاؤهم للاشتراك في دراسة التعديل.

التنبؤ (Prediction)

إن إجراء الربط الرابع الذي وصفه لين (1993) هو التنبؤ. علق لين أنه طالما يوجد درجة ما للعلاقة بين الإنجاز في أحد التقييمات مع الإنجاز في آخر، يكون ممكناً الربط بين التقييمين من خلال التنبؤ. بالطبع ستعتمد متانة ربط التقييمين على متانة العلاقة بين الدرجات النهائية المحصّلة من التقييمين. إن بعض النقاط الضعيفة في التنبؤ، عند استخدامه كإجراء ربط، تكون في أن تعادلات التنبؤ تعتمد على المجموعة. أيضاً تكون تعادلات التنبؤ وحيدة الاتجاه؛ مما يعني أنه يجب استخدام روابط منفصلة للتنبؤ بالدرجات النهائية لاختبارY من اختبار X والدرجات النهائية في اختبار X من اختبار Y.

تطبيق أربع طرق لربط درجات نهائية على أوراق اختبار معطاة في لغات مختلفة

من المهم أن نأخذ بعين الاعتبار التأسيس للدرجات النهائية القابلة للمقارنة من أجل الاختبارات التي قد جرى تكييفها وفق لغات مختلفة وتقررت على مرشحين في لغاتهم الأصلية من منظور نطاق الربط الذي قدمه لين (1993).

أولاً، من الواضح أنه من المحال تقريباً ربط الاختبارات التي تم تكييفها مع لغات مختلفة، ومن ثم إعطاء تلك الاختبارات إلى المتقدمين للاختبار في لغاتهم الأصلية واعتبار كون الاختبارات المربوطة متعادلة. السبب في هذا هو أن إعطاء المشكلات المقترنة بالاختبارات التي تم تكييفها من أجل لغة مختلفة ومجموعات ثقافية مختلفة (هامبلتون 1993) ليس من المحتمل أن افتراض الأشكال المتوازية (أشكال متشابهة جداً في المحتوى والصفات الإحصائية)، (مطلوبة لإجراء التساوي، أن يتحقق).

ينبغي أيضاً أن يكون واضحاً أنه من غير المحتمل أن يستطيع شخص ما أن يقول إنها مسألة لا فرق فيها بالنسبة للمتقدم للاختبار، فيما إذا أخذ أو أخذت اختباراً باللّغة الأصلية أو بلغة مترجمة (لورد 1980، بيّن هذه النتيجة كمحصلة لاختبارات درجات نهائية متساوية). ربما إذا كان بالإمكان تكييف اختبار بشكل تام، وكان المتقدمون للاختبار ثنائيي اللغة بشكل تام فإن حالة كهذه يمكن أخذها بالاعتبار. على أية حال، لا توجد أي من تلك الحالات بصورة حرفية في التقويمات الثقافية المتداخلة/ اللغوية المتداخلة.

إن المعاني المتضمنة في محاولة ربط الاختبارات المكيّفة مع لغات مختلفة وإعطاؤها إلى طلاب بلغتهم الأصلية ولغات مترجمة هي واضحة تماماً. من المحال لأية دراسة ربط، بغض النظر عن العناية التي بذلت لتنفيذها، أن تعطي درجات نهائية قابلة للمقارنة في مفهوم الدرجات النهائية المتساوية.

سؤال مهم يجب طرحه وهو فيما إذا كان ممكناً أو ليس ممكناً ربط الدرجات النهائية للاختبارات التي جرى استخدامها لتقويمات لغات متداخلة واعتبار هذه الدرجات معيارية باستخدام معيار لين (1993). وفقاً لـ"لين"، لا يتطلب الربط عن طريق المعايرة أن تكون الاختبارات ذات ثبات متساوٍ، ولكن أن تقيس المفهوم نفسه فقط. في سبيل الإجابة عن سؤال فيما إذا كانت المعايرة ممكنة أم ليست ممكنة، يجب أن يؤخذ بعين الاعتبار طبيعة التقييم. على سبيل المثال من المقبول بدرجة أكبر

كثيراً، أن يقيس اختبار رياضيات جرى إعطاؤه باللغة الإنكليزية لمجموعة ناطقة بالإنكليزية وباللغة الإسبانية لمجموعة ناطقة بالإسبانية المفهوم نفسه للمجموعتين من أن يقيس اختبار لمقدرة لفظية أعطى تحت نفس الظروف للمجموعتين نفس المفهوم. إذا لم تكن إحدى إجراءات التساوي. أو المعايرة قابلة للتطبيق بسبب طبيعة الاختبارات كما يتمنى المتدرب أن يقوم بالربط. فإنه يلجأ إلى إجراءات التساوي. يجب أن يجري تصميم دراسات التعديل الإحصائي بعناية شديدة لتحقيق الحاجات الفريدة لدراسة ربط اللغات المتداخلة. إن السبب في وجوب تصميم الدراسات بعناية هو أن اختباراً خارجياً عاماً ينبغي أن يؤخذ من قبل مجموعات في اللغة الأصلية وفي لغة الهدف المترجم إليها. من الصعب أن نرى كم يكون هذا ممكناً إذا كانت المجموعتان أحاديتي اللغة بالتمام. على أية حال، من المحتمل أن يوجد طريقان كأن يكون ممكناً محاكاة اختبار خارجي مشترك وأن يكون ممكناً تقرير النتائج، لدراسة التعديل. يتمثل أحد الطرق في تكييف اختبار مشترك وفي اللغة الأصلية ولغة الهدف مع عناية كافية وبذلك سيعمل كعامل ربط "عام" بين قياسين عند إعطائه إلى مجموعتي لغة خاصتين. الطريق الثاني بأن يعطى الاختبار العام إما باللغة الأصلية أو بلغة الهدف إلى مجموعة ثنائية اللغة. وبالتالي ربما يؤدى الاختبار كاختبار خارجي مشترك. وقد جرى استخدام كلا هذين الإجراءين مع بعض النجاح في دراسات ربط عبر لغوية.

إن الإجراء الرابع الموصوف من قبل لين (1993) هو التنبؤ. دراسات تنبؤ مماثلة تماماً أنتجت درجات نهائية للاختبار X جرى التنبؤ بها من الدرجات النهائية للاختبار Y. بغية تطوير معادلة التنبؤ، يجب أن يأخذ بعض المتقدمين للاختبار كلا الاختبارين. حذّر لين من أن النتائج لدراسات التنبؤ لديها عدد من التحديدات. إضافة إلى ذلك، حذّر من أن تكون العلاقة التي تم استخدامها لتقرير معادلات التنبؤ هي لمجموعة معينة. ولهذا عواقب معينة من أجل تطبيق هذا النمط من إجراء الربط لعمليات التقييم عبر اللغات وبافتراض، أن يكون الشرط الأساسي لتطوير

معادلات التنبؤ هو وجوب أن تأخذ مجموعة واحدة كلا الاختبارين، فإن ما يترتب على ذلك هو أن دراسة ربط لغات مختلفة مرتكزة على التنبؤ هو أن المجموعة المستخدمة من أجل دراسة لربط يجب أن تكون ثنائية اللغة. إن الضرر في استخدام مجموعة ثنائية اللغة لهذا النمط من الدراسة هو أن النتائج لهذه الدراسة قد لا تعمم على الحالة مركز الاهتمام، وهي الحالة التي تعطى فيها الاختبارات بلغات أصلية ولغات مترجم إليها إلى متقدمين للاختبار ناطقين بلغة واحدة من تلك اللغات.

إن إحدى ميزات دراسات التنبؤ كأساس لربط اختبارين مقدمين بلغات مختلفة هي أن الإجراءات تسمح باستخدام متغيرات معدّلة للغة وبالتالي ربما تقدم إجابة أكثر دقة للسؤال عن مدى جودة إنجاز الطالب (الفرد) إذا تم إعطاؤه الاختبار باللغة الهدف (المترجم إليها).

ربط اختبارات معطاة بلغات مختلفة:

مثالياً، يرغب أولئك المهتمون في ربط التقويمات التي كان قد تم تكييفها مع لغات مختلفة وجرى إعطاؤها إلى متقدمين للاختبار ناطقين بلغة واحدة في لغاتهم الخاصة، يتمنون أن يكونوا قادرين على مقارنة مهارات وقدرات المتقدمين للاختبار بأخذهم تقويمات مختلفة كما لو كانت الدرجات النهائية المحصلة في التقويمات قابلة للتبادل كلياً (متعادلة). على أية حال، وكما جرت الإشارة إليه سابقاً في هذا الفصل، يكون هذا الوضع المثالي صعباً (إن لم يكن محالاً) الحصول عليه لأن المعطيات التي جرى جمعها في دراسات الربط عبر اللغات لم يتم تزويدها جيداً بنماذج متوازنة نموذجية.

قدم سيرسي (1997) نظرة شاملة ممتازة للموضوعات التقنية المتعلقة باختبارات ربط تم استخدامها في تقويمات عبر اللغات. بدأت مراجعته بمناقشة حقيقة أن المتدربين يعتقدون أن ترجمة اختبار ببساطة من إحدى اللغات إلى أخرى هو شرط كاف لتقويم لغات مختلفة. أشار سيرسي، إلى المغالطة في سير المحاكمة

هذه بأنه لا شيء لأن الآثار غير المقصودة للترجمة قد تنتج مواد تختلف في الصعوبة وفي صفات أخرى عبر لغات مختلفة (انظر كسينجر 1994، هامبلتون، 1993، 1996، أوليدو 1981، بريتو 1992).

وفقاً لسيرسي (1997)، إن التصاميم المستخدمة لربط التقويمات المعطاة في لغات مختلفة تقع في فئات ثلاث:

(أ) تصاميم المجموعة أحادية اللغة المنفصلة.

