مجلة الروائز
ISSN: 2602-5914/ EISSN: 2716-9596
المجلد:07 العدد: 01 (السنة: 2023) ص ص: [08-20]

مصادر تباين الخطأ المؤثرة على ثبات وصدق تقييمات الكفاءة وفق نظرية إمكانية التعميم: مراجعة للبحوث

Sources of error variances affecting reliability and validity of competency assessments according to Generalizability Theory: A review of research

فاروق طباع *

جامعة محمد لمين دباغين –سطيف2 - ftebbaa05@yahoo.fr

عبد السلام دعيدش

جامعة محمد لمين دباغين –سطيف2 - daideche_ed@yahoo.fr

تاريخ الإرسال20/08/2022 : تاريخ القبول: 2023/01/01

ملخص:

تقييمات الكفاءة عبارة عن وضعيات معقّدة لتقييم أداء الطلاب باستخدام المهمات والمصحّحين والفترات وطرق التقييم، وهذا ما يؤثر على درجات الطلاب ويؤدي إلى خفض موثوقيتها، وقد كشفت البحوث بأن ثبات وصدق تقييمات الكفاءة يتأثران اعتماداً على نظرية إمكانية التعميم بمصادر تباين الخطأ الراجعة بالأساس إلى المصحّحين، والمهمّات، وتفاعل الطّلاب مع المهمّات، والفترات، وطرق التقييم. يتناول هذا المقال مراجعة للأدبيات البحثية التي استخدمت نظرية إمكانية التعميم في فحص مصادر تباين الخطأ المؤثّرة على ثبات وصدق تقييمات الكفاءة، وتقديم طرق لخفض تأثير كل مصدر من مصادر تباين الخطأ على موثوقية الدرجات.

الكلمات المفتاحية: تقييمات الكفاءة؛ مصادر تباين الخطأ؛ نظرية إمكانية التعميم؛ الثبات؛ الصدق.

Abstract:

Competencies assessments are complex situations for assessing students' performance using tasks, raters, occasions and assessment methods. This can affect student scores and decrease their generalizability. Based on Generalizability theory, research showed that reliability and validity of competency assessments are influenced by sources of error especially due to raters, tasks, person-task interaction, occasions and assessment methods. This article deals with a review of research literature used Generalizability theory in examining the sources of error that affect reliability and validity of competency assessments, and provide the ways to reduce each source of error variance on dependability of scores.

Keywords: Competency assessments; generalizability theory; reliability; sources of error variance; validity.

* المؤلف المرسل

مقدمة:

انتقدت التقييمات التقليدية القائمة على الاختيار من متعدد، والصواب والخطأ والاجابات القصيرة، وغيرها بسبب اعتمادها على اختيار الاجابة، والإجابات القصيرة، وسرعة الإجابة عليها، والتدريس من أجل الاختبار، وقياسها المعارف والمهارات المنعزلة للطلاب (Linn, Baker, & Dunbar, 1991)، وتركيزها بالأساس على قياس العمليات المعرفية الدنيا، واختلاف النواتج التعليمية المرغوبة من طرف المعلم، والتقييم المباشر القائم على المعارف دون الكفاءات (علام، 2004).

رغم هذه الانتقادات إلا أنها تتميز بدرجة عالية من الصدق والثبات بسبب أحادية أبعادها، وشموليتها للمحتوى الدراسي، ومعاينتها الكافية للبنود، لذلك لقيت تقييمات الكفاءة اهتماما كبيرا خلال العقدين الماضيين، والتي تنادي بضرورة أن تعكس التقييمات الأنشطة التعليمية، وأن تكون أكثر تحفيزا وتشجيعا للمهارات المعقدة للطلاب، وأكثر تركيزا على مهمات ذات معنى وأكثر استدلالا على الكفاءة.

تتميّز تقييمات الكفاءة بتعدد أبعادها، وإعطائها حرية للطلاب للإجابة على المهمات المعقدة، وواقعية سياقاتها، واستدعائها مهارات التفكير العليا، ومطالبتها بإصدار أحكام أثناء عمليات التصحيح، وهذه الأبعاد أدّت إلى إعادة النظر في تقدير صدق وثبات تقييمات الكفاءة حتى تتلاءم مع معايير شمولية المحتوى، وإمكانية التعميم، والواقعية، وتعقيد العمليات المعرفية (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006; Johnson, Penny, & Gordan, 2009).

وفي هذا الاطار أصبح تقدير صدق وثبات تقييمات الكفاءة بواسطة طرق النظرية الكلاسيكية غير ملائمة وغير كافية (De Ketele & Gerard, 2005)، لأن الطرق الكلاسيكية غير قادرة على ضبط المصادر المتعددة للخطأ التي تؤثر على ثبات وحتى صدق أدوات تقييم الكفاءات. كما أكد (Baartman, Prins, Kirschner, & van der Vleuten, 2007) بأن وجهات النظر التقليدية للثبات لا يمكن تطبيقها كما هي في تقييمات الكفاءة رغم أن فكرة الثبات مهمة إلا أنّها تحتاج إلى إعادة النظر، لذا من الضروري استخدام طرق ثبات تختلف عن الطرق التقليدية. وذلك على اعتبار أن بعض أساليب التقييم الجديدة تستخدم طرق تقليدية غير ملائمة لتقدير خطأ القياس لأن التقييم الحديث يتطلب اختبارات محكية المرجع ذات مهمات مفتوحة ومعقّدة، لذلك هناك حاجة ملحة للبحث عن قياسات أخرى جديدة لضمان إصدار أحكام متسقة لأداء الطلاب (Cronbach, Linn, Brennan, & Haertel, 1997).

