استخدام نظرية إمكانية التعميم في تقدير ثبات اختبار تقييم كفاءة الرياضيات لدى طلاب السنة الرابعة ابتدائي

فاروق طباع*

تاريخ تسلم البحث 2019/1/31 تاريخ قبوله 2019/8/25

ملخص: استخدمت الدراسة الحالية نظرية إمكانية التعميم في تقدير ثبات اختبار تقييم كفاءة الطلاب في الرياضيات، وأثناء الدراسة، طبق اختبار يشتمل تسع مهمات معقدة موزعة على ثلاث صيغ: أ) ثلاث مهمات محكمة البناء، ب) ثلاث مهمات غير محكمة البناء، ج) ثلاث مهمات ذات معلومات مشوشة، كأساس لتقييم كفاءة الأعداد والحساب، على عينة مكونة من (331) طالبًا في السنة الرابعة ابتدائي. وقد شارك في عملية تقييم أداء الطلاب ثلاثة مقدرين مدربين باستخدام شبكات تصحيح تحليلية، وتم تحليل البيانات بواسطة تصميم ثنائي البُعد متقاطع كليًا "طالب× مهمة× مقدر" باستخدام حزمة "EduG". أظهرت نتائج الدراسة وجود مصادر خطأ كبيرة راجعة إلى أثر تفاعل الطالب مع المهمة والتأثير الرئيسي للمهمة. ومن أجل ضمان مستويات ثبات مقبولة يجب زيادة عدد المهمات وليس عدد المقدرين. كما ينبغي العناية أثناء استخدام المهمات المعقدة في قياسات تقييم الكفاءة.

(**الكلمات المفتاحية**: نظرية التعميم، الثبات، اختبار تقييم الكفاءة، مصادر تباين الخطأ، المهمات المعقدة).

مقدمة: يعدّ تقييم الكفاءة قضيّة مهمة شغلت اهتمام الباحثين والمتخصصين في المجال التربوي خلال العقدين الماضيين، فمنذ ظهوره كرد فعل على الممارسات التقليدية أحدث تغييرات شاملة على إجراءات تقييم تعلّم الطلبة، فقد كانت الممارسات التربوية تركز على تقييم المعارف والمهارات المتفرقة باستخدام أسئلة الاختيار من متعدد، والصح والخطأ، والإجابات القصيرة التي تميّزت محتوياتها بسياقات مجردة، ثم انتقلت إلى استخدام أساليب بديلة لتقييم الكفاءات تعتمد على المهمات المعقدة، والمشكلات المفتوحة، والمقالات، والمحادثات الشفهية، والعروض، والتجارب العملية، والمشاريع، وملفات الأعمال...وغيرها (Alam, 2004 ; Feuer & Fulton, 1993 ; Johson, Penny & Gordon, 2009)، وتعتمد على مهمات واقعية متنوعة، وتتطلب بناء الإجابة من طرف الطالب، وتستدعي مهارات التفكير العليا، وتعتمد أثناء التصحيح على أحكام ذاتية، وتستخدم عدّة مقدرين أثناء التقييم، وتستدل على الأداء في مواقف متعددة (Scallon, 2004; Segers, Dochy & Cascallar, 2003).

وبالرغم من الخصائص المتميّزة لأساليب تقييم الكفاءة، إلا أن عملية فحص جودتها لا سيما الثبات، تواجه مشكلات باستخدام أساليب النظرية الكلاسيكية للاختبار كالاستقرار عبر الزمن (طريقة الاختبار-إعادة الاختبار)، والاتساق بين المهمات (طريقة الاتساق الداخلي)، واتساق تقديرات المصححين (طريقة الاتفاق بين المقدرين) (De Ketele & Gerard, 2005 ; Scallon, 2004; Tabba, 2017). ويعود ذلك حسب كرونباخ ولين وبرينان وهارتل (Cronbach, Linn, Brennan & Haertel, 1997) إلى أن أساليب تقييم الكفاءة تتصف بسمات أساسية بارزة، إنها تعتمد على اختبارات محكّية المرجع، والمهمات المستخدمة ذات نهاية مفتوحة ومعقّدة تجعل من الطرق الكلاسيكية لتقدير خطأ القياس غير ملائمة.

Using Generalizability Theory in Estimating Reliability of a Mathematical Competence Assessment Test of Fourth Year Primary School Students

Farouq Tebaa, *Mouloud Mammeri University, Algeria.*

Abstract: The current study used Generalizability Theory to estimate the reliability of a mathematical competence assessment test. During the study, the test was composed of nine different complex task formats: a) three well-defined tasks, b) three ill-defined tasks and c) three tasks with parasite data. These tasks were administered to a sample of (331) fourth year primary school students. Three trained raters participated in the scoring process by means of analytic scoring rubrics. Data collected were analyzed in terms of a fully crossed two-faceted design "person× task× rater" using "EduG" package. Research results showed substantial sources of error due to person-task interaction effect and task main effect. To ensure acceptable levels of reliability, the number of tasks should be increased but not the number of raters. As such, special caution should be put on the use of complex tasks in competence assessment measures.

(**Keywords**: Generalizability Theory, Reliability, Competence Assessment Test, Sources of Error Variance, Complex Tasks).

ويتطلب أثناء تقدير ثبات أساليب تقييم الكفاءة أن تكون الظروف التي يُجرى فيها التقييم نفسها قدر الإمكان لكل الطلبة، وتُصحح إجاباتهم بطريقة متسقة بواسطة المعايير نفسها، كما يتم تقييم الطلبة بواسطة مهمات متعددة، وتُصحّح إجاباتهم بواسطة عدّة مقيّمين، وفي عدّة مواقف لاتخاذ قرارات صائبة حول الطلبة (Baartman, Bastiaens, Kirschner & van der Vleuten, 2006; Baartman, Prins, Kirschner & van der Vleuten, 2007).

وفي ظل هذه الشروط يتم تقييم الطالب بواسطة مجموعة من المهمات الاختبارية التي يتم تقييمها باستخدام مجموعة من المقدّرين خلال فترات مختلفة، ويتحصل فيها على درجة شاملة أو كلية (Universe score) واحدة تمثّل متوسط درجاته إذا أنجز كل المهمات الممكنة، وصُحّحت إجاباته من طرف عدة مقدّرين، وأنجز مهمات الاختبار في فترات معينة (Shavelson, Baxter & Gao, 1993) ما قد يؤدي إلى اختلاف درجات الطالب باختلاف مجموعة المهمات الاختبارية التي تقدّم له، وتختلف باختلاف المقدّرين المعيّنين لتقييم أدائه، لأن درجاته تتأثر بمصادر خطأ قياس راجعة إلى اختلاف المهمات واختلاف المقدّرين.

* جامعة مولود معمري – الجزائر.