(ب) تصاميم المجموعة ثنائية اللغة.

(ج) تصاميم المجموعة أحادية اللغة المطابقة.

تتضمن مجموعة أحادية اللغة المنفصلة بالضرورة إجراءاً ما لتطوير المواد المتطابقة جزئياً، بينما يحوي التصميمان الآخران كمطلبهم المركزي تطوير المقاربات لتتطابق جزئياً مع مجموعات المتقدمين للاختبار.

تصاميم مجموعة أحادية اللغة المنفصلة:

تصل كل هذه التصاميم إدارة الاختبارات في اللغات الأصلية ولغات الهدف (المترجم إليها) بمجموعاتهم اللغوية الخاصة وتربط الاختبارات من خلال منظومة المواد التي يمكن إلى حد ما اعتبارها "عامة" لكلتا مجموعتي اللغة. وقد تم اعتبار تطبيقات نظرية إجابة المادة (IRT) لهذا النمط من التصميم واعدة تماماً. كان قد تم استخدام نماذج IRT لربط اختبارات مقررة لمجموعات أحادية اللغة في دراسات متعددة (مثلاً: آنغوف وكوك 1988).

إن النقد الأساسي لدراسات الربط الأحادية اللغة المرتكزة على IRT هو أن تلك الدراسات تصنع افتراضاً غير مستقر حول تكافؤ أجهزة قياس المفردة عند سكان البلدين. بكلمات أخرى، يبدو أن نماذج جهاز قياس المفردة العائدة لـ IRT لم تصمد أمام عينات لغوية مختلفة. وسّع سيرسي (1997) مشكلات استخدام IRT لربط اختبارات لغوية مختلفة. أشار إلى أن "تقديم برهان تجريبي لعدم تغيّر المفردة

عبر اللغات، يتطلب معياراً متفقاً عليه وساري المفعول. إن مقياس الكفاءة (IRT مقياس ثيتا) هو معيار متفق عليه وعرضة للخطأ بسبب عدم وجود مفردات عامة" (صفحة 14). استمر سيرسي في الإشارة إلى أن سير عمليات القياس IRT مثل المعايرة المتزامنة وسير عملية ستوكنغ - لورد (1983) لا يحل المشكلة؛ لأنها تتطلب قياس ما للفروق في الكفاءة بين مجموعتين لغويتين؛ من المحال نظرياً الحصول على هذا القياس دون منظومة مفردات عامة صحيحة.

طريق آخر لتبيان المشكلات المقترنة باستخدام إجراءات IRT للتصاميم أحادية اللغة هو أن تلك الإجراءات تدعي بناء تكافؤ عبر المفردات العامة، وبصورة لا متناهية عبر اختبارات مختلفة مقررة لمجموعات أحادية اللغة.

باستخدام نقاش لين (1993) لتصنيف دراسات الربط، يتم تصنيف النتائج لمعظم دراسات الربط أحادية اللغة، في أحسن الحالات، كدراسات تعديل إحصائية (دراسات تتضمن اختبارات ربط لتراكيب مختلفة) وتكون خاضعة لكل التحذيرات التي يتم تطبيقها حرفياً عند تفسير نتائج دراسات التعديل.

على الرغم من الآراء النقدية لتصميمات مجموعة أحادية اللغة التي أثيرت في وقت مبكر، تجدر الإشارة إلى أن تطبيق نمط التصميم هذا يحدث بصورة متكررة ويقدم غالباً نتائج مفيدة جداً. تدور الموضوعات المقترنة بهذا النمط للتصميم حول تفسير نتائج الدراسة. يتم تفسير تلك النتائج أحياناً كما لو كانت حصيلة دراسة توازن. من المهم أن نلاحظ أنه ببساطة بسبب أن تصميماً متساوياً قد استخدم، لا يعني أن الدراسة قد أنتجت علامات نهائية تعادلية. إن درجات نهائية متساوية، في مفهوم تلك التي تم الحصول عليها من دراسة تعادلية حرفية، تحدث فقط إذا كانت الافتراضات المتضمنة لنموذج التساوي تلتقي مع المعطيات. على أية حال، غالباً ما ينجم عن تطبيقات تصاميم أحادية اللغة درجات نهائية يمكن اعتبارها قابلة للمقارنة بدرجة كافية للأغراض المستخدمة لأجلها.

تصاميم مجموعة ثنائية اللغة:

وصف سيرسي (1997) متغيرات ثلاثة لتصميم مجموعة ثنائية اللغة. الأول هو التصميم الذي تأخذ فيه مجموعة وحيدة ثنائية اللغة للمتقدمين للاختبار كلا الترجمتين اللغويتين للاختبار في ترتيب متساوٍ. أشار سيرسي إلى أنه ربما يكون عائق واحد لهذا النمط من هذا التصميم من آثار التدريب. هذا صحيح على الأخص إذا مثل الاختبارات تكييفات قريبة جداً من اختبار واحد. التصميم ثنائي اللغة الثاني هو التصميم الذي فيه كل واحدة من المجموعات المتساوية ثنائية اللغة تأخذ عشوائياً نسخة من الاختبارات ليتم ربطها. قدم سيرسي الحجة على أن الانطلاق المحتمل لهذا التصميم هو الإمكانية في أنه قد تزول مجموعات عشوائية لأنها ليست متساوية. الثالث هو التصميم الذي تستجيب فيه عشوائياً مجموعات متساوية ثنائية اللغة إلى خليط من مفردات لغة أصلية ولغة هدف.

استمر سيرسي (1997) في القول بأن إحدى المشكلات الرئيسة مع التصاميم ثنائية اللغة هي تعريف "ثنائية اللغة" إجرائياً. من الصعب تعيين متقدمين للاختبار يملكون كفاءة متساوية في كلتا اللغتين موضع الاهتمام، خاصة عندما يعتبر المرء كفاءة اللغة وكأنها تمت إلى التركيب الذي جرى تقييمه. موضوعات إضافية متصلة باستخدام مجموعات ثنائية اللغة في تقييمات لغات مختلفة يجري وصفها بالتفصيل في سيرسي (الفصل 5 من هذا المجلد).

عائق أساسي لتصاميم الربط ثنائية اللغة هو أنه ربما لا تمثل مجموعة ثنائية اللغة أياً من مجموعات أحادية اللغة التي هي المجموعات موضع الاهتمام في الدراسة المقارنة. إن لهذا التحديد مضامين جدية من أجل تعميم نتائج دراسة ربط لغات مختلفة جرى إنجازها باستخدام مجموعة ثنائية اللغة لمجموعات أحادية اللغة.

تصاميم مطابقة لأحادية اللغة:

إنه بإعطاء المشكلات الموصوفة سابقاً مع تصاميم بسيطة لمجموعات أحادية

اللغة ومجموعات ثنائية اللغة، تكون الإمكانية في استخدام تصميم يطابق مجموعات منفصلة أحادية اللغة على بعض المتغيرات التي ربما تؤثر على نتائج الربط تكون الإمكانية مغرية تماماً. على أية حال، تصاميم كهذه نادراً ما جرى استخدامها بنجاح. تحاول التصاميم المطابقة لأحادية اللغة أن تتجاوز الحاجة إلى مواد عامة لكي يقيم الفرق في المهارات/ القدرات وذلك باستخدام مجموعات لأجل دراسة الربط الذي يتفق مع الآراء النقدية الوثيقة الصلة أياً كانت المهارات أو القدرات التي يجري تقييمها بواسطة اختبارات لغة مختلفة.

كما أشار سيرسي (1997)، كان قد جرى التحري بشكل موسع تماماً عن تأثيرات مجموعات مطابقة في تصاميم متعادلة لأشكال تقليدية (انظر كوك، إيغنور، وشميت 1989، إيغنور، ستوكنغ، وكوك 1990، كولن 1990؛ ليفنغستون، دورانز، ورايت 1990؛ سكاغز 1990). لقد جرى خلط نتائج هذه الدراسات. اقترح ليفنغستون وآل أنه ربما يتم تحسين التساوي عبر التطابق في ميل نزوع الدرجات النهائية (روزبنون وروبن 1983)، بينما حذّر كوك وآل من تقنيات كهذه. إن التحذيرات نفسها المذكورة عند تقييم استخدام مجموعات ثنائية اللغة من أجل دراسات ربط لغات مختلفة ينبغي أن تذكر في سياق استخدام مجموعات متطابقة لأجل أنماط دراسات الربط تلك.

نقطة رئيسة أثارها لين (1993)، عند مناقشة دراسة الربط للتصنيفات، وكانت في أن كل دراسات ربط أخرى غير دراسات التساوي الصحيحة تعاني من مشكلة اعتماد النتائج على المجموعة. بناء على ذلك لا يمكن تعميم النتائج لدراسة ربط لغات مختلفة تم إنجازها باستخدام مجموعات أحادية اللغة مبنية بحيث إنها تتطابق مع متغيرات خاصة هي مفتاح القدرة المقاسة. لا يمكن تعميم تلك النتائج على مجموعات أحادية اللغة متغايرة الخواص بدرجة أكبر والتي تكون بشكل نهائي المجموعات مركز الاهتمام.

في القسم التالي من هذا الفصل تجري مناقشة دراسات ربط ثلاث جرى العمل بها عبر العشرين سنة الماضية بغرض ربط الدرجات النهائية في SAT و PAA يتم نقد كل دراسة من منظور المناقشة السابقة حول تصاميم الربط.