يتطلب التحقق من جودة تقييمات الكفاءة أساليب ملائمة لأنها تتصف من الناحية العملية بأداء معقد يتم معاينته في وضعية خاصة، وفي فترة معينة، ويتم تقييمه من طرف مصحّح معين يتصرّف كمحكّم (Scallon, 2004) وبواسطة استخدام طرق تقييم مختلفة (Shavelson, Baxter, & Gao, 1993)، ونتيجة لذلك تعكس عملية تقدير الثبات والصدق أبعاد متعددة يتم قياسها ضمن نطاق مركب من المهمات، والمصحّحين، والفترات، وطرق التقييم.

ومما سبق يتضح أنه من الضروري الرجوع إلى نظرية إمكانية التعميم لتقدير صدق وثبات أدوات تقييم الكفاءة (طباع و ليفة، 2015) لأن تقييم أداء الطالب يتم ضمن نطاق واسع يتكوّن من مهمات ومصحّحين وفترات وطرق تقييم، بالإضافة إلى التفاعلات التي تحدث بين هذه الأبعاد.

أجريت العديد من الدراسات لتقدير صدق وثبات تقييمات الكفاءة باستخدام نظرية إمكانية التعميم، وتوصلت إلى انخفاض مستويات ثباتها، وقد شمل انخفاض الثبات مختلف المستويات والمجالات الدراسية، في العلوم (Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson et al., 1993)، وفي الرياضيات (Hébert, Valois, Scallon, & Frenette, 2014; Klein, McCaffrey, Stecher, & Koretz, 1995; McBee & Barnes, 1998; Nie, Yeo, & Lau, 2007; Taylor & Pastor, 2013)، وفي اللغات (Chen, Niemi, Wang, & Mirocha, 2007; Gao & Brennan, 2001; Gebril, 2009)، كما تناول العديد من المؤلفين مشكلة ضعف ثبات تقييمات في مقالات منشورة في دوريات القياس التربوي (Brennan & Johnson, 1995; Linn & Burton, 1994; Parkes, 2001).

وبما أن الاتساق بين المهمات والاتساق بين طرق التقييم عنصران مهمان في تقدير الصدق بالتحقق من أهمية معاينة المحتوى وطرق التقييم المستخدمة في تطوير قياسات متعددة للكفاءة، فانّ نظرية إمكانية التعميم استخدمت في تأييد الصدق، وتوصلت بعض الدراسات أيضا إلى انخفاض في أدلتها (طباع، 2016؛ Huang, 2012 ; Shavelson et al., 1993).

ومن الأسباب التي يمكن أن تؤدي إلى انخفاض مستويات ثبات وصدق تقييمات الكفاءة تأثّرها بمصادر متعددة للخطأ، والتي ترجع إلى المصحّحين أو المهمات أو الفترات أو طرق التقييم أو إلى التفاعلات بين الأبعاد فيما بينها، وبين هذه الأبعاد والطلاب. وللتغلّب على انخفاض مستويات الثبات تم اقتراح طرق لخفض مصادر الخطأ ورفع إمكانية التعميم. وفي هذا الاطار يهدف هذا المقال إلى تحليل نتائج بحوث تقدير صدق وثبات تقييمات الكفاءة وفق نظرية إمكانية التعميم بالتركيز على مصادر تباين الخطأ، وطرق التغلب على مشكلة انخفاض ثباتها وصدقها.

1. مساهمات نظرية إمكانية التعميم:

في تقييمات الكفاءة يتحصّل الفاحص على درجة لأداء الطالب مستمدة من نطاق معقّد من المهمات والمصحّحين والفترات وطرق التقييم، حيث تُطبق على عينة من الطلاب مجموعة من المهمات خلال فترات مختلفة، وتصاغ هذه المهمات بطرق متنوعة، ويتم تصحيح أداء الطلاب بواسطة عينة من المصحّحين.

ويمكن اعتبار درجة الطالب من بين الدرجات التي يمكن أن يحصل عليها من أي من المهمات والطرق والمصحّحين والفترات، فيمكن أن يحصل على درجة شاملة بمعنى متوسط الدرجات التي يحصل عليها إذا أنجز كل المهمات الممكنة، والمقدرين الذين يمكن الاعتماد عليهم في تصحيح إجابات الطلاب، والفترات التي يمكن أن تُجرى فيها مهمات الاختبار، وطرق التقييم الممكنة التي يؤديها الطالب (Shavelson et al., 1993)

فاختيار المهمات والمصحّحين والفترات قد يؤدي إلى تغيّر درجات الطالب من مهمة إلى أخرى، ومن مصحّح إلى آخر، ومن فترة إلى أخرى، ومن طريقة تقييم إلى أخرى، وعندما يختلف الأداء من عينة إلى أخرى فان تغيّر اختيار هذه الأبعاد يؤدي إلى خطأ القياس، ومن ثم فان القياس الذي يمكن تعميمه قد يختلف من عينة لأخرى.

وفي ضوء هذه الأبعاد المعقّدة التي تتميز بها وضعيات القياس تتلاءم نظرية إمكانية التعميم مع وضعيات تقييم الكفاءة (Hébert & Duclos, 2007; Bain, 2014) والتي تسمح بإيجاد الدرجة التي يمكن أن تُعمم بها درجات تقييم

الطلاب مع نطاق المهمات والمصحّحين والفترات وأساليب التقييم، ويمكن أن تعالج أبعاد متعددة في وضعية القياس في الوقت نفسه (Cardinet & Tourneur, 1985).