ومن بين الطرق التي يمكن أن توفر أدوات تقييم جيدة معرفة وتحديد حجم خطأ القياس الذي يمكن أن يسهم به كل بُعد من أبعاد خطأ القياس، والتفاعل بين هذه الأبعاد في الدرجات الملاحظة التي يتم الحصول عليها باستخدام أدوات التقييم. وتمثّل المهمات الاختبارية والمقدّرين أحد الأبعاد التي يمكن أن تؤثّر على ثبات درجات اختبارات تقييم كفاءة الطلبة التي تعتمد على المهمات المعقدة، وتمثل هذه المهمات مصدرًا مهمًا لخطأ القياس كون أداء الطالب يمكن أن يتباين في هذه المجموعة من المهمات عن مجموعة أخرى يتم اختيارها من نفس النطاق الواسع من المهمات (Dunbar, Koretz & Hoover, 1991)، حيث يمكن أن تكون بعض المهمات صعبة عند بعض الطلاب وسهلة عند بعضهم، كما يمكن أن يغيّر الطالب من استراتيجيات معالجته للمهمات من مهمة إلى أخرى (Shavelson, Baxter & Pine, 1992). ومن ثمّ تمثّل عينة المهمات مصدرًا لخطأ القياس عند محاولة تعميم درجات تقييم أداء الطلاب من عينة من المهمات إلى نطاق أوسع من المهمات، وهو ما يعرف لدى بعض المؤلفين بخطأ معاينة المهمة (Huang, 2009 ; Shavelson et al., 1993).

وفيما يتعلّق بالأشخاص المكلفين بتقييم كفاءة الطلبة، فيعتمد على المعلمين كمقيّمين لاتخاذ قرارات تحديد مستوى الأداء، فهم المسؤولون عن تقييم أداء طلبتهم، وهناك العديد من الأدلة البحثية التي تشير إلى ضرورة الاعتماد على المعلمين في عملية التقييم، كما يمكن أن يمثلوا أحد أبعاد خطأ القياس لا سيما إذا طلب إليهم إصدار أحكام حول أداء الطلبة (Scallon, 2004). ومن ثم فإن المقدّرين يمثلون مصدرًا لخطأ القياس؛ لأن تقييماتهم للطلبة يمكن أن تتأثّر بالذاتية، بحيث تتسم تقديراتهم بالتشدّد أثناء التصحيح فيميل بعضهم إلى الصرامة، في حين يميل الآخرون إلى التساهل، كما يمكن أن تتأثر أيضا تقديراتهم بأثر الهالة الناتج عن اختلاف فترات تصحيح المهمات أو الناتج عن استخدام موازين تقدير غير واضحة (Casanova & Demeuse, 2011)، فعينة المقدّرين أو المقيّمين قد تمثل أيضًا مصدرًا لخطأ القياس أثناء تعميم أداء الطلاب من عينة مقدّرين إلى نطاق أوسع من المقدرين، وهو ما يعرف بخطأ معاينة المقدّر.

وبهدف التحكم في مختلف مصادر خطأ القياس التي تؤثر على ثبات درجات الطلبة في الاختبار، وتقديم مؤشرات إحصائية عن الثبات التي تعبّر عن دقة تعميم درجة الفرد الملاحظة في اختبار معين مقارنة بالدرجة المتوسطة التي يحصل عليها ضمن الشروط الممكنة التي سوف يخضع لها (مهمات، ومقدرين)، فإنّه يتطلب استخدام نظرية إمكانية التعميم التي تمتلك إطارًا مفاهيميًا واسعًا ومجموعة قوية من الأساليب الإحصائية التي تسمح بمعالجة مختلف مشكلات القياس (Brennan, 2001 ; Cardinet, Johnson & Pini, 2010; Cardinet & Tourneur, 1985 ; Shavelson & Webb, 1991).

وتعدّ نظرية إمكانية التعميم امتدادًا للنظرية الكلاسيكية للاختبارات من خلال تطبيق أساليب تحليل التباين (ANOVA)، فهي إطار مفاهيمي وإحصائي موسع للنظرية الكلاسيكية للاختبارات تسمح بتقدير دقة القياسات في مواقف تتأثر فيها بمصادر متعددة للخطأ، وتُقدّم طرقًا فعّالة في تحسين إجراءات القياس المستقبلية (Brennan, 2001 ; Cardient et al., 2010)، وقد قدمها كرونباخ وراجاراتنام وجليزر (Cronbach, Rajaratnam & Gleser, 1993) لتحرير النظرية الكلاسيكية للاختبارات عن طريق إيجاد طرق كفيلة لمعالجة مختلف مصادر أخطاء القياس في الوقت نفسه، والتي قد تكون راجعة إلى خصائص الفاحصين، أو صيغ الاختبار، أو فترات تطبيق الاختبار، أو نوع عبارات الاختبار...إلخ.

وتلعب نظرية إمكانية التعميم دورًا فريدًا من نوعه مقارنة بالنظرية الكلاسيكية للاختبارات في تصميم إجراءات القياس، فهي تتميّز بأربعة خصائص بالنظر إلى النظرية الكلاسيكية للاختبارات، حيث بإمكانها إدماج مصادر متعددة للخطأ وتُقيَم في الوقت نفسه (ثبات الاختبار وإعادة الاختبار، وثبات الاتساق الداخلي، والصدق الظاهري، والثبات بين التقديرات)، وتقدير ليس فقط تأثير أبعاد القياس على حدة ولكن تأثيرات التفاعل أيضا، وتوسيع ثبات القياسات مع الاقتصاد من التكاليف والوقت، وتقديم معلومات عن ثبات التفسيرات النسبية وحتى التفسيرات المطلقة أثناء تقييم أداء الطالب (Yin & Shavelson, 2008).

وفي إطار نظرية إمكانية التعميم يقوم الباحث بجمع الملاحظات (مثلاً: درجات الطلبة) في موقف معين، وهذه الدرجات هي بمثابة عينة من "الملاحظات" التي يشتمل عليها نطاق شامل من الملاحظات الذي يتضمن "ظروف أو شروط جمع الملاحظات" أي المثيرات أو المواقف المعينة التي يتم في ضوئها جمع الملاحظات أو الحصول على القياسات (Allam, 2000)، ومجموعة الشروط التي يتم فيها قياس أداء الطلاب في الاختبار عبارة عن "نطاق الملاحظات المقبولة" (Universe of Admissible Observations)، والذي يحدده كل باحث حسب سياق التقييم المرغوب، وكل قياس يتم فيه أخذ عينة يُفترض أن يكون قابلاً للاستبدال بنطاقات أخرى. ويشير نطاق الملاحظات المقبولة إلى كل الأبعاد التي يمكن اختيارها من أجل تصميم وضعية القياس، فهو إطار معاينة يتم فيه تحديد خصائص كل بُعد يشتمل عليه القياس (Meyer, 2010)، ويمكن أن تتباين الدرجات التي تنتمي إلى نطاق شامل معين في أكثر من جانب أو بُعد (facet)، ويمكن أن تكون الأبعاد مهمات الاختبار التي تُطبّق على الطلبة أو المقدّرين المكلّفين بتصحيح إجابات الطلبة أو الفترات التي يُجرى فيها الاختبار.