تطوير علاقة بين الدرجات النهائية لـ PAA** وSAT*

لقد تم تصميم سلسلة دراسات جرت لتطوير علاقة بين الدرجات النهائية لاختبار SAT والدرجات النهائية لاختبار PAA لتطوير مقياس عام يسهّل مقارنات الدرجات النهائية المحصّلة في الاختبارين.

كان الباحثون العاملون على كل الدراسات مدركين أن الفروق الأساسية في اللغة، والعادات، والقيم ربما تضعف بصورة ممكنة مقارنات بين مجموعات تأخذ الاختبارين. على أية حال، كان أولئك الباحثون ملتزمين بتطوير علم منهج مثالي يمكن استخدامه لبناء مقياس غير منحاز على قدر الإمكان.

من المهم في تلك النقطة التأكيد على أن اختبارات PAA ليست ترجمة مباشرة أو تكييف لـ SAT بالرغم من أن اختبارات PAA مصممة لقياس التراكيب نفسها مثل الـ SAT، فاختبارات PAA تحتوي على مفردات مختلفة ويتم تطويرها بشكل مستقل تماماً عن اختبارات SAT تم اتخاذ قرار من قبل مجلس الكلية، باكراً في تاريخ برنامج الاختبار لـ PAA، أنه بسبب التعقيدت والصعوبات المتضمنة في تكييف اختبار من إحدى اللغات إلى الأخرى، يكون من الأفضل حفظ "التساوي" بين الاختبارين إذا كان كل اختبار مصمماً لقياس "التركيب نفسه" لكن بلغة مختلفة.

هناك ظاهرة مميزة لاختبارات PAA وهي أنه جرى تصميمه ليستخدم في مضامين متعددة للناطقين بالإسبانية. فسكان بلاد «الهسيبانك» المتنوعين، على

(*) Scholastic Assessment Test (SAT).

(**) Prueba de Aptitude Academica (PAA) (النسخة الإسبانية للاختبار الأمريكي AST)

سبيل المثال، المكسيكيون البروتريكيون يختلفون الواحد منهم عن الآخر بدرجة كبيرة بالطريقة نفسها، مثلاً لنقل اختلاف رعايا الولايات المتحدة ورعايا بريطانيا العظمى. تتكلم كلتا هاتين المجموعتين الإنكليزية، لكن الفوارق الطفيفة في اللغة تختلف في بلدان مختلفة. جرى نقل تحليلات متفاوتة لتوظيف المفردة (DIF) إلى الـ PAA لتأكيد صدق البنية التي يقيسها الاختبار عبر سكان بلاد الهسبانك المختلفين (انظر مثال، سيرسي وآلالوف 2003).

كانت الدراسة الأولى التي أجراها آنغوف ومودو (1973) في خريف 1971 قد تم توجيهها من أجل غرض ربط الدرجات النهائية في الـ PAA بالدرجات النهائية في الـ SAT وقد جرى استخدام نتائج دراسة آنغوف/ مودو لمقارنة الدرجات النهائية في الـ PAA بالدرجات النهائية لـ SAT لحوالي مدة عشر سنوات. قادت التقدمات في التقنية، كما في التحقق من أن تدريباً جيداً لتكرار ومراجعة نتائج دراسات الربط بشكل دوري. قادت إلى تكرار ربط SAT/PAA بدراسة آنغوف وكوك (1988). اتبعت دراسة آنغوف - كوك التصميم الأساسي للدراسة الأبكر، ولكنها استبدلت نظرية الاختبار الكلاسيكية لعلم المنهج بتقنيات IRT جرت إدارة دراسة الربط الأكثر حداثة والمعتمدة من الـ PAA والـ SAT من قبل شميت، دورانز، ماغرينا، وكوك (1998).

وكان الغرض من هذه الدراسة تقديم عامل ربط حديث للاختبارين الذي عكس تغييرات حديثة في مواصفات خواص الاختبار. وظفت الدراسة الثالثة علم منهج مختلف تماماً من أجل ربط الاختبارين من علم المنهج الذي استخدم في الدراستين السابقتين. ما يتبع هو نقاش موجز ونقد للدراسات الثلاث للربط.

دراسة آنغوف-مودو:

طوّر آنغوف ومودو (1973) علم منهج لتأمين تحويل درجات نهائية رياضية ولفظية باللغة الإسبانية PAA إلى درجات نهائية رياضية ولفظية خاصة بالـ SAT جرى تقرير كلا الاختبارين على طلاب مدرسة ثانوية من أجل أغراض القبول في

كلية. وكما جرى ذكره سابقاً، بالرغم من أن PAA و SAT تشاركت في البنية نفسها والشكل نفسه، فقد تم تأليف كل منهما من مفردات أصلية مستقلة، مما يعني أن الاختبارات لم تكن تراجم مكيّفة الواحدة منهم من الأخرى. كان الغرض من الدراسة المنفذة من قبل آنغوف ومودو تأمين لوائح تحويل بين الـ PAA والـ SAT تسهل مقارنات مباشرة للمجموعات الفرعية من مجموعتي اللغة اللتين أخذتا اختباراً ملائماً بلغاتهم الأصلية. بالإضافة إلى ذلك، كان من المتوقع أن تساعد لوائح التحويل في تقييم الأرجحية للنجاح في كليات البلد الرئيس الذي ربما تم الحصول عليه من قبل طلاب من بورتوريكو.

تألفت دراسة آنغوف ومودو (1973) من جانبين. تضمن الجانب الأول انتقاء مفردات "عامة" تمّ استخدامها كاختبار معتمد في دراسة "التوازن" ويتألف الجانب الثاني من "تساوٍ" فعلي. كان على الطريقة المستخدمة في الجانب الأول أن تختار منظومتي مفردات، الواحدة بالإنكليزية أصلياً والثانية بالإسبانية أصلياً، وأن تترجم كل منظومة إلى اللغة الأخرى. وبعد الترجمة، تم تقرير منظومتي المفردات (واحدة بالإسبانية وواحدة بالإنكليزية) على طلاب ملائمين أحادي اللغة من أجل أغراض اختبار تمهيدي. تم تنفيذ إدارات الاختبار التمهيدي لهذه الدراسة في خريف عام 1970، وعلى أساس تحليل معطيات الاختبار التمهيدي، جرى انتقاء منظومتي مفردات، الواحدة لفظية والثانية رياضية، كمنظومات مفردات "عامة" لكي يجري استخدامها من أجل "عمليات تساوٍ" رياضية ولفظية خاصة.

في "التساوي" من الدراسة، جرى تقرير المفردات "العامة"، الظاهرة في كل من الإسبانية والإنكليزية، بلغة ملائمة إلى جانب ومع الشكل الإجرائي لـ PAA في نوفمبر 1971، ومع الإجرائي لـSAT في يناير الثاني 1972 وقد تم استخدام المعطيات من تلك الإدارات لتوجيه كل عمليات التساوي الخطية و Equipercentile ذات التصنيف المئوي لقيم المتغير إلى الاختبارات الرياضية واللفظية لـ PAA وSAT.

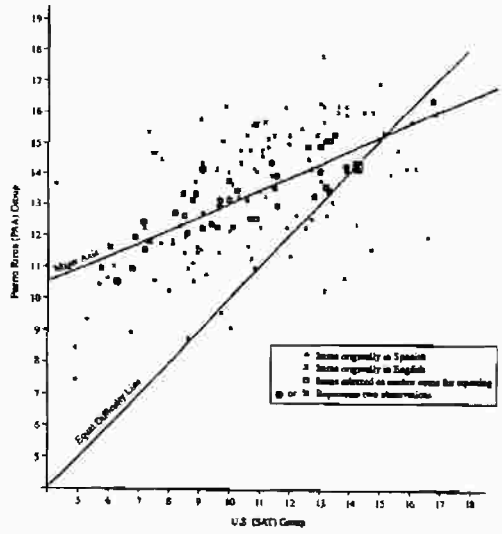
هناك مظاهر عدة من هذه الدراسة المبكرة تستحق الوصف بالتفصيل. تتألف المرحلة I للدراسة من بناء مفردات "عامة" أو اختبار رابط يستخدم لتقييم الفروق في المقدرة بين المجموعة PAA الناطقة بالإسبانية والمجموعة SAT الناطقة بالإنكليزية. وقد تم رسم إطار المجموعة الأولية للمفردات التي سيجري استخدامها لتشكيل اختبارات الربط في أعداد متساوية تقريباً من أطر مفردات الـ PAA والـ SAT. جرت ترجمة تلك المفردات إلى اللغة الثانية من قبل مجموعة صغيرة من خبراء ثنائي اللغة. جرى بذل جهد لإنتاج منظومة مفردات، بالإنكليزية والإسبانية، التي كانت، تقريباً على قدر الإمكان، مساوية بالمعنى في اللغتين. وفي وقت لاحق جرت إعادة ترجمة كل المفردات إلى لغتهم الأصلية ومقارنة الترجمات المعادة (النسخ التي خضعت إلى ترجمتين) مع النص الأصلي.

ومن ثم جرى اختبار تمهيدي لمجموعة المفردات العامة وذلك بتطبيقها على مجموعات طلاب تقدموا لكل من اختباري الـ PAA أو الـ SAT باللغة المناسبة. وتلي الاختبار التمهيدي، عملية غربلة المفردات إحصائياً وذلك برسم العلاقة البيانية بين صعوبة المفردات وقيم دلتا لكل من المفردات الرياضية واللفظية المأخوذة من قبل مجموعات ناطقة بالإنكليزية وبالإسبانية. (انظر آنغوف ومودو، 1973، من أجل وصف "المثلث"). كان الغرض من رسوم دلتا البيانية هو التمكّن من تعيين المفردات التي تملك معنى مختلف للمجموعتين، PAA و SAT وقد اعتبرت المفردات ملائمة بدرجة متساوية للمجموعات الناطقة بالإنكليزية كما هي للناطقة بالإسبانية على أساس قربهم من المحور الرئيس للقطع الناقص من رسم دلتا البياني كما في الأشكال 6-1 و 6-2 المأخوذة من آغنوف ومودو (1973)، توضح نتائج الرسوم البيانية المثلثية لمنظومات مفردات الربط الرياضية واللفظية.