تقدّم نظرية إمكانية التعميم في سياق تقييم الكفاءة إطارا فكريا وإحصائيا لفحص تأثير مصادر تباين الخطأ على دقة درجات الطلاب (Cardinet, Sandra, & Pini, 2010; Brennan, 2001; Shavelson & Webb, 1991)، فنظرية إمكانية التعميم ليست نموذجا بسيطا يزوّد الفاحص بمصادر الخطأ الرئيسية التي تؤثر على درجات الكفاءة (مثلا: المقدرين والمهمات والفترات)، ولكنها تقدم تقديرات إحصائية لحجم تغير المعاينة الراجعة إلى المصادر المختلفة والتفاعلات فيما بينها، ويمكن استخدام هذه المعلومات لتصميم قياسات بديلة فعالة من حيث التكلفة. بالإضافة إلى أنها تقدم معاملات مختصرة (معاملات إمكانية التعميم ومعاملات الموثوقية) التي تعبّر عن ثبات التفسيرات المعيارية المرجع والتفسيرات المحكية المرجع (Briesch, Swaminathan, Welsh, & Chafouleas, 2014).

ومن الناحية الاحصائية تستخدم نظرية إمكانية التعميم تحليل التباين من أجل تجزئة التباين بين الدرجات إلى مصادر متعددة وإلى تفاعلاتها، وفي مجال تقييمات الكفاءة يتم تجزئة التباين إلى المصادر الراجعة إلى المهمات والمصحّحين والفترات وطرق التقييم، ومصادر راجعة إلى التفاعل بين هذه الأبعاد فيما بينها. ويمكن وفقا لنظرية إمكانية التعميم أن تؤثر مصادر التباين على تقديرات الثبات، والتي لفتت انتباه العديد من الباحثين منذ بداية التسعينات، وبداية الألفية الجديدة (Brennan, 2000; Brennan & Johnson, 1995; Miller & Linn, 2000; Ruiz-Primo et al., 1993)، كما يمكن أن يتأثر الصدق وفقا لنظرية إمكانية التعميم بأساليب التقييم المستخدمة لأن درجات الطلاب يمكن أن تتغير من طريقة إلى أخرى أو إلى تفاعلات الطريقة مع الطلاب والمهمات.

يتم التركيز في بمزيد من التفصيل على عرض نتائج الدراسات التي اهتمت بتقدير مصادر تباين الخطأ المؤثرة على ثبات وحتى صدق درجات تقييمات الكفاءة وفقا لنظرية إمكانية التعميم، وذلك بتحليل تأثيرات كل مصدر من مصادر تباين الخطأ على إمكانية تعميم درجات الطلاب.

2. مصادر تباين تقييمات الكفاءة

1.2 مصدر تباين المصححين:

تختلف درجات الطلاب في المهمات باختلاف المصحّحين الذين يقيّمون أداءهم، ويمكن أن تعود تباينات المصحّحين إلى الذاتية التي تنتج من ميل بعض المصححين إلى الصرامة وميل الآخرين إلى التساهل أثناء معاينة انجازات الطلاب، أو تعود إلى أثر الهالة لاختلاف فترات التصحيح أو تتأثر بشبكات أو معايير التصحيح غير المحددة.

منذ فترة طويلة طُرحت مسألة الاتفاق بين المصحّحين كمصدر مهم من مصادر تباين الخطأ، فقد أشار (Clausser, Clyman, & Swanson, 1999) إلى أن تباين المصحّحين لا زال موضوعا جديرا بالاهتمام خاصة إذا امتزج بمصادر أخرى كفترة التصحيح ولجنة المصحّحين، كما خلصت مراجعة (Parkes, 2001) إلى أن تقديرات المصحّحين مشكلة لا زالت تلقى اهتماما متزايدا من طرف المهتمين وصانعي القرارات.

بالرغم أن الاتفاق بين المصحّحين مشكلة تتطلب العناية الشديدة إلا أنّها في الواقع ليست مشكلة حقيقية تواجه استخدام تقييمات الكفاءة (Linn & Burton, 1994)، ولكن رغم ذلك توصلت العديد من الدراسات إلى وجود اختلافات حول اتساق تصحيحات المقيّمين، فقد توصلت مراجعة (Dunbar, Koretz, & Hoover, 1991) لتسع دراسات إلى معاملات اتفاق في مجال الكتابة تراوحت بين (0.33) و(0.91)، وتوصلت إلى معاملات ثبات منخفضة تراوحت بين (0,26) و(0,60) مقارنة بالتصحيحات بين المقيّمين. وأشارت دراسة (Klein et al., 1995) إلى أن الاتفاق بين المصحّحين في ملفات الانجاز ضعيف، ويعود ذلك إلى الاختلافات المنتظمة بين المصحّحين في تفسير وتطبيق شبكات التصحيح، وطبيعة هذه الشبكات، وعدم تقنين المهمات عبر ملفات الانجاز المصحّحة.

وبالرغم من ذلك، أظهرت العديد من الدراسات التي تم مراجعتها انخفاضا أو انعداما في تباين المصحّحين، وانخفاضا في تفاعل المصحّحين مع الطلاب أو مع المهمات أو الفترات (Gao & Brennan, 2001; Gao, Shavelson, & Baxter, 1994; Güler & Gelbal, 2010; Lane, Liu, Ankenmann, & Stone, 1996; Lee & Kantor, 2007; McBee & Barnes, 1998; Taylor & Pastor, 2013) ويعود ذلك بالأساس إلى الاهتمام بالتخطيط الجيد لشبكات التصحيح، وتدريب المصحّحين، والعناية الشديدة بعملية التصحيح (Gipps, 1994).