وتُعدّ هذه الأبعاد منفردة أو بالتفاعل فيما بينها مصادر مهمة من مصادر خطأ القياس المتعددة، التي يمكن أن تأخذها نظرية إمكانية التعميم بعين الاعتبار في الوقت نفسه، وفي دراسات إمكانية التعميم (Generalizability Studies) يهتم الفاحص بدراسة تأثير شروط أو ظروف الاختبار على الدرجات من أبعاد مختلفة (مثلاً: المهمات، والمقدرين)، ومن خلال استخدام تحليل التباين يتم تقدير مكونات تباين الدرجات الملاحظة، التي يمكن أن تسهم في التباين الحقيقي الذي مصدره الفرد، ويعكس الفروق بين أداء الأفراد، ويُعرف بموضوع القياس (Object of Measurement)، وتباين الخطأ (Variance of Error) الذي مصدره كل بُعد من الأبعاد، وتفاعل الأبعاد مع بعضها البعض، وتفاعل الأبعاد مع الفرد.

ومن خلال مكونات التباين يمكن تقدير ثبات أداة القياس عن طريق حساب معاملات إمكانية التعميم (Generalizability Coefficients)، والتي تعبّر عن نسبة تباين الدرجة الشاملة (التباين الحقيقي) إلى تباين الدرجة الملاحظة أو التباين الكلي (التباين الحقيقي + تباين الخطأ)، وتعتمد مكونات تباين الخطأ على نوع الخطأ الذي سوف يتم مراعاته، ففي الخطأ النسبي (Relative Error) يهدف إجراء الاختبار إلى تصنيف الطلبة فيما بينهم، ويأخذ بعين الاعتبار أثر التفاعل بين الأبعاد، ويتم على أساسه حساب معامل إمكانية التعميم النسبي، بينما في الخطأ المطلق (Absoulte Error) يهدف الاختبار إلى تحديد مستوى الطلبة وفق مقياس معين، ويأخذ بعين الاعتبار الخطأ المطلق، والذي بدوره يأخذ بعين الاعتبار الآثار المباشرة للأبعاد وتفاعلاتها التي تسمح بحساب معامل إمكانية التعميم المطلق (Shavelson & Webb, 1991).

وتهدف دراسات القرار (Decision Studies) إلى الحصول على أفضل طريقة قياس تتمتع بالثبات في وضعية معينة بالاعتماد على المعلومات التي تقدمها دراسات إمكانية التعميم للوصول إلى أفضل الطرق، وأكثرها فعالية لاستخدام أداة القياس بالتقليل من أخطاء القياس لتحقيق هدف معين (Briesch, Swaminathan, Welsh & Chafouleas, 2014)، والتساؤل الذي تسعى دراسات القرار للإجابة عنه هو: ماذا لو؟ بمعنى ماذا لو تم استخدام خمس مهمات اختبارية عوضًا عن ثلاث مهمات؟، وماذا لو تم استخدام أربعة مقيّمين بدلاً من مقيّمَيْن اثنين؟ وذلك بهدف تحديد أثرها على الثبات لاتخاذ قرارات أكثر دقة في المستقبل.

يتضح مما سبق أنه من الضروري الرجوع إلى نظرية إمكانية التعميم لتقدير ثبات الاختبارات نظرًا لملاءمتها، وقدرتها على معالجة مختلف الظروف التي تتأثر فيها وضعيات التقييم بمصادر متعددة لخطأ القياس، والاستجابة لمختلف أنواع القرارات المتخذة من مختلف أنواع اختبارات تقييم الكفاءة (Tebba & Lifa, 2015; Bain, 2014; Brennan, 2000).

وفي هذا السياق أجريت العديد من الدراسات التي هدفت إلى تقدير ثبات تقييمات أداء الطلبة في مختلف المستويات والمجالات الدراسية، من خلال فحص مصادر تباين الخطأ الراجعة إلى المهمات والمقدّرين وتفاعلاتها فيما بينها وتفاعلات هذه الأبعاد مع الأفراد، وتقدير معاملات إمكانية التعميم النسبية والمطلقة، وزيادة عدد مستويات الأبعاد للحصول على مستويات ثبات مقبولة.

وتعدّ الدراسة التي قام بها شافلسون وزملاؤه (Shavelson et al., 1993) من أوائل الدراسات في هذا المجال، حيث هدفت دراستهم إلى الكشف عن تغير معاينة تقييمات الأداء في العلوم لدى عينة مكوّنة من (186) طالبًا من طلبة الخامس والسادس ابتدائي في ولاية كاليفورنيا، وتم تصحيح أداء الطلبة بواسطة مقدّرين في ثلاث مهمات تقييم من المهمات البديلة، وتم تحليل البيانات باستخدام تصميميْن أحدهما ثنائي البُعد "فرد × مقدر × مهمة"، والآخر ثلاثي البُعد "فرد × مقدر × مهمة × فترة". وقد أظهرت النتائج أن أكبر مصدر تباين راجع إلى تفاعل الفرد مع المهمة وتفاعل الفرد مع المهمة مع المقدر، أما مصادر تباين المقدر وتفاعلاته جاءت منخفضة، وكانت معاملات إمكانية التعميم منخفضة، ومن أجل بلوغ معامل إمكانية التعميم (0.80) يتطلب (8) مهمات.

وفي دراسة أخرى لشافلسون وزملائه (Shavelson et al., 1993) نُشرت في نفس المقال، وهدفت إلى الكشف عن تغيّر معاينة تقييمات الأداء في الرياضيات لدى عينة تكونت من (105) طالبًا من طلبة السادس الابتدائي في ولاية كاليفورنيا، وبالإجابة عن ثلاث مهمات، وصُحّحت بواسطة مقدريْن، وتم تحليل البيانات وفق تصميم ثنائي البُعد "فرد × مقدر × مهمة"، وقد كشفت النتائج أن أكبر مصدر تباين راجع إلى تفاعل الفرد مع المهمة، أما مصادر تباين المقدر وتفاعلاته منخفضة، وجاءت أيضًا معاملات إمكانية التعميم منخفضة، ومن أجل بلوغ معامل إمكانية تعميم مقبول يتطلب ذلك (15) مهمة.

كما قام لين وليو وأنكمان وستون (Lane, Liu, Ankenmann & Stone, 1996) بإجراء دراسة هدفت إلى التحقق من إمكانية تعميم وصدق اختبار تقييم أداء الرياضيات لدى عينة مكوّنة من طلبة السادس والسابع اختيرت من ست مدارس في أمريكا، وذلك بالإجابة عن تسع مهمات، وتم تحليل البيانات بواسطة تصميميْن أحادي البُعد "فرد × مهمة"، وثنائي البُعد "فرد × مهمة × مقدر"، وكشفت النتائج عن ارتفاع في مكون تفاعل الطالب مع المهمة، وانخفاض في مكونات تباين المقدر وتفاعلاته، كما كشفت عن معاملات إمكانية التعميم مقبولة، وأن زيادة عدد المهمات ترفع معاملات إمكانية التعميم أفضل بكثير من زيادة عدد المقدّرين .