نقطة مهمة ينبغي ملاحظتها وهي أنه عند مقارنة الرسم البياني للمفردات اللفظية مع الرسم البياني للمفردات الرياضية نجد درجة أعظم من التشتت في المفردات اللفظية عن المحور الأساسي في القطع الناقص لرسم دلتا البياني.

فسّر آنغوف ومودو (1973) التشتت الكبير للمفردات اللفظية كمشير على أن المفردات اللفظية لا تملك تماماً المعنى النفسي لمجموعتي اللغة. وتابعا القول في أن التشتت كان كافياً لإبقاء الشك حول نوعية أي توازن تم القيام به مع تلك المفردات. جرى تحسين الوضع بحذف المفردات الأكثر شذوذاً؛ على أية حال، استمر المؤلفان في إبداء الاهتمام بأن تفاعلات المجموعة بالمفردات والمشار إليه في المعطيات الموجودة في رسوم دلتا البيانية قد جعل التوازن "أقل جدارة كثيراً بالثقة مما هو متوقع من تعادل اختبارين متساويين جرى إعدادهما من أجل أعضاء في ثقافة اللغة نفسها" (صفحة 14).

الوجه II من الدراسة يتألف من "تساوٍ" فعلي لـ PAA و SAT، باستخدام منظومة مفردات الربط المبنية في الوجه I جرى انتقاء أربعين من المفردات اللفظية وخمس وعشرين من المفردات الرياضية كمفردات عامة لأجل الوجه II من الدراسة. تمّ إقرار المفردات، في موازاة ومع نسخة معدلة إجرائية للاختبارات الخاصة في نوفمبر 1971 الإدارة لـ PAA، ويناير 1972 الإدارة لـ SAT إجراءات «تساو» متفقة مع القواعد المقررة التي جرى استخدامها لربط الدرجات النهائية في الـ PAA مع الدرجات النهائية في الـ SAT وقد جرى استخدام مقياس النشاط الطولي الخطي (Tucker).

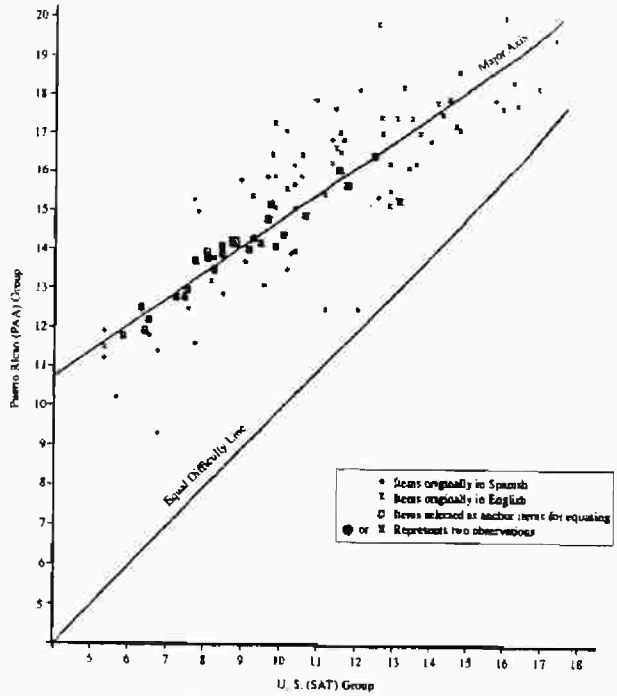


U.S (SAT) Group

الشكل 1.6 رسم دلتا البياني للمفردات اللفظية.

ومقياس التصنيف المئوي لقيم المتغير (آنغوف، 1984؛ تجري الآن الإشارة إليه كمقياس تصنيف مئوي "متسلسل")، وسير العملية الطولي لـ ليفن (1955). نظراً لأن المعطيات لا تقابل أياً من نماذج التوازن الثلاث المتعارف عليها، تقرر إيجاد معدل نتائج النماذج الثلاث، بإعطاء وزن أعظم لنتائج التصنيف المئوي لقيم المتغير. تظهر الأشكال 306 و 406 رسوماً بيانية متبعثرة للتعادلات الرياضية واللفظية.

أشارت نتائج "التساوي" اللفظي إلى أن مقياساً وسطياً قيمة (500) لـPAA كان مساوياً لمقياس لفظي درجة نهائية (350) لـ SAT فعلياً أدنى من القيمة الوسطية. أشارت النتائج في "تساوي" الرياضيات إلى أن الدرجة النهائية لـPAA لـ 500 نتجت حتى في أخفض درجة رياضيات نهائية لـ (319) SAT.



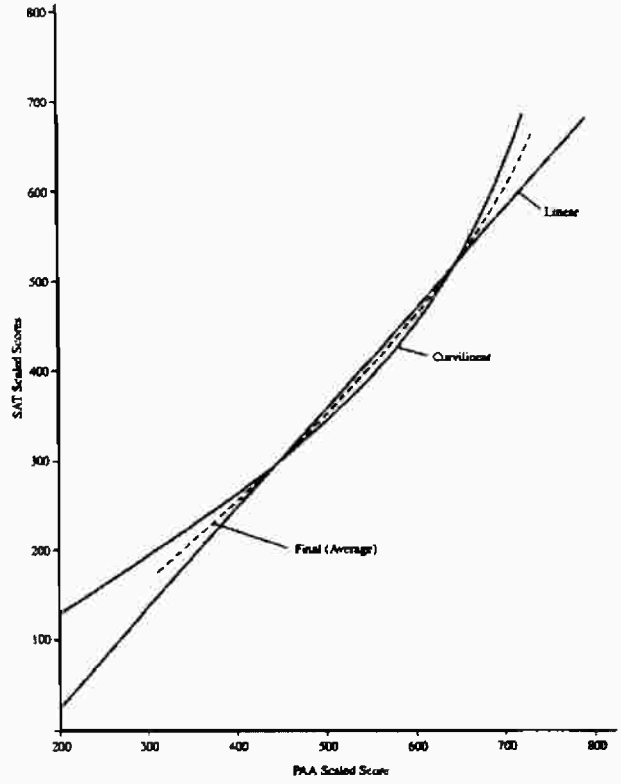
U.S (SAT) Group

الشكل 2.6 رسم دلتا البياني لمفردات الرياضيات.

حذّر آنغوف ومودو (1973) من أن "الدقة في تلك التحويلات هي محددة بملاءمة الطريقة المستخدمة في اشتقاقها والمعطيات المجمعة في أثناء مجرى الدراسة، من المأمول أن تكون تلك التحويلات مفيدة في تنوع للمحتويات لكن....لكي تكون مفيدة ستحتاج في كل لحظة إلى أن تعزز بمعطيات إضافية متميزة للمحتويات" (صفحة 41).

دراسة آغنوف-كوك:

استخدمت الدراسة التي قام بها آنغوف وكوك (1988) التصميم الأساسي نفسه مثل ذلك الذي جرى استخدامه من قبل آنغوف ومودو (1973).

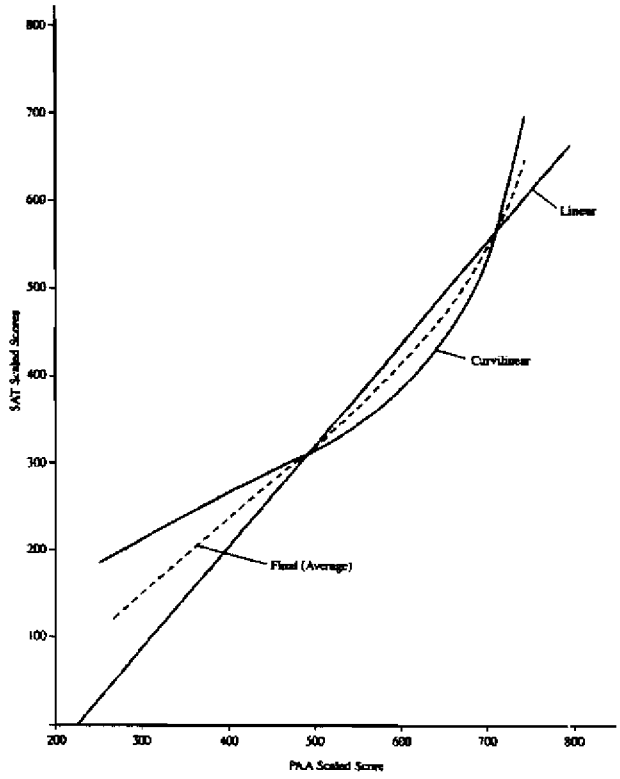


الدرجات النهائية المقاسة PAA

الشكل 6-3 نتائج التعادل للاختبارات اللفظية.

لكنها وظّفت علم المنهج IRT لتأخذ مكان كل من تشتت مفردة الرسم البياني المثلثي المستخدمة لانتقاء اختبار المفردات العامة ومن علم منهج التساوي المتعارف عليه والمستخدم في الدراسة الأبكر.