2.2 مصدر تباين المهمات:

خلصت مراجعة (Linn & Burton, 1994) إلى أن الاتساق بين المصحّحين ليس مشكلة بالمقارنة مع الاتساق بين المهمة، حيث يمكن أن يكون تباين المهمات مشكلة لأن بعض المهمات أكثر صعوبة لدى بعض الطلاب، في حين لا تكون مهمات أخرى كذلك لدى نفس الطلاب، لذلك اهتم الباحثون بمصدر تباين المهمات باعتبار أن الاعتماد على مهمات متعددة للاستدلال على كفاءة الطلاب ضروري، وأن استخدام عدد محدود من المهمات غير كافي.

طُرحت مشكلة تغير معاينة المهمة في تقييمات الكفاءة، حيث أثبت بعض الدراسات ارتفاع في تباين الخطأ الراجع إلى المهمات(Huang, 2009; Shavelson et al., 1993; Taylor & Pastor, 2013) كما أظهرت نتائج (Huang, 2009) بأن نسبة لا يستهان بها من مقدار تباين المهمة بلغت (12%) من مجموع الدراسات التي قام بمراجعتها، وتوصلت كذلك مراجعة حديثة أُجرت من طرف (In'nami & Koizumi, 2015) لـ (36) دراسة في مجال الحديث والكتابة في الانجليزية كلغة ثانية بأن تأثيرات المهمة أكبر من تأثيرات المصحّح.

تؤدي مشكلة ارتفاع تباين المهمة إلى انخفاض في اتساق التقييمات، حيث جاءت نسبته في أغلب الدراسات التي راجعها الباحث كثاني أو ثالث أكبر مصدر بعد تفاعل الطالب مع المهمة، وتفاعل الطالب مع المهمة والفترة، فقد توصل (Shavelson et al., 1993) إلى نسبة بلغت (16%) من تباين المهمة، وتوصل (Lane et al., 1996) إلى نسب تباين المهمة تراوحت بين (7 و17%) في صيغ المهمات المقدمة للطلاب.

ومن الأسباب التي تؤدي إلى ارتفاع تباين المهمة انخفاض عدد المهمات المستخدمة في التقييم، وانخفاض في درجة تماثلها من حيث الخصائص، ومستوى التعقيد، ومستوى الصعوبة، حيث أثبت دراسة (McBee & Barnes, 1998) أن استخدام المهمات الأكثر تجانسا وتكافؤا يؤدي إلى خفض مصادر تباين المهمة وتفاعلاتها مع الأبعاد الأخرى.

ومن جهة أخرى توصلت دراسات أخرى إلى انخفاض تباين المهمة (Chen et al., 2007; Gebril, 2009; Güler & Gelbal, 2010; Nie et al., , 2007; Shavelson et al., 1993). ويمكن أن يرجع ذلك إلى محدودية عدد المهمات المستخدمة في هذه البحوث مما ساعد الباحثين في التحكم في إعدادها بطريقة متكافئة، وفي الواقع لم يكن تباين المهمة مصدرا ذات تأثير كبير على الثبات بالمقارنة مع نسبة تأثير تفاعل الطالب مع المهمة الذي سوف يتم مناقشته.

3.2 مصدر تباين تفاعل الطلاب مع المهمات:

بالمقارنة مع تباين المصحّحين والمهمات اعتبر تباين تفاعل الطالب مع المهمة مشكلة في تقييمات الكفاءة، وقد أرجع المؤلفون انخفاض الثبات إلى ارتفاع في تفاعل الطالب مع المهمة، واعتبر نقطة ضعف في تقييمات الكفاءة (Shavelson, Ruiz-Primo, & Wiley, 1999) لأنّه يعتبر عاملا مؤثرا يؤدي إلى خفض إمكانية تعميم الدرجات، وخاصية غير مرغوبة في تقييمات الكفاءة.

يعود ارتفاع تباين تفاعل الطالب مع المهمة إلى اختلاف أداء الطلاب من مهمة إلى أخرى، فإذا تحصل طالب على درجة مرتفعة في مهمة معينة فانه يمكن أن يحصل على درجة منخفضة في مهمة أخرى، ويكون أداؤه مستقلا عن المصحّح أو فترة الانجاز، فمنذ التسعينات كشف (Shavelson, Baxter, & Pine, 1992) عن حصول بعض الطلاب على درجات عالية في مهمة معينة ودرجات منخفضة في مهمة أخرى.

توصلت معظم الدراسات إلى ارتفاع تفاعل الطالب مع المهمة، حيث بيّنت مراجعة (Huang, 2009) أن حوالي (26%) من التباين راجعة إلى تفاعل الفرد مع المهمة، والذي ساهم بشكل واضح في خفض معاملات إمكانية التعميم، مما أدى إلى مشكلة في عملية تطوير تقييمات متسقة، كما أثبتت مراجعة (In'nami & Koizumi, 2015) بأن تأثيرات تفاعل الطالب مع المهمة أكبر من تأثيرات تفاعل الطالب مع المصحّح، وكانت تأثيرات تفاعل الطالب مع المهمة كبيرة بالمقارنة مع المهمات أو المصحّحين.

أثبتت الدراسات السابقة التي اطُّلع عليها أن مصادر التباين الأكثر تأثيرا على إمكانية تعميم تقييمات الكفاءة راجعة إلى تفاعل الطالب مع المهمة في مختلف تصميمات البحث سواء في مجال الرياضيات (Güler & Gelbal, 2010; Hébert et al., 2014; Klein et al., 1995; McBee & Barnes, 1998; Nie et al., 2007; Shavelson et al., 1993; Taylor & Pastor, 2013) أو في مجال العلوم (Gao et al., 1994; Shavelson et al., 1992; Shavelson et al., 1999; Webb, Schlackman, & Sugrue, 2000) أو في مجال اللغات (Gao & Brennan, 2001; Taylor & Pastor, 2013) حيث تتفق نتائج هذه الدراسات مع مراجعة (Huang, 2009) لمختلف دراسات تقييمات الأداء، ومع مراجعة (In'nami & Koizumi, 2015) في مجال تقييمات الحديث والكتابة.