وفي المجال نفسه، أجرى مكبي وبارنس (McBee & Barnes, 1998) دراسة للتحقق من إمكانية تعميم الأداء لقياس تحصيل الرياضيات لدى عينة مكوّنة من (101) طالبًا من طلاب الصف الثامن في مقاطعة ويدواسترن بأمريكا، وقام الباحثان باختيار أربع مهمات منها مهمتان أكثر تجانسًا للإجابة عليها من طرف الطلبة، وتصحيح أدائهم من طرف مقدريْن اثنين، واستخدم تصميم ثلاثي البُعد "فرد × مهمة × مقدر × فترة" لتحليل البيانات، وقد أظهرت نتائج الدراسة ارتفاعًا في تباين تفاعل الطالب مع المهمة وتفاعل الطالب مع المهمة مع المقدر الممزوج بالخطأ العشوائي، وانخفاضًا في مصادر تباين الخطأ بتحليل بيانات المهمات الأكثر تجانسًا، أما عدد المهمات المطلوبة لتحقيق مستويات مقبولة الثبات لا يمكن بلوغها حتى أثناء استخدام المهمات الأكثر تجانسًا.

كما أجرى ويب وشلكمان وسيرجي (Webb, Schlackman & Sugrue, 2000) دراسة على عينة مكوّنة من (57) طالبًا من طلبة الصف السابع والثامن في مقاطعة لوس أنجلس، وهدفت إلى فحص تقديرات إمكانية تعميم درجات تقييم العلوم وإمكانية استبدال صيغ الاختبار، وذلك بالإجابة عن أربع مهمات صُحّحت من طرف مقدريْن، وتم تحليل البيانات وفق تصميميْن الأول ثنائي البُعد "فرد × مهمة × مقدر" والثاني ثلاثي البُعد "فرد × مهمة × مقدر × فترة"، وقد أظهرت النتائج أن أكبر مصدر تباين راجع إلى تفاعل الفرد مع المهمة، وساهمت الفترة في خفض تأثير تفاعل الفرد مع المهمة، فأصبح تفاعل الفرد مع المهمة مع الفترة أكبر مصدر تباين، وجاءت معاملات إمكانية التعميم مرتفعة ولكن انخفضت بعد تضمين الفترة في التصميم، ومن أجل بلوغ معاملات مقبولة يتطلب زيادة عدد المهام.

وأجرى قاو وبرينان (Gao & Brennan, 2001) دراسة هدفت إلى تقييم الجودة الفنية لتقييمات الأداء بفحص تغير معاينة مكونات التباين المقدرة في الاستماع والكتابة، وذلك بتطبيق أعداد مختلفة من المهمات على عينات الطلبة بين السنوات (1992 و1993 و1994)، وتم تصحيح إجابات الطلبة بواسطة مقدّرين، وتحليل البيانات وفق تصميم ثنائي البُعد "فرد × مهمة × مقدر"، وقد أظهرت النتائج أن أكبر مصدر تباين راجع إلى تفاعل الفرد مع المهمة وتفاعل الفرد مع المهمة مع المقدر، وتباين المهمة، وجاءت مصادر التباين الأخرى ضعيفة، كما أظهرت انخفاضًا في معاملات إمكانية التعميم في الاستماع وارتفاعًا في الكتابة.

وفي دراسة أجراها ني ويو ولو (Nie, Yeo & Lau, 2007) حول استخدام نظرية إمكانية التعميم لفحص جودة كتابة مقالة في الرياضيات على عينة مكوّنة من (29) طالبًا من طلبة المتوسطة في سنغافورة، وذلك بإنجاز الطلبة لمهمتيْن وتصحيح أدائهم من طرف مقدرين، واستخدام تصميم ثنائي البُعد "فرد × مهمة × مقدر"، وقد كشفت النتائج عن معاملات إمكانية تعميم مقبولة، وارتفاع في مصدر تباين تفاعل الطالب مع المهمة، وجاءت مصادر تباين المهمات والمقدّرين ضعيفة، ومن أجل رفع معاملات إمكانية التعميم يتطلب زيادة عدد المهمات بدلاً من زيادة المقدرين.

وقام شين ونيمي وونج وميروشا (Chen, Niemi, Wang & Mirocha, 2007) بنشر تقرير عن دراسة هدفت إلى التحقق من إمكانية تعميم مهمات التقييم المباشرة في الكتابة، بحيث طبقت أربع مهمات (ثلاث مهمات في الأدب، ومهمة واحدة في القصة) على عينة مكوّنة من (397) طالبًا من طلبة الصف التاسع الذين أنجزوا مهمتيْن عشوائيًا، وتم الاعتماد في تصحيح المهمات على أربعة مقدّرين، وتم تحليل البيانات بواسطة تصميم ثنائي البُعد متقاطع كليًا "طالب × مقال × مقدر". كشفت النتائج أن أكبر مصدر تباين راجع إلى تفاعل الطالب مع المقال مع المقدر الممزوج بالخطأ العشوائي، وتفاعل الطالب مع المقال، وجاءت مصادر التباين الأخرى منخفضة، كما كشفت أن الطريقة المثلى لرفع معاملات إمكانية التعميم يتطلب زيادة عدد المقالات.

وهدفت دراسة لي وكانتور (Lee & Kantor, 2007) إلى فحص تأثيرات المهمات والمقدّرين على درجات الطلبة في الكتابة، وتأثير عدد المهمات والمقدرين، وتصميمات التقدير على ثبات الاختبار بواسطة نظرية إمكانية التعميم، استخدمت ستّ مهمات مدمجة في الكتابة (الاستماع والقراءة) طبقت على (448) طالبًا من طلبة اللغة الإنجليزية في خمسة بلدان (أمريكا، وكندا، وهونغ كونغ، ومكسيكو، وتايوان)، وصُحّحت المهمات من طرف ثلاث أزواج من المقدرين، وتم تحليل البيانات باستخدام تصميميْن، الأول ثنائي البُعد متقاطع جزئيا "فرد : مقدر × مهمة"، والثاني ثنائي متقاطع كليا "فرد × مهمة × مقدر"، وقد أظهرت نتائج الدراسة أن معاملات إمكانية التعميم كانت مقبولة، وأن أكبر مصدر تباين راجع إلى تفاعل الفرد مع المهمة مع المقدر الممزوج بالخطأ غير العشوائي، وتفاعل الفرد مع المهمة، في حين أظهرت انخفاضًا في مصادر التباين الأخرى، بالإضافة إلى أن زيادة عدد المهمات أكثر تأثيرا من زيادة عدد التقديرات في رفع معاملات الثبات.