على وجه شبه الدراسة السابقة، جرى القيام بدراسة آنغوف وكوك (1988) في جانبين. تألّف الجانب الأول من انتقاء



الدرجات النهائية المقاسة PAA

الشكل 6 - 4 نتائج التعادل لاختبارات الرياضيات

مفردات الربط التي سيجري استخدامها في الجانب II، وجه "التساوي" للدراسة. لأجل دراسة آنغوف وكوك، تمّ اتباع علم المنهج المؤسس في دراسة أسبق من أجل التكييف، إعادة التكييف، الاختبار التمهيدي للمفردات. كان الفرق بين

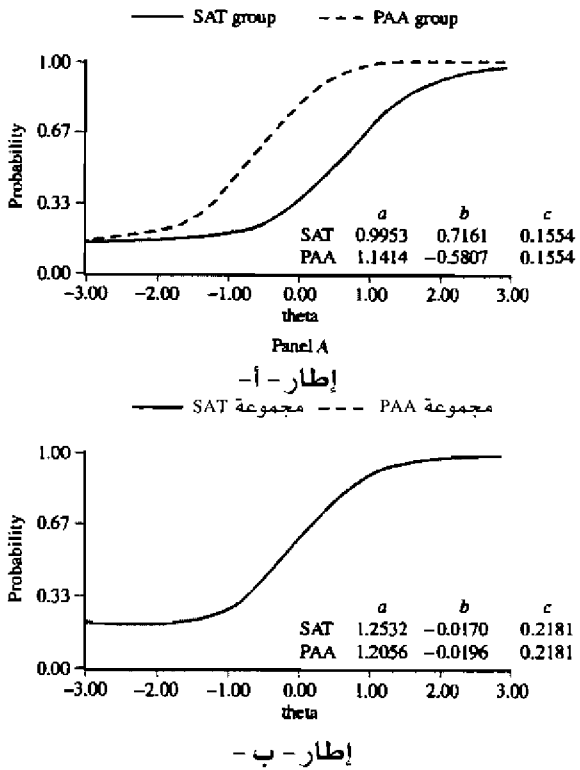
الوجه I في الدراستين هو علم المنهج المستخدم في تشتت المفردات لأجل الانتقاء كمفردات "عامة". قيّم آغنوف وكوك المفردات بمقارنة الفروق، بصرياً وإحصائياً، بين المنحنيات الخاصة للمفردة (ICCs) IRT.

يعطي الشكل 6-5 أمثلة على الرسوم البيانية لـ ICCs من أجل مفردات رياضيات ولفظية جرى تقديرها لمجموعات ناطقة بالإنكليزية، وبالإسبانية. يظهر الإطار (أ) في الشكل 6-5 أنه من أجل كل مستويات القدرة (Theta) تملك مجموعة PAA احتمالاً أعلى في الحصول على إجابة صحيحة للمفردة منه في احتمال مجموعةSAT مفردة كهذه لا يمكن بوضوح اعتبارها مفردة "عامة" للمجموعتين وبالتالي سقطت أثناء جانب تبعثر المفردة في الدراسة. يحتوي الإطار "ب" في الشكل 6-5 مقارنة لـ ICCs المحصلة لأجل مادة رياضيات معطاة إلى مجموعات SAT و PAA.

بالمقابلة مع المنحنيات الظاهرة على الإطار (أ)، تكون ICCs لأجل مفردة رياضيات معطاة إلى مجموعتين من المتقدمين للاختبار تقريباً مطابقة؛ وذلك يعني أن الأفراد بكافة مستويات القدرة في كلتا المجموعتين يملكون الاحتمال نفسه في الحصول على إجابة صحيحة للمفردة. ولا تفضّل المفردة أياً من المجموعتين. مفردات كهذه المفردة يمكن أن تعتبر مثالية من أجل الإدخال في منظومة المفردات "العامة" المستخدمة في ربط اختباري الرياضيات.

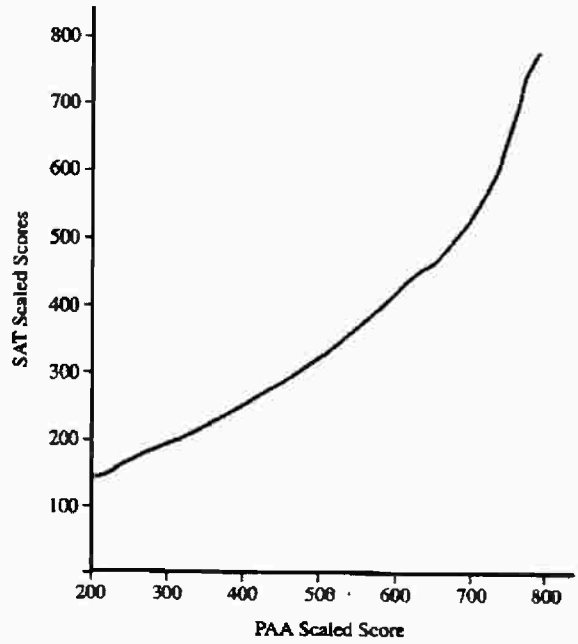
يحاذي جانب "التساوي" في الدراسة نظيره في دراسة آنغوف ومودو (1988) باستثناء استخدام إجراءات IRT تم إقرار منظومات مفردة "عامة" لفظية ورياضيات إلى الأمام مع اختباراتهم الإجرائية الخاصة (SAT) أو (PAA) على مجموعات ملائمة أحادية اللغة. جرى جمع معطيات SAT في كانون الأول 1985 ومعطيات PAA في الإدارة في تشرين الأول 1986 كانت طريقة التساوي الـ IRT المستخدمة في هذه الدراسة هي التساوي لـ IRT المطبق (كوك وإيفنور، 1983؛ بيترسن، كوك، وستوكنغ، 1983) فقط تم الإبلاغ عن نتائج التساوي المشكلة بخط منحن لـ IRT من أجل الدراسة. يجري تمثيل هذه النتائج لاختبارات الرياضيات واللفظية في الأشكال 6-6 و6-7 .

من الواضح من مراجعة الخطوط البيانية المعروضة في الأشكال 6-6، 6-7، أن العلاقة بين مقاييس الـ PAA والـ SAT متسمة بخطوط منحنية بشكل ملحوظ. هذه كانت أيضاً الحالة بالنسبة إلى نتائج تعادل محصّلة في دراسة آنغوف ومودو (1973)؛ على أية حال، اختار آغنوف ومودو أن يوجدا معدل النتائج المشكّلة خطاً منحنياً مع النتائج الخطية. وتشير نتائج دراسة آنغوف وكوك (1988) إلى أن الفروق بين مقاييس الـ PAA والـ SAT حسب الدرجة النهائية 500 لـ PAA كانت حوالي 180 إلى 185 علامة. أشارت نتائجهما إلى فروق مشابهة بالنسبة لربط الرياضيات، وهذا يعني، في الدرجة النهائية PAA من 500، كانت الفروق بين مقياس الـ PAA والـ SAT حوالي 180 إلى 185 علامة.



الشكل 6-5 منحنيات استجابة المفردة. رسوم بيانية لوظائف استجابة المفردة من أجل المفردات اللفظية (الإطار أ) والرياضية (الإطار ب) مفردات معطاة إلى مجموعات SAT ومجموعات PAA، موضحة اتفاق جيد أو ضعيف بين المجموعات.

سلمت نتائج الدراسة لـ آنغوف وكوك (1988) تحويلات الدرجات النهائية اللفظية PAA إلى المقياس اللفظي SAT أدنى جوهرياً من الدراسة الأبكر، على الأخص في المجال الأوسط لمقياس الدرجة النهائية. أظهرت التحويلات إلى المقياس الرياضي لـ SAT اتفاقاً أفضل مع النتائج الأبكر. فكّر المؤلفون أن الفروق في النتائج قد تعزى إلى فروق في علم المنهج أو إلى صعوبات متأصّلة لاختبارات "التساوي" المثقلة لفظياً نسبة إلى مجموعات لغة مختلفة.



شكل ٦,٦ نتائج تعادل الاختبارات اللفظية

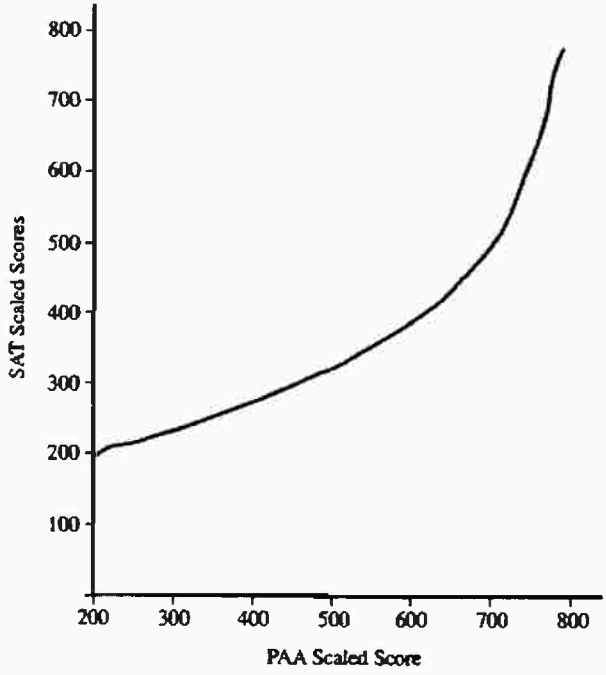
كتابة نقدية للدراستين السابقتين:

إن علوم المنهج لربط الدرجة النهائية الموظّفة في الدراستين، سير عمليات التساوي المتّسمة بخطوط منحنية وسير عمليات توكر وليفين والتصنيف المئوي لقيم المتغير، هي سير عمليات معقول للاستخدام إذا كان السؤال موضع الاهتمام هو مقارنة توزيعات الدرجات النهائية لمجموعتين من المتقدمين للاختبار (أو في حالة

طرائق خطية، الحركتين الأوليتين للتوزيع). تذكر أن الأغراض الأصلية لربط /SAT PAA، كما جرى وصفها من قبل آنغوف ومودو (1973)، كانت:

(أ) أن نقارن توزيعات الدرجات النهائية للمجموعات الفرعية من سكان بلدين.