حاول بعض الباحثين مناقشة مشكلة ارتفاع مصدر تباين تفاعل الطالب مع المهمة، فقد أفرد كلا من الباحثيْن (Huang, 2009; Parkes, 2001) جزءا هاما من تحليلاتهما في تناول هذه المسألة التي لا زالت لم تجد لها الحل النهائي، حيث تتطلب تأملات نظرية ودراسات ميدانية للإجابة عن التساؤل الهام حول ما إذا كان تفاعل الطالب مع المهمة مشكلة موجودة فعلا أم خرافة؟

وبالنظر إلى ارتفاع تفاعل الطالب مع المهمة الذي يؤثر على إمكانية تعميم تقييمات الكفاءة حاول بعض الباحثين معالجة المشكلة باستخدام طرق تجريبية معتمدة على خرائط المفاهيم لخفض تباين تفاعل الطالب مع المهمة (Parkes, Zimmaro, Zappe, & Suen, 2000)

وبناء عليه تمثل مشكلة ارتفاع تفاعل الطالب مع المهمة أحد العبقات التي تحدّ من ثبات تقييمات الكفاءة، وتعتبر من العوامل التي تساهم في خفض إمكانية التعميم، لذا اعتبر ارتفاع تفاعل الطالب مع المهمة مشكلة حقيقية وليست خرافة تحتاج إلى مزيد من البحث.

4.2 مصدر تباين الفترات:

يمكن أن تتأثر درجات الطلاب بالفترة التي أنجزت فيها التقييمات، ولكن استثنيت الفترة من مصادر تباين تقييمات الكفاءة ولم تلق اهتمامات بحثية مقارنة بمصادر تباين المهمة والمصحّح، فقد أشار (Brennan, 2000) إلى عدد قليل من الدراسات التي أدمجت الفترة في أبعاد وضعية القياس، يمكن أن تكشف عن إمكانية التعميم في فترة واحدة على فترات مختلفة. فبالرغم من إمكانية ارتفاع تغير معاينة الفترة إلا أن عددا ضئيلا من الأدبيات المتوفرة في تقييمات الكفاءة عن استقرار الدرجات التي توضّح الأهمية النسبية لاختلاف الفترة (Webb et al., 2000).

وتجدر الاشارة إلى أن مسألة تغير معاينة الفترة لم تلق اهتمامات بحثية، وحتى إن لقيت لم تعالج بطريقة مناسبة، فحسب (Cronbach et al., 1997) تفسير تغير معاينة الفترة كمصدر رئيسي للخطأ في بعض الدراسات مسألة معقدة، حيث يمكن أن تكون الفترات مصدرا خفيا للتباين، فإنجاز الطلاب للمهمات في فترة واحدة لا يمكن التعرف ما إذا كان الاختلاف راجع إلى التغير بين المهمات أو إلى التغير بين الفترات أو كلاهما.

فإجراء التقييم في فترة واحدة يكون تغير معاينة المهمة ممزوجا بتغير معاينة الفترة، فلا يمكن التعرف على ما إذا كانت الفروق في أداء المهمات في فترة واحدة قابلة أن تتكرر في فترات أخرى (Webb et al., 2000). فإذا كانت الفروق في المهمات مستقرة بين الفترات فان تغير معاينة المهمة يصبح المصدر الرئيس للخطأ، وإذا كانت الفروق في المهمات ليست مستقرة بين الفترات حينها يكون كلا من تغير معاينة المهمة وتغير معاينة الفترة مصادر مهمة لتباين الخطأ.

لذلك يؤدي امتزاج تغير معاينة الفترة مع تغير معاينة المهمة إلى ارتفاع تباين تفاعل الطالب مع المهمة ومع الفترة، فقد أكدت دراسة (McBee & Barnes, 1998) بأن تحليل البيانات خلال فترة واحدة يكون تغير معاينة المهمة مصدرا رئيسيا لخطأ القياس، في حين أن تحليل البيانات في فترات متعددة يكون مزيجا من تغير معاينة المهمة والفترة مصدرا رئيسيا للخطأ، وقد أثبتت نتائج (Shavelson et al., 1993) بأن ارتفاع تباين تفاعل الطالب مع المهمة ومع الفترة قد بلغ (59%) من مجموع مصادر التباين.

أكدت معظم تقييمات الكفاءة التي أدرجت الفترة في تصميمات القياس أن أكبر مصدر للتباين راجع إلى تفاعل الطالب مع المهمة ومع الفترة، والذي تجاوز مقدار تفاعل الطالب مع المهمة (طباع، 2016 ;McBee & Barnes, 1998 Ruiz-Primo et al., 1993; Shavelson et al., 1993; Shavelson et al., 1999; Webb et al., 2000) ورغم أن تباينات

تفاعل الطالب مع الفترة، والمهمة مع الفترة، والمصحّح مع الفترة ضئيلة إلا أن معاملات استقرار درجات الطلاب بين الفترات في بعض الدراسات منخفضة أو متوسطة (McBee & Barnes, 1998; Ruiz-Primo et al., 1993).

يعتبر مصدر تباين الفترة وتفاعله مع الأبعاد الأخرى خاصة مع المهمة (أي تفاعل الطالب مع المهمة ومع الفترة) مصدر رئيسي للتباين لا يمكن للباحث تجاهله أثناء تقدير ثبات تقييمات الكفاءة لأن أداء الطلاب للمهمات يمكن أن يتغير من فترة إلى أخرى، وفي بعض الأحيان يميل الطلاب إلى تغيير استراتيجيات انجازهم من فترة إلى أخرى.