وفي دراسة أجراها باين (Bain, 2008) هدفت إلى إعداد اختبار مشترك لتقييم كفاءات الرياضيات وفق نموذج إمكانية التعميم لدى عينة مكوّنة من (43) طالبًا من طلبة السنة التاسعة بجنيف، وطبق عليهم اختبار مكوّن من (15) مهمة، وتم تحليل البيانات وفق تصميم متقاطع جزئيًا "طالب × مهمة: مدرسة × مجموعة المستوى"، وقد كشفت نتائج الدراسة ارتفاعًا في معاملات إمكانية التعميم، وارتفاعًا في مصدر تباين تفاعل الطالب مع المهمة المتداخل مع المدرسة والمجموعة، وتباين المهمة، في حين كشفت عن انخفاض في مصادر التباين الأخرى.

وفي دراسة أجراها جبريل (Gebril, 2009) هدفت إلى الكشف عن أثر مختلف الأبعاد على ثبات درجات الكتابة باستخدام نظرية إمكانية التعميم لدى عينة مكوّنة من (115) طالبًا من طلبة السنة الرابعة في جامعة سوهاج بمصر، وتم الاعتماد على أربع مهمات كتابية منها مهمتان مستقلتان ومهمتان مندمجتان أخذت من اختبار الإنجليزية كلغة أجنبية، وتم تصحيحها من طرف ثلاثة مقدرين، واعتمد الباحث في تحليل البيانات على تصميم ثنائي البُعد متقاطع كليا "طالب × مهمة × مقدّر"، وكشفت النتائج أن أكبر مكوّن تباين كان راجعًا إلى تفاعل الطالب مع المهمة والمقدر الممزوج بالأخطاء العشوائية غير المقدرة في التصميم ثم يليها مكوّن تباين تفاعل الطالب مع المهمة، بينما جاءت مكونات التباين الأخرى منخفضة، وكشفت أيضًا عن معاملات إمكانية تعميم منخفضة لكنها مشجّعة، وأن زيادة عدد المقدرين لا يرفع من معاملات إمكانية التعميم.

وفي تحليل ما ورائي أجرى هوانج (Huang, 2009) مراجعة للدراسات التي استخدمت نظرية إمكانية التعميم بين سنتي (1980-2006)، وهدفت إلى فحص تغيّر معاينة المهمة في تقييمات الأداء، وقام الباحث بتحليل (50) دراسة تضمّنت (130) مجموعة بيانات مستقلة، وتم تحديد قيمة مكونات تباين المهمة، ومكوّنات تباين تفاعل الفرد مع المهمة في كل الدراسات لتقدير حجم الأثر، وقد كشفت النتائج أن نسبة تباين المهمة بلغت حوالي (12%)، وبلغت نسبة تباين تفاعل الفرد مع المهمة (26%)، كما كشفت أنه من أجل خفض تغيّر معاينة المهمة يتطلب إدماج أكبر عدد من الأبعاد في وضعية القياس، واستخدام تصميمات متقاطعة بدلاً من تصميمات متداخلة، وإدماج الفترة كبُعد من أبعاد القياس.

كما هدفت دراسة جولر وجيلبال (Güler & Gelbal, 2010) إلى تقدير ثبات أسئلة ذات نهاية مفتوحة بواسطة النظرية الكلاسيكية ونظرية إمكانية التعميم، واستخدم الباحثان اختبارًا مكوّناً من (24) مهمة مفتوحة في الرياضيات التي استخدمت في دراسة التوجّهات العالمية في الرياضيات والعلوم، وطبّق على عينة مكوّنة من (203) من طلبة الصف الثامن والتاسع في أنقرة، وصُحّحت إجاباتهم من طرف أربعة مقدرين، وتم تحليل البيانات باستخدام تصميم ثنائي البُعد متقاطع كليا "فرد × مهمة × مقدر"، وقد كشفت النتائج أن مصدر التباين الأكبر راجع إلى تفاعل الفرد مع المهمة، أما مصادر التباين الأخرى فهي منخفضة، كما كشفت عن معاملات إمكانية تعميم مرتفعة، وأن زيادة عدد المهام ترفع من معاملات إمكانية التعميم أفضل من زيادة عدد المقدرين .

وفي دراسة أجراها كازانوفا وديماس (Casanova & Demeuse, 2011) هدفت إلى فحص مصادر الخطأ المؤثرة على ثبات اختبار التعبير الكتابي في الفرنسية كلغة أجنبية باستخدام نظرية إمكانية التعميم ونموذج راش متعدّد الأبعاد، وطبق الاختبار على عينة مكوّنة من (33) طالبًا من الطلبة الأجانب في جامعة باريس، وتم تحليل البيانات وفق تصميميْن ثنائيين "طالب × مهمة × مقدّر" من خلال تقديم مهمتيْن في الكتابة، وتم تصحيحها من طرف أربعة مقدرين. وكشفت الدراسة أن مصادر تباين الخطأ في التصميم الأول راجعة إلى المقدرين وتفاعل المقدرين مع المهمات، وفي التصميم الثاني راجعة إلى تفاعل المقدرين مع المهمات، في حين كانت تباينات الخطأ الأخرى المتبقية ضعيفة، أما معاملات إمكانية التعميم المطلقة فقد جاءت مرتفعة، كما أن زيادة عدد المصححين ترفع من معامل إمكانية التعميم المطلق أفضل من زيادة عدد المقدرين.

وفي دراسة أجراها تيلور وباستور (Taylor & Pastor, 2013) هدفت إلى تطبيق نظرية إمكانية التعميم لفحص ثبات التقييمات البديلة في القراءة والرياضيات لدى طلبة صفوف السنوات الخامسة والثامنة والعاشرة الذين يعانون من إعاقات متنوعة، وقد أخذت بعين الاعتبار مهمتان في ثلاث محاولات تقييم لكل طالب اعتمد في تصحيحها على مقدريْن، واستخدم في تحليل البيانات على تصميم متقاطع جزئيًا "المهمة : الطالب × محاولة التقييم"، وقد أظهرت نتائج الدراسة أن معاملات إمكانية التعميم المطلقة منخفضة، ومصدر تباين الخطأ الأكبر راجع إلى تداخل المهمة مع الطالب، والتفاعل بين المهمة والمحاولة الممزوج بالتفاعل بين الطالب مع المهمة مع المحاولة، في حين كانت مصادر التباين الأخرى منخفضة، كما أظهرت النتائج أن زيادة عدد المهمات وعدد محاولات التقييم تسمح برفع معاملات إمكانية التعميم المطلقة.