(ب) أن نقيّم النجاح المحتمل لطلاب بورتو ريكان الذين كانوا مؤخراً مهتمين بالمواظبة على الحضور في كليات البلد الرئيس. وكانوا يخضعون إلى الدرجات النهائية لـ PAA من أجل أغراض القبول.



الشكل 6 -7 نتائج تساوي اختبارات رياضيات

بالاعتماد على مدى الجودة التي لاقت فيها المعطيات افتراضات النماذج الإحصائية المستخدمة لأجل الربط، ومدى الجودة التي تمّ بها تنفيذ الربط، من الممكن أن علم المنهج الموظّف في دراسات آنغوف ومودو (1973)، وآنغوف وكوك

(1988)، يستطيع أن يقدم أساساً معقولاً لمقارنات وتوزيعات الدرجات النهائية. سبب ذلك، أن الغرض من إجراءات IRT، الخطية والتصنيف المئوي لقيم المتغير، الموظّفة في هاتين الدراستين هو تحويل توزيع الدرجات النهائية المحصّلة في اختبار واحد لتلائم تلك المحصّلة في الاختبار الثاني، لأجل مجموعة خاصة من المتقدمين للاختبار.

بالرغم من أنه يمكن نظرياً لتقنيات إحصائية مستخدمة في هذه الدراسات أن تقدم نتائج نهائية تقابل هدف الدرجات النهائية المقارنة، من المحتمل ألا يقابل الهدف من قبل نتائج الدراستين بسبب طبيعة المعطيات المستخدمة لأجل الربط. على أية حال، سيقدم الإجراء المستخدم حلاً للمشكلة شريطة أن تتلاقى الافتراضات الأساسية، في علم المنهج.

خذ بعين الاعتبار السبب الثاني لتنفيذ الربط SAT/PAA، وهذا يعني تقييم نجاح الطلاب من بورتوريكو في كليات البلد الرئيس والجامعات، كيف يكون بالإمكان استخدام العلاقة المؤسسة بين الـ PAA و SAT في الدراستين السابقتين للتنبؤ بمدى الجودة التي ربما يحققها طالب ثانوي في سان جوان عندما ينتظم هو أو هي في كلية، لنقل، في ميامي! ماذا يعني أن نقول إن طالباً حصل على 500 في الـ PAA سيحصل على درجة نهائية 320 في الـ SAT إذا لم يتكلم الطالب الثانوي في سان جوان الإنكليزية، بالتأكيد لن يحصل هو أو هي على درجة نهائية 320 في الـSAT وسيقضي وقت دراسة شديد الصعوبة في كلية أو جامعة في الولايات المتحدة.

بحثت بينوك - رومان (1995) العلاقة بين الدرجات النهائية لاختبارات قبول مستوى الخريجين معطاة بالإنكليزية وبالإسبانية لمجموعة من الطلاب الذين كانوا أكثر تمكناً بالإسبانية منهم بالإنكليزية ووصلت إلى أن التمكن من الإنكليزية ساهم في درجات الطالب النهائية في اختبار سجل الخريجين (GRE) اللفظي واختبارات علم الأحياء، علم النفس، التحليلية، الرياضيات. وجدت بينوك-رومان أن التمكن من الإنكليزية ساهم بشكل مختلف معتمداً على مستوى تمكن الطالب ومعتمداً على

فحوى الاختبار. وقد ناقشت حقيقة أنه من الممكن لطالب موهوب في اللغة الثانية أن يحصل على درجة نهائية دون المعدل في الاختبار اللفظي GRE ببساطة بسبب فهم قراءة أبطأ.

إن المعاني المتضمنة في دراسة بينوك - رومان (1995) من أجل دراسات ربط (مقارنة) SAT/PAA هي أنه إذا كان الغرض الرئيس من الدراسة هو تقديم وسيلة من أجل تقييم مدى الجودة التي سيؤديها طالب في بورتوريكو في كلية البلد الرئيس أو جامعة، ثم قياس ما للمقدرة باللغة الإنكليزية يجب أن يؤخذ بالاعتبار. يحاول علم المنهج المستخدم لأجل دراسة الربط SAT/PAA الموصوف لاحقاً أن يأخذ هذه المعاني المتضمنة في الحسبان.

دراسة شميت، دورانز، ماغرينا، وكوك:

في ربيع 1994، جرى إدخال SATI جديد. يحتوي الاختبار الجديد على أنماط مفردات جديدة وتم بناؤه وفقاً لمواصفات إحصائية ومحتوى منقّح. (انظر كوك، 1995)، لأجل وصف مراجعات لـ (SAT) في أكتوبر 1996، تم أيضاً تنقيح الـ PAA لتشمل أنماط مفردات جديدة، محتوى منقّح، ومواصفات إحصائية. تغييرات في الـ PAAموازية للتغيرات المدخلة في SATI على الأخص، لم تشمل الـ PAA اللفظية الجديدة على مفردات مناقضة وتملك نسبة مئوية أعلى من المفردات اللفظية التي تتصل بمقاطع قراءة نقدية (56 ٪ مقابل 31 ٪). بالإضافة إلى المفردات المتعددة الاختيار التقليدية، يتضمن الـ PAA الجديد للرياضيات مفردات حيث ينتج المتقدم للاختبار استجابته أو استجابتها الخاصة. (انظر مجلس الكلية، 1995، لوصف موسّع للتغييرات في PAA الجديد). الفرق الوحيد بين الـ SATI والـ PAA الجديد هو أن الـ SATI يسمح باستخدام الآلات الحاسبة في اختبار الرياضيات.

قدمت دراستا الربط SAT/PAA السابقتين جداول اتفاق بين الدرجات النهائية في نسخ سابقة لـ PAA و SAT آنغوف وكوك، 1988؛ آنغوف ومودو،

1973 استخدمت دراسات القياس هذه مفردات "عامة" لتتلاءم مع أية فروق بين مجموعات الـ SAT و PAA. وبسبب الموضوعات الموصوفة باكراً، اقتربت آخر دراسة من التشابه في الدرجات النهائية بين الـ SATI والـ PAA من منظور مختلف. تم الحصول على جداول اتفاق بالرغم من أن طرائق "التساوي" المستخدمة في الدراستين الأبكر تدّعي أن الاختبارين كانا جوهرياً أشكال بديلة تمثّل البناء نفسه. نظراً لكون الـ PAA باللغة الإسبانية والـ SATI باللغة الإنكليزية وتحتوي كل منهما على مفردات تمّ تطويرها بصورة دقيقة واختبارها مسبقاً على السكان المعنيين الخاصين في كل بلد، فليس بالإمكان اعتبار الـ PAA والـ SATI أشكالاً بديلة.

لأجل الدراسة الثالثة PAA/SAT من قبل شميت وآل (1998)، جرى استخدام طريقة تنبئية. جرى افتراض أن تطوير جدول الاتفاق لم يكن أساسياً لنجاح الدراسة. بناء على ذلك قدمت طريقة التنبؤ الموظّفة قياس مقدرة اللغة الإنكليزية، اختبار الإنجاز للإنكليزية كلغة ثانية (ESLAT) ، التصميم الأساسي لأجل الدراسة.

دراسات سابقة وثيقة الصلة بهذا البحث، تم استخدام طريقة الارتداد من قبل ألدرمان (1981) وبولدت (1969) لدراسة العلاقة بين اختبارات معطاة بالإنكليزية وبالإسبانية إلى مجموعات ثقافية مختلفة. في دراسة ألدرمان، جرى اختبار طلاب على الـ PAA، (SAT TOEFL اختبار الإنكليزية كلغة ثانية)، و ESLAT جرى اعتبار التمكن من اللغة في (TOEFL) أو (ESLAT) كمتغير وسيط في التنبؤ بنتائج اختبار SAT من نتائج PAA نجم عن التمكن الأعلى من اللغة، كما تم قياسه بواسطة تلك الاختبارات، علاقة أقوى بين الدرجات النهائية لـ PAA و SAT تحدد تلك النتائج أهمية استخدام قياس التمكن من اللغة عند خلق معادلة تنبؤ بين SAT و PAA نظراً لأن جميع المتقدمين لاختبار PAA يأخذون أيضاً اختبار إنجاز الإنكليزية كلغة ثانية (ESLAT) من أجل أغراض القبول، يمكن أن تستخدم نتائج ESLAT كمتغير وسيط للغة. كانت عينة الدراسة الثالثة PAA/SAT شميت وآل، 1998 مرشحين متوفرين أخذوا اختبار PAA الجديد في بورتوريكو من يناير

الثاني 1996 إلى يونيو 1997، في التحاليل تم فقط اعتبار الدرجة النهائية الأخيرة للطلاب الذين أعادوا الاختبار ضمن مدة زمنية محددة. كان لكل طالب مشمول بالدراسة درجات الاختبار النهائية التالية:

(أ) SATI اللفظية والرياضيات، من يناير 1996 على يونيو 1997 (تقديم الاختبار).

(ب) PAA اللفظية والرياضيات، من أكتوبر 1996 إلى يونيو 1997 (تقديم الاختبارات).

(ج) ESLAT، من أكتوبر 1996 إلى يونيو 1997 (تقديم الاختبارات).

جرى اعتبار كل من نماذج التنبؤ* الخطي والمنحني من أجل التنبؤ بالدرجات النهائية SATI من الدرجات النهائية PAA وESLAT.