5.2 مصدر تباين طرق التقييم:

بما أن درجات الطلاب يجب أن تكون ثابتة عبر المهمات والمصحّحين والفترات يجب أن تكون استدلالاتها صادقة كذلك عبر طرق التقييم المختلفة، لهذا حاول بعض المؤلفين تطوير أدلة جديدة للصدق قائمة على نظرية إمكانية التعميم تتفق مع تقييمات الكفاءة (Kane, 1982; Messick, 1995).

وفي هذا الشأن ساهمت نظرية إمكانية التعميم في تقديم أدلة عن صدق تقييمات الكفاءة باقتراح (Kane, 1982) نموذجا للصدق التقاربي قائما على نظرية إمكانية التعميم، حيث يعتمد على تقدير حجم التباين بين الدرجات الناتجة عن معاينة طرق التقييم، وتقديم مؤشر عن مدى تقارب طرق القياس البديلة ،حيث يعبّر تغير المعاينة الراجع إلى طريقة التقييم عن الصدق التقاربي، والذي يشير إلى ارتفاع تغير معاينة الطريقة إلى عدم تقارب طرق التقييم على نحو مشترك مقارنة بتقييمات الاختيار من متعدد (Shavelson et al., 1993).

وفي هذا الصدد أجريت دراسات حول مصادر تباين أساليب تقييم الكفاءة، وتوصلت إلى ارتفاع في تباين طرق التقييم (طباع، 2016 ; Hébert, 2006; Shavelson et al., 1993) فقد كشفت دراسة (Shavelson et al., 1993) بأن مصدر تباين طرق التقييم (الملاحظة، دفاتر الكتابة، المحاكاة بواسطة الكمبيوتر، الإجابة القصيرة) بلغ (16%) من مجموع التباينات، وبلغ تفاعل الطالب مع المهمة ومع الطريقة الممزوج بالخطأ العشوائي غير المقاس (29%). وتوصلت دراسة (Hébert et al., 2014) إلى ارتفاع في تباين طريقة بناء المهمات (درجة واقعية المهمة، وملاءمة معلوماتها، وبناء المهمة، والمجال الذي تنتمي إليه) الذي بلغ (62%) من مجموع مصادر التباين. وفي دراسة حديثة توصلت نتائج (طباع، 2016) إلى نسبة تباين بلغت (16%) ونسبة تفاعل الطالب مع المهمة ومع الصيغة الممزوج بالخطأ العشوائي بلغت حوالي (39%).

أظهرت نتائج الدراسات ارتفاع تباين طرق التقييم وتفاعل الطالب مع المهمة ومع الطريقة الممزوج بالخطأ العشوائي، ما يؤكد صعوبة الاستدلال على طرق تقييمات الكفاءة، وهذا الأمر يمكن أن يساهم في مشكلة تتعلق بضعف أدلة الصدق على غرار مشكلة ضعف الثبات بسبب التباين بين المهمات والمصحّحين والفترات وتفاعلاتها.

3. طرق خفض مصادر تباين تقييمات الكفاءة:

تؤدي مصادر تباين الخطأ الراجعة إلى المصحّح والمهمة والفترة والتفاعلات الحادثة بينها إلى خفض مستويات ثبات تقييمات الكفاءة، لذا أصبح المهتمون في مواجهة هذه المشكلة، وانصبت جهودهم على إيجاد طرق كفيلة بتحسين الثبات لاتخاذ قرارات صائبة حول أداء الطلاب، وأصبحت طرق خفض مصادر تباين الخطأ المؤثرة على الثبات من

أولويات دراسات القرار في نظرية إمكانية التعميم، والتي اقترحت عددا من الطرق تتلاءم مع مصادر الخطأ أكثر تأثيرا على وضعيات التقييم.

يتفق معظم الباحثين والمؤلفين أن استخدام شبكات تصحيح واضحة ومحكمة البناء، والعناية بتدريب المصحّحين، ومتابعة عملية التصحيح يمكن أن تخفّض تباينات الخطأ الراجعة إلى المصحّحين (Gipps, 1994; Johnson et al., 2009; Linn & Burton, 1994)، بالإضافة إلى زيادة عدد المصحّحين في دراسات القرار لبلوغ مستويات إمكانية التعميم مقبولة، ولكنها أقل فعالية من الطرق التي سوف نتناولها.

فمن أجل خفض مصدر تباين المهمة اقترحت طرق متعددة أكثرها فعالية تعتمد على زيادة عدد المهمات (Shavelson et al., 1999; Taylor & Pastor, 2013)، كما يمكن تقنين المهمات المقدمة للطلاب بإعدادها بطريقة محكمة (Parkes, 2001) وإعادة بنائها وصياغتها وفحصها بشكل أفضل (Scallon, 2004)، كما اقترح بعض المؤلفين تخفيض مجال محتوى انتماء المهمات (Dunbar et al., 1991)، والعناية باختيار مهمات محدّدة للتقييم لضمان درجات ذات موثوقية.

رغم تعدّد الطرق المقترحة لخفض تباين المهمة إلا أن زيادة عدد المهمات أكثرها استخداما وفعالية في رفع ثبات درجات الاختبارات، ولكنها يمكن أن تنتج عنها أثار سلبية مرتبطة بالتكاليف والجهود الاضافية لأن زيادة المهمات تحتاج إلى جهود إضافة في إعداد المهمات وتوفير وسائل لإعدادها وتطبيقها وتصحيحها (Parkes, 2000).