وهدفت دراسة هيبرت وفالوا وسكالون وفرينيت (Hébert, Valois, Scallon & Frenette, 2014) إلى تقدير ثبات أداة لتقييم مهارة تحديد سلسلة عمليات حسابية لدى تلاميذ الثانوية في كيبك على عينة مكونة من (82) طالبًا من طلبة السنة الأولى، وطبقت عليهم (12) مشكلة، واستخدم في تحليل البيانات تصميم متقاطع جزئيًا "طالب × مشكلة: مجال الرياضيات × بناء المشكلة"، وقد أظهرت النتائج أن معاملات إمكانية تعميم النسبية والمطلقة ضعيفة، وأن أكبر مصدر تباين راجع إلى تفاعل الطالب مع المشكلة المتداخل بمجال وبناء المشكلة الممزوج بالأخطاء العشوائية غير المقدّرة، والتباين بين المشكلات، بينما جاءت مصادر التباين الأخرى منخفضة .

وفي دراسة أجراها إنّامي وكويزومي (In'nami & Koizumi, 2015) هدفت إلى تحليل دراسات تأثير المهمة والمقدر في المحادثة والكتابة باللغة الثانية، حيث قام الباحثان بمراجعة (38) دراسة استخدمت نظرية إمكانية التعميم في مجال المحادثة والكتابة باللغة الإنجليزية، وقد كشفت الدراسة أن أثر تفاعل المهمات أو تفاعل المقدرين أكبر من التأثيرات المستقلة للمهمات أو التأثيرات المستقلة للمقدرين، وفسّرت أيضًا تأثيرات المهمة والتفاعل المرتبط بالمهمة نسبة أكبر من تباين الدرجات بالمقارنة مع تأثيرات المقدر والتفاعل المرتبط بالمقدر على التباين الكلي للدرجات.

يلاحظ من الدراسات السابقة أن معظمها أظهرت بأن أكبر مصدر للتباين راجع إلى تفاعل الطالب مع المهمة (Bain, 2008 ; Gao & Brenan, 2001; Güler & Gelbal, 2010; 2014 ; Lane et al., 1996 ; McBee & Hébert et al., Barens, 2008 ; Nie et al., 2007; Webb et al., 2000; Shavelson et al.,1993)، وهذا ما يتفق مع التحليل الما ورائي أي مجموعة من الطرق الكمية لجمع وتحليل نتائج الدراسات التي تركز على نفس الموضوع، والذي أجراه هوانج (Haung, 2009) حول الدراسات التي اهتمت بتغيّر معاينة المهمة في تقييم الأداء، والذي خلص إلى أن نسبة كبيرة من مصدر الخطأ راجع إلى تفاعل الطالب مع المهمة، وتتفق أيضًا مع ما توصلت إليه مراجعة إنّامي وكويزومي (In'nami & Koizumi, 2015) بأن أثر المهمة وتفاعل الطالب مع المهمة فسّرت نسبة أكبر من تباين الدرجات بالمقارنة مع تأثيرات المقدر والتفاعل المرتبط مع المقدر.

في حين أظهرت دراسات أخرى (Chen et al., 2007; Gebril, 2009; Lee & Kantor, 2007) بأن أكبر مصدر لتباين الخطأ راجع إلى تفاعل الطالب مع المهمة ومع المقدر الممزوج بالأخطاء العشوائية غير المقدرة، بالإضافة إلى دراسة كازانوفا وديماس (Casanova & Demeuse, 2011) التي كشفت أن أكبر مصدر للخطأ راجع إلى المقدرين، وتفاعل المقدرين مع المهمات.

وفيما يتعلق بمستوى ثبات الاختبارات المستخدمة في تقييم أداء الطلاب، فقد تباينت نتائج الدراسات، ففي حين أظهرت بعض الدراسات (Bain, 2008 ; Güler & Gelbal, 2010 ; Lane et al., 1996; Lee & Kantor, 2007) بلوغ درجات الاختبارات مستويات مقبولة من الثبات، نجد أن معظم الدراسات الأخرى (Hébert et al., 2014; McBee & Barens, 1998; Nie et al., 2007; Shavelson et al., 1993; Taylor & Pastor, 2013) كشفت عن ضعف في ثبات الاختبارات، وكشفت دراسة لين وزملائه (Lane et al., 1996) ودراسة قاو وبرينان (Gao & Brennan, 2001) عن مستويات ثبات مقبولة في اختبارات فرعية وغير مقبولة في اختبارات فرعية أخرى.

ومن ناحية أخرى أثبتت معظم الدراسات (Chen et al., 2007; Güler & Gelbal, 2010; Lane et al., 1996; Lee & Kantor, 2007; McBee & Barens, 1998; Nie et al., 2007; Shavelson et al., 1993; Taylor & Pastor, 2013; Webb et al., 2000) أن زيادة عدد المهمات الاختبارية تساهم أكثر في رفع معاملات إمكانية التعميم مقارنة بزيادة عدد المقدرين، ويعدّ هذا الاجراء أكثر فعالية من زيادة عدد المقدرين لأن مصادر تباين الخطأ التي تؤثر على ثبات درجات الاختبار راجعة بالأساس إلى تفاعل الطالب مع المهمة، وتأثير المهمة. في حين أثبتت دراسة كازانوفا وديماس (Casanova & Demeuse, 2011) عكس ما توصلت إليه الدراسات السابقة، حيث أشارت إلى أن زيادة عدد المقدرين أفضل من زيادة عدد المهمات في رفع معاملات إمكانية التعميم، ودراسة جبريل (Gebril, 2009) التي اعتمدت على زيادة عدد المقدرين، ولكنها كشفت عن عدم فعالية زيادة عدد المقدرين في رفع الثبات، كما كشفت دراسة هوانج (Huang, 2009) عن استخدام طرق أخرى لخفض مكوّنات تباين تفاعل الفرد مع المهمة، وتباين المهمة تمثّلت إدماج أكبر عدد من الأبعاد في وضعية القياس، واستخدام تصميمات متقاطعة بدلا من تصميمات متداخلة، وإدماج الفترة كبُعد من أبعاد القياس.

إن الاختلاف في نتائج الدراسات السابقة حول مدى تحقيق الاختبارات لمستويات مقبولة من الثبات، وانعدام الدراسات في البيئة الجزائرية والعربية -حسب علم الباحث- يكون مبررًا لإجراء الدراسة الحالية، التي يتوقع أن تضيف نتائجها أساسًا للباحثين لإجراء المزيد من الدراسات حول ثبات اختبارات كفاءة الطلاب باستخدام نظرية إمكانية التعميم، وتساعد التربويين على تسليط الضوء على استخدام نظرية إمكانية التعميم في تقدير ثبات الاختبارات لأساليب التقييم الجديدة القائمة على الكفاءة.

وفيما يتعلق بمصادر تباين الخطأ المؤثرة على ثبات درجات الاختبار يلاحظ أيضا اختلاف في النتائج، ففي حين أظهرت معظم الدراسات أن أكبر مصدر للخطأ راجع إلى تفاعل الطالب مع المهمة، نجد أن بعضها الآخر أظهر أن أكبر مصدر للخطأ راجع إلى تفاعل الطالب مع المهمة والمقدر الممزوج بالأخطاء العشوائية غير المقدرة، كما أشارت نتائج دراسات أخرى إلى أن أكبر مصدر للخطأ راجع إلى تفاعل الطالب مع المقدر، مما يتطلب إجراء مثل هذه الدراسة.