وجد شميت وآل (1998) أن الارتباطات بين SATI و ESLAT/PAA في العينة تستحق الملاحظة بدرجة كبيرة. بلغ معامل الارتباط بين PAA رياضيات وSATI رياضيات 0.82 مشيراً إلى أن الاختبارات تقيس تراكيب بنيوية متشابهة، لكن ليست نفسها. إضافة إلى ذلك، ارتبطت SATI رياضيات بـ 0.57 درجة مع ESLAT، ارتباط يشير إلى أن ESLAT تعمل كمتغير وسيط للتمكن من اللغة من جل الدرجات النهائية للرياضيات. الجدول 6-1 يحوي الارتباطات لدرجات الاختبار النهائية تلك.

Test Score	*ESLAT*	*PAA-M*	*PAA-V*	*SAT-M*	*SAT-V*
ESLAT	1.00	.51	.45	.57	.74
PAA-MATH	.51	1.00	.61	.82	.60
PAA-VERBAL	.45	.61	1.00	.56	.62
SAT-MATH	.57	.82	.56	1.00	.69
SAT-VERBAL	.73	.60	.62	.69	1.00

جدول 6-1 الارتباطات بين SAT و PAA اللفظي والرياضيات والـ ESLAT

(*) المقصود استعمال:

Linear and Curvilinear multiple regression models.

تظهر الارتباطات بين ESLAT و SATI اللفظي اقتراحاً أقوى لمتغير وسيط للتمكن من اللغة، بالنسبة إلى SATI الارتباط 0.73 درجة مع ESLAT وأقل بدرجة معتبرة، 0.62 مع PAA اللفظي. لاحظ أن PAA رياضيات تملك فقط ارتباطاً أدنى قليلاً مع PAA اللفظي (0.60) من الارتباط الذي أظهره SATI اللفظي وPAA اللفظي. من الواضح من هذه المعطيات أن أي جدول ارتباط جرى تطويره مستخدماً اختبارات بهذا الترتيب لدرجات الارتباط ستكون له قيمة مثيرة للتساؤل، بغض النظر عن علم المنهج المختار للربط. كانت إحدى الأسئلة لهذه الدراسة فيما إذا استطاع PAA اللفظي أم لم يستطع إضافة الكثير إلى تنبؤ الدرجات النهائية اللفظية SATI إلى درجة أبعد من الذي استطاعت ESLAT فعله بنفسها.

جرت مقاربة معادلات للتنبؤ بدرجات نهائية SAT رياضيات و SAT لفظية من درجات نهائية PAA ودرجات نهائية ESLAT بالصيغة التالية لـ SAT اللفظي:

SAT اللفظي المقدّر = .371* PAA اللفظي + .797* ESLAT-284

لاحظ أن الوزن المعين لـ ESLAT هو أكثر مرتين من ذلك المعين لـ PAA اللفظي. المعادلة التقريبية للتنبؤ بـ SAT رياضيات من PAA رياضيات وESLAT تجعل من الواضح أن PAA رياضيات هو عامل التنبؤ الأكثر أهمية:

SAT رياضيات المقدر= .688* PAA رياضيات + .259* ESLAT - 150

من المستطاع رؤية أن PAA رياضيات تملك وزناً أكثر من مرتين بالسعة من ذلك الذي تملكهESLAT.

تجدر الملاحظة إلى أن المعادلات للتنبؤ SAT لفظي ورياضيات تظهر أن لديها قابلية تطبيق محدودة بسبب كونها غير قابلة للاستخدام من أجل درجات نهائية للاختبار بدرجات نهائية ESLAT أدنى من 550 في العينة لديها علاقة شاردة مع المتغيرات موضع الاهتمام. كان تفسير هذا بما معناه، أن مستوى معيناً من المقدرة

في اللغة الإنكليزية، كما تم قياسها بواسطة ESLAT يكون مطلوباً قبل أن تصبح الدرجات النهائية متصلة نظامياً بالدرجات النهائية الأخرى للاختبار، وأكثر أهمية، قبل أن تستقر العلاقات بين الدرجات النهائية في تراكيب بنيوية مشابهة جرى قياسها بالإنكليزية و الإسبانية. إن الدور البارز لـ ESLAT في التنبؤ بالدرجات النهائية SAT لفظي حتى في هذه المجموعة من الدرجات النهائية العالية ESLAT (550) هو بالتقريب معيار الانحراف [118] فوق المتوسط 446 في عدد السكان الكامل (PAA) يدفع إلى المقدمة، المشكلات في محاولة ربط الدرجات النهائية في الاختبارات مثل الـ PAA والـ SATI التي يجري إعطاؤها في لغات مختلفة إلى مجموعات من خلفيات ثقافية مختلفة. إن ربط درجات نهائية باستخدام علم منهج التنبؤ قد يبدو أنه يقدم نتائج أكثر قابلية للتفسير منه في المحاولة لتأسيس جدول اتفاق باستخدام علم منهج للتساوي التقليدي للاختبار المعتمد الذي جرى توظيفه في الدراستين الأوليتين.

إن تصميم الدراسة الثالثة لديه بالتأكيد عدد من العوائق التي ينبغي لفت النظر إليها. إن الدراسة هي دراسة تنبؤ مستخدمة لتصميم مجموعة ثنائية اللغة. إن عوائق دراسات التنبؤ وتصاميم مجموعة ثنائيّ اللغة التي جرت الإشارة إليها سابقاً في هذا الفصل، لكن تصميم المجموعة ثنائية اللغة المستخدم في هذه الدراسة هو مختلف عن التصاميم التي جرت مناقشتها سابقاً والتي فيها تم استخدام ESLAT كمتغير ملازم. إن العائق الظاهر لهذا التصميم هو أن معادلات التنبؤ المحصّلة من هذه الدراسة تكون مختصّة بمجموعة والعينة المستخدمة لأجل هذه الدراسة غير ممثلة لجميع المتقدمين للاختبار الآخذين اختبار الـ PAA تألفت من طلاب قدموا في المقام الأول من مدارس عليا خاصة في بورتوريكو بمستويات أعلى في التمكن من اللغة الإنكليزية من تلك المستويات الموجودة إلى حد نموذجي بين طلاب المدارس العليا في بورتوريكو. على أية حال، إنه فقط هذا النمط من المتقدم للاختبار الذي يتطلع على نحو نموذجي إلى تعليم بعد الثانوي في الولايات

المتحدة. وهكذا، مع أن نتائج الدراسة قد لا تصف العلاقة بين الدرجات النهائية المحصّلة في الـ PAA والـ SATI من أجل كل طلاب المدارس العليا في بورتوريكو، ربما تكون النتائج صالحة تماماً من أجل عدد منتقى من الطلاب الذين يعزمون على إكمال دراستهم في الولايات المتحدة.

أشار شميت و آل (1998) إلى أن التعميم إلى مجموعات أخرى أبعد من تلك الممثلة في العينة (تشمل مجموعات تأخذ بعين اختبار الـ PAA في بلدان أمريكا اللاتينية، أخرى غير بورتوريكو)، ربما لا يكون ملائماً نظراً لأن العلاقة بين SATI و PAA و ESLAT ربما تختلف في تلك البلدان الأخرى.

عائق إضافي لعلم المنهج الممثل بهذه الدراسة هو أنه لم ينجم عنه جدول اتفاق يسمح بمقارنات مباشرة لمجموعات فرعية من طلاب يأخذون الـ PAA مع مجموعات فرعية لطلاب يأخذون الـ SATI كنتيجة لهذه الدراسة، تم تزويد مستخدمي الدرجات النهائية بجدول يتطلب إدخالاً مع الدرجة النهائية PAA و ESLAT كلتيهما والدرجة النهائية للقراءة SATI المتنبأ بها من النص المطبوع للجدول. بناء عليه كان الربح في صحة تنبؤات الدرجة النهائية يتوازن مع خسارة للقابلية العملية أو الملائمة لمستخدم الدرجة النهائية.

نظراً لأن تطبيق علم المنهج المستخدم في تطوير العلاقة بين الدرجات النهائية PAA و SATI في الدراسة الثالثة لم تظهر نتائجه في جدول الاتفاق، ليس بالإمكان أن نقارن نتائج هذه الدراسة مع تلك المحصّلة في الدراستين السابقتين المقامة من قبل آنغوف ومودو (1973) وآنغوف وكوك (1988). بالرغم من أنه تجدر الإشارة إلى أنه نظراً لأن كلا الاختبارين PAA و (SAT) قد تم تعديله إلى حد بعيد منذ أن تم إكمال الدراستين السابقتين، قد تكون مقارنة النتائج عبر الدراسات الثلاث موضع تساؤل، حتى إذا أيّد علم المنهج المستخدم لربط الاختبارات في الدراسة الثالثة، تطوير جدول الاتفاق.

مناقشة

بالإمكان تعلم عدد من الدروس المهمة من تقييم العمل الذي تم إجراؤه عبر العشرين سنة الماضية والذي ركّز على ربط الدرجات النهائية المحصّلة في PAA بدرجات نهائية محصّلة في الـ SAT. حاولت كل من الدراسات الثلاث التي تجري مناقشتها هنا تحسين نتائج الدراسات السابقة بتطبيق التفكير الأكثر شيوعاً في نظرية القياس السيكولوجي والتطورات التقنية الأقرب حداثة. مع ذلك عرضت حتى الدراسة الأقرب حداثة، التي أجراها شميت وآل، عدداً من العوائق الجدية. بالتأكيد أن الخبرات المكتسبة من دراسات ربط PAA/SAT الثلاث تعرض بقوة كم يكون صعباً الحصول على درجات نهائية صحيحة وقابلة للمقارنة من اختبارات تم إعطاؤها إلى مجموعات تختلف في اللغة والثقافة.