كما يمكن أن تؤدي الطرق المقترحة الأخرى إلى خفض من فاعلية التقييم، فتقنين المهمات يؤدي إلى تقييمات غير واقعية، وإعادة بناء وصياغة وفحص المهمات يؤدي إلى جهود وتكاليف معتبرة. كما يؤدي خفض مجال انتماء المهمات وإعادة بنائها إلى خفض شمولية مجال محتواها، وإلى تعميمات مضللة (Dunbar et al., 1991).

ومن جهة أخرى اقترحت زيادة الفترات لخفض تغيّر معاينة الفترة، إلا أن هذه الطريقة تؤدي إلى جهود وتكاليف إضافية لتمرير وتصحيح المهمات، وكما يمكن التغلب على مشكلة ارتفاع تفاعل الطالب مع المهمة ومع الفترة يمكن استخدام عدد واسع من المهمات و/ أو عدد الفترات (Brennan, 2000; Huang, 2009).

ومن أجل التغلب على مشكلة ارتفاع تفاعل الطالب مع المهمة اقترحت طرق متعددة اهتمت أغلبها بزيادة المهمات كاستراتيجية فعالة، كما واقترحت طرق أخرى تجريبية تعتمد على خرائط المفاهيم لزيادة انتقال أثر التعلم (Parkes, 2001; Parkes et al., 2000)

أكدت أدبيات البحث بأن طرق تحسين إجراءات القياس تساهم بشكل أو بآخر في خفض مصادر تباين الخطأ، بحيث يتطلب خفض تباين المصحّح إعداد شبكات تصحيح وإعطاء عناية كافية لتدريب المصحّحين، ويتطلب خفض تباين المهمة وتفاعل الطالب مع المهمة زيادة عدد المهمات، ويتطلب خفض مصدر تباين الفترة زيادة عدد الفترات، في حين أن الطرق أكثر فعالية تعتمد على إدماج أكبر عدد ممكن من الأبعاد واستخدام تصميمات متقاطعة بدلا من تصميمات متداخلة وإدماج الفترة في أبعاد وضعية القياس.

يتضح مما سبق أن توسيع عدد مستويات أبعاد القياس أحد الحلول الملائمة لبلوغ مستويات مقبولة من إمكانية التعميم، ورغم فعالية الطرق المقترحة وعزيمة الباحثين للتغلب على مشكلات ارتفاع مصادر التباين إلا أن مشكلة تفاعل الطالب مع المهمة تحتاج إلى دراسات تجريبية لتدريب الطلاب على معالجة المهمات ضمن برامج ملائمة.

خاتمة:

إن الانتقال إلى تقييمات الكفاءة ليست بالعملية السهلة، فرغم أهمية الاستدلال على كفاءة الطلاب في مجال أو عدة مجالات باستخدام عدد ملائم من مهمات التقييم، والتي تؤدي إلى تقييمات صادقة وموثوقة، وتؤدي إلى نجاح الإصلاحات التربوية إلا أنها تواجه تحديات في مراقبة جودة قياساتها.

وبالتفكير في العديد من المهمات المطلوب تقديمها للطلاب في تقييمات الكفاءة لتغطية المحتوى الدراسي المطلوب، وعدد المصحّحين المعتمدين في تصحيح منتوج الطلاب، وعدد الفترات التي يقيّم فيها الطلاب، وعدد طرق التقييم المستخدمة فإن الباحثين وصانعي القرارات يواجهون العديد من التحديات المتعلقة بوجود مصادر متعددة لتباين الخطأ التي يمكن أن تؤثر على ثبات وصدق درجات الطلاب، منها ما يرتبط بالمهمات، والمصحّحين، والفترات، وصيغ التقييم، ومنها ما يرتبط بالتفاعلات التي تحدث بينها.

أثبتت العديد من الدراسات التي تم مراجعتها بأن ثبات وصدق تقييمات الكفاءة يمكن أن تتأثرا بشكل أو بآخر بمصادر متعددة للخطأ، مع العلم أن هذه التقييمات تحتاج إلى مستويات مقبولة من إمكانية التعميم لاتخاذ قرارات صائبة حول أداء الطلاب، كما تتطلب في أحيان كثيرة رفع عدد مستويات أبعادها كزيادة عدد المهمات أو عدد المصحّحين أو عدد الفترات لبلوغ معاملات إمكانية تعميم مقبولة في الدراسات المستقبلية.

وفي ضوء ما تقدم يحتاج المهتمون بتقييمات الكفاءة العناية الكافية بفحص أداء الطلاب في وضعيات أكثر ملاءمة، وضمان انغماس الطلاب في انجاز المهمات، وإعداد شبكات تصحيح واضحة، وإعداد مهمات وطرق تقييم واقعية ومتجانسة، واختيار فترات ملائمة لتقييم أداء الطلاب.

ومن هنا يبدو واضحا بأن ظروف التقييم وخصائص المهمات وتفاعل تلك الخصائص مع الخبرات التعليمية للطلاب لها تأثير كبير على مستويات إمكانية التعميم المطلوب تحقيقها (Linn & Burton, 1994). ولكن في الواقع الأدلة المتاحة بشأن القضايا المتعلقة بتفاعل الطلاب مع خصائص المهمات وظروف التقييم المختلفة محدودة للغاية في أدبيات القياس، لذا فمن المهم جدا أن تكون هذه الخصائص هي محور اهتمام الدراسات المستقبلية لأن الحصول على تقييمات ذات إمكانية تعميم مقبولة أصبحت ضرورة لاتخاذ قرارات صائبة حول كفاءات الطلاب.

المراجع:

- طباع، ف. (2016). *تقييم نموذج إمكانية التعميم لاختبار تقييم كفاءات الرياضيات وفق الوضعيات المركبة*. رسالة دكتوراه غير منشورة، جامعة سطيف2.