ولا بد من الاشارة إلى أن الدراسة الحالية تتشابه مع دراسة هيبرت وزملائها (Hébert et al., 2014) في بعض الجوانب (من حيث نوع المهمات) إلا أنها تختلف في بعضها الآخر، فقد صُمّمت الدراسة الحالية وفق تصميم ثنائي البُعد متقاطع كليًا، ويتضمن المهمات والمقدرين كأبعاد، وأجريت كلا من دراسات إمكانية التعميم (تحديد مصادر تباين الخطأ) ودراسات القرار (زيادة عدد المهمات، وعدد المقدرين)، في حين استخدمت دراسة هيبرت وزملائها (Hébert et al., 2014) تصميم ثلاثي البُعد متقاطع جزئيًا، واهتمت فقط بإجراء دراسات إمكانية التعميم.

وفيما يتعلق بالإجراءات، فقد ركزت أغلب الدراسات السابقة على إعداد مهمات من نوع واحد، وغير معقّدة، وقليلة خاصة في الرياضيات، وهو مجال اهتمام الدراسة الحالية، حيث استخدمت أغلب الدراسات اختبارات لا يتعدى عدد مهماتها عن أربع، باستثناء دراسة لين وزملائه (Lane et al., 19996)، ودراسة جولر وجلبال (Guler & Gelbal, 2010)، ودراسة باين (Bain, 2008)، ودراسة هيبرت وزملائها (Hébert et al., 2014)، والتي كانت مهماتها قد أعدت في مشاريع بحث سابقة، وتم تحليل معظم بياناتها وفق تصميمات متداخلة، في حين استخدمت الدراسة الحالية مهمات معقّدة صيغت وفق ثلاثة أنواع (محكمة البناء، غير

محكمة البناء، ذات معلومات مشوّشة)، وعددها كبير بالنظر إلى المهمات التي استخدمت في معظم الدراسات السابقة، ومن إعداد الباحث، وهذا مما يتطلب إجراء الدراسة الحالية.

واعتمدت معظم الدراسات السابقة على عينات متنوعة اختيرت من مختلف المراحل الدراسية من الابتدائية، والمتوسطة، والثانوية، والجامعية، وأجريت أغلبها على عينات من طلاب المرحلة الثانوية والجامعية، كما اقتصرت على عينات صغيرة باستثناء الدراسات التي أجريت في مجال اللّغات كدراسة شين وزملائه (Chen et al., 2007)، ودراسة لي وكانتور (Lee & Kantor, 2007)، وباستثناء أيضًا الدراسات التي أجريت في مجال الرياضيات كدراسة جولر وجيلبال (Guler & Gelbal, 2010)، ودراسة تيلور وباستور (Taylor & Pastor, 2013) التي أجريت على عينات كبيرة، اعتمدت الدراسة الحالية على طلاب المرحلة الابتدائية نظرًا لقلة الدراسات التي أجريت على هذه المرحلة، وتحديدًا السنة الرابعة باعتبارها فترة تسمح للطلاب بالانتقال إلى السنة الخامسة التي تعدّ مرحلة حاسمة للطلاب للالتحاق بالمتوسطة، كما اعتمدت على عينة كبيرة تم تحليل بياناتها وفق تصميم متقاطع كليًا، مما يبرر إجراءها.

مشكلة الدراسة وأسئلتها

يتضح مما سبق بأن مصادر خطأ القياس التي تؤثر على ثبات درجات تقييم الكفاءة راجعة إلى المهمات الاختبارية، وإلى المقيّمين أو المقدرين، فالحصول على تقييمات موثوقة تعكس واقع كفاءة الأعداد والحساب لدى الطلبة يحتاج بالدرجة الأولى إلى التعرف على أبعاد خطأ القياس، ومدى إسهام كل واحد منها منفردًا أو من خلال التفاعل فيما بينها في اختلاف أو تباين الدرجات الملاحظة المحصلة من الاختبار.

ومن خلال ما تم ذكره فإن الدراسة الحالية تهدف إلى تقدير نسب تباين الخطأ التي يمكن أن تفسّرها كل من المهمات والمقدرين كأبعاد محتملة لخطأ القياس في التباين الكلي لدرجات طلاب الرابع ابتدائي على اختبار تقييم كفاءة الأعداد والحساب، كما تهدف إلى فحص أفضل الشروط التي تتيح لهذا الاختبار تحقيق أفضل المستويات من الثبات، وفي هذا الاطار تحاول هذه الدراسة الاجابة على السؤالين الآتيين:

1- ما مقدار تباين الخطأ الذي يفسّره كل من أبعاد (المهمات، والمقدرين) في التباين الكلي لدرجات طلاب الرابع ابتدائي في اختبار كفاءة الأعداد والحساب في الرياضيات؟

2- ما شروط تحقيق اختبار كفاءة طلاب الرابع ابتدائي في الأعداد والحساب في الرياضيات لأفضل مستويات الثبات؟

أهميـة الدراسة

بيّنت الأدبيات التربوية أن هناك العديد من الدراسات التي تناولت استخدام نظرية إمكانية التعميم في تقدير ثبات اختبارات تقييم الكفاءة، إلا أن هذه الدراسات تباينت في بعض نتائجها من سياق تقييم إلى آخر، ومن أنواع الاختبارات إلى أخرى، ومن تصميم إلى آخر، ومن هنا تبرز أهمية الدراسة الحالية بما يمكن أن تضيفه إلى المعرفة، وما يمكن أن يستفاد منها في الممارسات التربوية، وذلك من خلال ما ستقدمه إلى الدراسات المنشورة باللغة العربية في مجال القياس والتقويم التربوي.

وكذلك من خلال توجيه اهتمام الباحثين والمهتمين بالمجال التربوي إلى استخدام نظرية إمكانية التعميم في تقدير ثبات (وحتى صدق) أساليب تقييم الكفاءة، وكيفية تصميم هذا النوع من الدراسات وتفسير مصادر تباين الخطأ، ومعاملات إمكانية التعميم النسبية والمطلقة، ومعرفة إجراءات تحسين القياسات في المستقبل.

بالإضافة إلى إمكانية الاستفادة من نتائج هذه الدراسة عن طريق تشجيع المدرسين على تقديم محتوى نظرية إمكانية التعميم للطلبة على مستوى كافة التخصصات التربوية، وعلى إجراء بحوث على مستوى الماجستير والدكتوراه في مختلف التخصصات التربوية للاستفادة منها في إعداد اختبارات وفق هذه النظرية، الأمر الذي سينعكس ايجابًا على جودة أدوات القياس والتقييم المستخدمة في البحوث التربوية أو في الممارسة العملية.