من المحتمل أن التقدم الأكثر أهمية في دراسة آغنوف - مودو (آنغوف ومودو 1953) كان تطبيق تقنية الرسم البياني المثلى لأجل كشف مفردات في منظومة تعادل المفردة "العامة" التي لم تسلك طريقة متشابهة في المجموعات الناطقة بالإسبانية والناطقة بالإنكليزية. استخدم آغنوف ومودو هذه التقنية الجديدة، التي كانت قد تم تطويرها لغربلة المفردات من أجل انحياز عرقي آنغوف وفورد، (1973) تتضمن سير العملية تعيين قيم صعوبة المفردة (المثلثات) من أجل مفردات جرى تقريرها على المجموعتين موضع الاهتمام، وشطب تلك المفردات التي تقع بعيداً عن المحور الرئيس للقطع الناقص المشكل بالرسم البياني. أدرك آنغوف ومودو بوقت مبكر أن مأزقاً جديداً واحداً في دراسات حضارية متداخلة/لغوية متداخلة كان على الرغم من الترجمة الأكثر شدة في التدقيق والترجمة المعادة، يجب أن يتم إحصائياً غربلة المفردات التي يتوقع أن تتصرف بصورة مشابهة (مثل مفردات "عامة" في تصميم اختبار معتمد، على الأخص تلك المفردات التي لديها عنصر لغوي/ لفظي متين). يظهر بحث أخير لـ ميونز، هامبلتون، وإكسنغ (2001) أيضاً رسوماً بيانية مثلثية لا تزال تستطيع أن تكون مفيدة في كشف المفردات الإشكالية ولو بأحجام عينات صغيرة.

إن الدراسة التي قام بها آنغوف وكوك (1988) مبنية على تصميم دراسة سابقة مع بعض التحسينات التقنية والمنهجية. أمل المؤلفون في أن استخدام إجراءات IRT، لتحل مكان استخدام إجراءات نظرية الاختبار الكلاسيكية، التي كانت تستخدم في الدراسة الأولى، ستقدم نتائج محسنة. في الحقيقة، أثبتت إجراءات IRT لاكتشاف DIF انظر لورد، 1980 كونها إجراءات قوية جداً من أجل غربلة المفردات العامة. كان المؤلفون واثقين جداً من أنه في الوقت الذي يكونون فيه قد أكملوا غربلة المفردات سيكونون قادرين على تشييد اختبار "عام" يمكن استخدامه لربط الأغراض دون خطر تمييز أية مجموعة مركز الاهتمام. على أية حال، بدأ المؤلفون يشكّون في الافتراضات الضمنية للعمل الذي يقومون به. هل كان الاختباران PAA و (SAT) يقيسان تركيبات بنيوية متشابهة بدرجة كافية لتأييد تطوير جدول الاتفاق؟ ماذا كان يعني استخدام درجة نهائية في اختبار PAA لطالب كان يتكلم الإسبانية فقط لتقدير درجة نهائية للطالب في القسم اللفظي من اختبار الـ SAT؟

كنتيجة للأمور المقلقة المثارة من قبل مؤلفي دراسة الربط SAT/PAA الثانية، جرت مراجعة علم المنهج المستخدم في الدراسة الثالثة بشكل كلي. استخدم مؤلفو الدراسة الثالثة شميت وآل، (1998) إجراءات الارتداد لتطوير العلاقة بين الدرجات النهائية PAA و SAT في مجرى تحليل المعطيات للدراسة الثالثة، وجدوا أن الارتباطات بين الدرجات النهائية المحصّلة في الاختبارات PAA اللفظية و SATI اللفظية (لأجل العينة ثنائية اللغة المستخدمة في الدراسة) كانت فقط أعلى قليلاً من الارتباطات بين الدرجات النهائية المحصّلة في الاختبارات PAA اللفظية و PAA الرياضيات بالرغم من أنه، كما جرت الإشارة سابقاً، توجد تقنيات إحصائية يمكن استخدامها لتطوير جداول الاتفاق عندما لا تقيس الاختبارات الشيء نفسه (و في الحقيقة، العمل الذي تم إجراؤه في الدراسة الثانية هو مثال ممتاز لهذا النمط من العمل) يبقى السؤال، كيف يفسّر المرء النتائج من التطبيق لجدول اتفاق مطور تحت تلك الظروف؟

اختار شميت وآل (1998) تطوير معادلات التنبؤ من أجل التنبؤ بالدرجات النهائية SAT من الدرجات النهائية PAA أخذت المعادلات في الحسبان ليس فقط المقدرة اللفظية أو الرياضية للمتقدم للاختبار، كما تم قياسها بـ PAA، لكن أيضاً اعتبرت المقدرة باللغة الإنكليزية للمتقدم للاختبار، كما تم قياسها بـESLAT بالرغم من أن معادلات التنبؤ غير صالحة للاستخدام ولا يمكن استخدامها بجاهزية في المقارنة لمجموعات متقدمين للاختبار، فإنها بالتأكيد تزود بإجابة أكثر دقة على السؤال كيف لطالب يحصل على درجات نهائية بمستوى معين في الـPAA أن يحصل على درجات نهائية في الـ SATI.

إن الأسئلة التي تبقى ليجري اكتشافها عند اعتبار ربط PAA و SATI هي: الوصف للتشابه أو الفروق بين التركيبات البنيوية المقاسة من الـ PAA و SATI وكيف تأثرت هذه التشابهات أو الفروق بالمقدرة في اللغة الإنكليزية. أضف إلى ذلك أنه من المهم أن نبقي في الذهن أن الكليات لا تهتم كثيراً بتنبؤات الدرجات النهائية SATI من درجة نهائية PAA بقدر ما تهتم في اتخاذ قرارات صالحة حول كم سيكون الطلاب ناجحين إذا تم قبولهم في كلية معينة. إن العلاقة بين الدرجات النهائية PAA وSATI، والتي جرى تطويرها في الدراسة SATI/PAA الثالثة، لا تتطلب مصادقة رسمية باختبار العلاقة بين الدرجات النهائية SATI المتنبأ بها والأداء في الكلية، كما جرى في "متوسط درجة طالب الصف الأول الجامعي"، أو في معيار آخر ما ذو أهمية.

من المهم أن نبقي النتائج والدروس المتعلّمة من الدراسات SAT/PAA الثلاث في أذهاننا عند مراجعة عمل مشاريع أخرى لتكييف الاختبار. ليس تكييف الاختبارات عملية تافهة، بل إنها عملية مهمة جداً تؤثر بدرجة عظيمة على صحة الدرجات النهائية للاختبار. استناداً إلى الاستخدام اللا متناه للدرجات النهائية للاختبار، يمكن أن يكون من الأهمية بمكان للطلاب، والموظفين، وسكان بلدان أخرى

أن يجري إعطاؤهم الفرصة لعرض مهاراتهم وقدراتهم في اختبارات يتم إعطاؤها بلغتهم الأصلية. درسان هامان تم تعلمهما من الدراسات SAT/PAA الثلاث يشيران إلى أنه:

1- من المهم أن نأخذ بالحسبان كيف سيجري استخدام وتفسير الدرجات النهائية للاختبار. سيعتمد بدرجة كبيرة تصميم دراسة الربط والنماذج المختارة على الاستخدامات الممكنة وتفسيرات الدرجات النهائية للاختبار.

2- لا توجد طريقة بسيطة للقيام بوظيفة ذات نوعية عالية لتكييف الاختبارات من أجل لغات وثقافات مختلفة. الاختبارات المكيّفة هي عملية مجهدة تتطلب تنفيذاً حذراً، ليس فقط انتباهاً دقيقاً فقط للعملية التطويرية للاختبار، بل يكون الانتباه لعملية إدارة الاختبار وتفسير الدرجات النهائية على قدر مساو من الأهمية.

لحسن الحظ يجري القيام بعمل جيد في موضوع تكييف الاختبار وبالاهتمام المتزايد في المجال. ستكون الحلول لما قد يبدو مشكلات متفاعلة على الأغلب متوفرة لدينا في المستقبل.

شكر

يقدر المؤلفون الإسهامات لـ أنتوني ماغرينا، فيل دورانز، ودانيل إيغنور في الإعداد لهذا الفصل.

*****

المراجع

Alderman, D. L. (1981). *Language proficiency as a moderator variable in testing academic aptitude* (TOEFL Research Rep. No. 10, RR81-41). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

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

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–106.

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

Boldt, R. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade County high school volunteers* (Statistical Rep. No. 69-31). Princeton, NJ: Educational Testing Service.

The College Board. (1995). *Cambiod en el examen de admision del College Board: La nueva PAA* [Changes in the College Board admissions tests: The new PAA]. San Juan: Oficina de Puerto Rico y de Actividades Latinamericanas.

Cook, L. L. (1995, April). *Lessons learned: Implementing change in the SAT.* Paper presented at the meeting of the National Council on Educational Measurement, San Francisco.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1989, April). *Equating achievement tests using samples matched on ability.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3,* 37–55.

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

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

Hambleton, R. K. (1996, April). *Guidelines for adapting educational and psychological tests.* Paper presented at the meeting of the National Council on Educational Measurement, New York.

Kolen, M. J. (1990). Does matching in an equating work? A discussion. *Applied Measurement in Education, 3,* 97–104.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. 23). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73–95.

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

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

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

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

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737–756.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Schmitt, A. P., Dorans, N. J., Magrina, A. & Cook, L. L. (1998). *Predicting scores on the English Language SAT from the Spanish Language PAA and the Spanish Language English as a Second Language Achievement Test.* Paper presented at the meeting of the American Educational Research Association, San Diego.

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

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

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3,* 105–113.

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.

van de Vijver, F. J. R. & 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.