- طباع، ف. (2017). انتقادات استخدام مصطلح الكفاءات في الممارسات التربوية المرتبطة بالتقويم. *مجلة العلوم الاجتماعية*، 24، 162-177.

- طباع، ف & ليفة، ن. (2015). تقييم الكفاءات من منظور نظرية إمكانية التعميم. *مجلة دراسات نفسية وتربوية*، 12، 207-226.

- علام، ص. (2004). *التقويم التربوي البديل: أسسه النظرية والمنهجية وتطبيقاته الميدانية*. القاهرة: دار الفكر العربي.

- Baartman, L., Bastiaens, T., Kirschner, P., & van der Vleuten, C. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluation, 32*, 153-177.

- Baartman, L., Prins, F., Kirschner, & van der Vleuten, C. (2007). Determining the quality of competence assessment programs: a self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258-281.

- Bain, D. (2014). Généralisabilité et évaluation des compétences : Pistes et fausses pistes. Dans C. Dierendonck, *L'évaluation des compétences en milieu scolaire et en milieu professionnel* (pp. 141-165). Bruxelles: De Boeck.

- Brennan, R. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

- Brennan, R. (2001). *Generalizability Theory.* New York: Springer-Verlag.

- Brennan, R., & Johnson, E. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 9-12.

- Briesch, A., Swaminathan, H., Welsh, M., & Chafouleas, S. (2014). Generalizability theory: A practical guide to study design implementation, and interpretation. *Journal of School Psychology, 52*(1), 13–35.

- Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure.* Berne: Peter Lang.

- Cardinet, J., Sandra, J., & Pini, G. (2010). *Applying generalizability theory using EDUG.* New York: Routledge.

- Chen, E., Niemi, D., Wang, H., & Mirocha, J. (2007). *Examining the generalizability of direct writing assessment tasks.* Technical Report, CRESST: University of California.

- Clausser, B., Clyman, S., & Swanson, B. (1999). Components of rater error in complex performance assessment. *Journal of Educational Measurement, 36*(1), 29-45.

- Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

- De Ketele, J., & Gerard, F. (2005). La validation des épreuves selon l'approche par compétences. *Mesure et Évaluation en Éducation, 28*(3), 1-26.

- Dunbar, S., Koretz, D., & Hoover, H. (1991). Quality Ccntrol in the dvelopment and ue of prformance asessments. *Applied Measurement in Education, 4*(4), 289-303.

- Gao, X., & Brennan, R. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*(2), 191–203.
- Gao, X., Shavelson, R., & Baxter, G. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(4), 323-342.
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531.
- Gipps, C. (1994). *Beyond testing: towards a theory of educational assessment.* London: Routledge-Falmer.
- Güler, N., & Gelbal, S. (2010). Studying reliability of open-ended mathematics items according to the generalizability theory. *Educational Sciences: Theory & Practice, 10*(2), 1011-1019.
- Hébert, M. (2006). *L'interaction statistique élèves × tâches comme source d'erreur de mesure en évaluation des apprentissages des compétences.* Mémoire de maitrise, Université de Laval.
- Hébert, M., & Duclos, V. (2007). Fiabilité d'outils d'évaluation des apprentissages dans une approche par compétences : L'apport de la théorie de la généralisabilité. *Journal of Educational Measurement and Applied Cognitive Sciences, 1*(1), 1-11.
- Hébert, M., Valois, P., Scallon, G., & Frenette, E. (2014). Fiabilité d'un dispositif d'évaluation de l'habileté à déterminer le résultat d'une chaîne d'opérations chez des élèves québécois du secondaire. *Mesure et évaluation en éducation, 37*(1), 21-41.
- Huang, H. (2009). Magnitude of task-Sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement, 69*(6), 887-912.
- Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123-139.
- In'nami, Y., & Koizumi, R. (2015). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341-366.
- Johnson, R., Penny, J., & Gordan, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: Guilford Press.
- Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125–160.
- Klein, S., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The realiability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education, 8*(3), 243-260.
- Lane, S., Liu, M., Ankenmann, R., & Stone, C. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.
- Lee, Y., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7*(4), 353-385.
- Linn, R., & Burton, E. (1994). Performance-based assessment: implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-15.
- Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

- McBee, M., & Barnes, L. (1998). The generalizability of a performance assessment measuring achievement in eight-grade mathematics. *Applied Measurement in Education, 11*(2), 179-194.
- Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.
- Miller, M., & Linn, R. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367-378.
- Nie, Y., Yeo, S., & Lau, S. (2007). Application of generalizability theory in the investigation of the quality of journal writing in mathematics. *Studies in Educational Evaluation, 33*(3-4), 371-383.
- Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Education Policy Analysis Archives, 16*(8), 1-14.
- Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational Assessment, 7*(2), 143–164.
- Parkes, J., Zimmaro, D., Zappe, S., & Suen, H. (2000). Reducing task-related variance in performance assessment using concept maps. *Educational Research and Evaluation, 6*(4), 357–378.
- Ruiz-Primo, M., Baxter, G., & Shavelson, R. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41-53.
- Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences.* Bruxelles: De Boeck.
- Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* California: Sage Publications.
- Shavelson, R., Baxter, & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.
- Shavelson, R., Baxter, G., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.
- Shavelson, R., Ruiz-Primo, M., & Wiley, E. (1999). Note on sources of sampling variability in science performance assessment. *Journal of Educational Measurement, 36*(1), 61-71.
- Taylor, A., & Pastor, D. (2013). An application of generalizability theory to evaluate the technical quality of an alternate assessment. *Applied Measurement in Education, 26*(4), 279-297.
- Webb, N., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*(3), 277–301.