التعريفات الاجرائية

- **الثبات:** هو مدى اتساق درجات الطلاب في الاختبار عبر المهمات الاختبارية والمقدرين، ويعبّر عنه في الدراسة الحالية بمعامل إمكانية التعميم النسبي الذي يشير إلى اتساق درجات الطلاب عبر المهمات والمقدرين لتحديد المكانة النسبية للطلاب بين زملائهم، ومعامل إمكانية التعميم المطلق الذي يشير إلى اتساق درجات الطلاب عبر المهمات والمقدرين بهدف مقارنة أدائهم بمحك أداء خارجي.
- **نظرية إمكانية التعميم:** هي مختلف الإجراءات والأساليب الإحصائية المتبعة في الدراسة الحالية لتقدير مصادر تباين الخطأ الموضحة في الجدول رقم (1) ومعاملات إمكانية التعميم النسبية والمطلقة (الموضحة في الصيغتين رقم 3 و4)، والطرق المتبعة في تحسين ثبات الاختبار بزيادة عدد المهمات وزيادة عدد المقدرين.
- **مصادر تباين الخطأ:** هي تقديرات كمية لتباين متوسط درجات الطلاب عبر المهمات والمقدرين، ويعبّر عنها في الدراسة الحالية بواسطة قيم ونسب تباين الخطأ النسبي وتباين الخطأ المطلق التي تتضمن ست مصادر للتباين: تباين المهمة، تباين المقدر، تباين تفاعل الطالب مع المهمة، وتباين تفاعل الطالب مع المقدر،

تفاعل المهمة مع المقدر، تفاعل الطالب مع المهمة والمقدر الممزوج بالأخطاء العشوائية غير المقدرة في التصميم.

- **اختبار تقييم الكفاءة:** هو مجموعة مكوّنة من تسع مهمات اختبارية أعدت في مجال الأعداد والحساب بهدف حلّها من طرف الطالب، ويتم تقدير كفاءته بواسطة سلم تقدير تفصيلي تتراوح درجاته في كل مهمة بين صفر وأربع درجات.
- **المهمة المعقدة:** هي مشكلة مركبة يجب على الطالب الاجابة عنها وفقًا لما هو مطلوب في العبارة (أو السؤال) التي تتضمنها لإثبات معرفته ومهارته في مجال الأعداد والحساب، وقد أعدّت المهمة المعقّدة وفق ثلاث صيغ: مهمة محكمة البناء، ومهمة غير محكمة البناء، ومهمة ذات معلومات مشوشة.

الطريقة

منهج الدراسة

بما أن الدراسة ذات طبيعة سيكومترية اهتمت بتقدير ثبات اختبار لتقييم كفاءة الطلاب باستخدام نظرية إمكانية التعميم، فإن أغلب هذا النوع من الدراسات في الأصل وصفية استكشافية، فالمنهج المستخدم في هذه الدراسة وصفي يهدف إلى الكشف عن أحجام مصادر تباين الخطأ، وتقدير معاملات إمكانية التعميم النسبية والمطلقة، بالإضافة إلى الكشف عن الأبعاد الأكثر كفاءة في رفع ثبات درجات الاختبار.

عينة الدراسة

تكوّنت عينة الدراسة من (331) طالبًا من طلاب السنة الرابعة ابتدائي في مقاطعة سطيف بالجزائر للعام الدراسي 2014/2013، وقد تم اختيار ست مدارس بطريقة عشوائية بسيطة من بين (10) مدارس اشتملت عليها المقاطعة، وتراوحت نسب تمثيل الطلاب في كل مدرسة بين (9-20%) من مجموع المدارس التي أجريت عليها الدراسة. وفي أثناء اختيار المدارس، تم الحصر الشامل لكل الطلاب الذين بلغ عددهم (358) طالبًا، وقد انخفض عدد الطلاب إلى (331) في العينة النهائية بسبب غياب (27) طالبًا في إحدى المرات الثلاث التي طبقت فيها المهمات.

أداة الدراسة

استخدم الباحث في الدراسة الحالية اختبارًا تحصيليًا مكوّن من تسع مهمات مركبة في مجال الأعداد والحساب في الرياضيات، وهي مشكلات معقّدة تسمح بقياس كفاءة "حل مشكلات في ميدان الأعداد والحساب تتعلق بتعيين الأعداد، ومقارنتها، وترتيبها، والحساب عليها لدى طلاب الرابع ابتدائي" (Ministry of National Education, 2011).

وتوزعت كل المهمات التي احتوى عليها الاختبار على نفس مجال محتوى الرياضيات، وهو مجال الأعداد والحساب، وتقيس المهمات كفاءة الطلاب في إنجاز العمليات الحسابية الأربع (الجمع، والضرب، والقسمة، والطرح) ومقارنة الأعداد، وتوزعت وفق ثلاث صيغ: ثلاث مهمات محكمة البناء، وثلاث مهمات غير محكمة البناء، وثلاث مهمات ذات معلومات مشوّشة حتى تكون شاملة لأنواع المهمات المعقدة.

وتشتمل المهمات محكمة البناء على كل المعلومات الضرورية للحل، ويكون العمل المطلوب من الطالب واضحًا، وتشتمل المهمات غير محكمة البناء أيضًا على كل المعلومات الضرورية، إلا أن العمل المطلوب غير بديهي وغير واضح لدى الطلاب، وتشتمل المهمات ذات المعلومات المشوّشة على معلومات ضرورية للحل ومعلومات إضافية أخرى غير ضرورية للحل. وقد وُضع سياق كل صيغة من صيغ المهمات في سياق واقعي مألوف، حيث يمكن أن يصادفها الطالب في حياته اليومية.

استعان الباحث في إعداد المهمات على مجموعة من معلمي ومفتشي التعليم الابتدائي بمراعاة الشروط الأساسية، وهي أن تكون واضحة الصياغة، وجديدة (لم يصادفها الطالب من قبل)، ومعقّدة (مركبة) وذات سياق عملي (أن تكون المهمة واقعية)، وتتضمن المهمة السند (وثيقة، أو معلومات، أو إحصائيات...)، والسياق (بناء المشكلة)، والتعليمة (السؤال).

اعتمد الباحث في التحقق من صدق محتوى الاختبار على أحكام الخبراء حول انسجام محتوى الاختبار مع المفهوم المقاس، ومدى تمثيل المحتوى لمجال المفهوم المقاس، وجودته الفنية (Messick, 1995)، وذلك بالإجابة عن مجموعة من الشروط المطلوبة في المهمات، والتي جُمعت في "قائمة مراجعة" استوحيت من خصائص المهمات المعقدة (Gerard, 2006; Roegiers, 2004; Scallon, 2004)، بحيث يقوم المحكمون بفحص مدى توفر هذه الشروط في كل مهمة من مهمات الاختبار على حدة.

استُخدم في التحقق من صدق محتوى الاختبار على معاملات الاتفاق بين (16) محكمًا ذوي خبرة، منهم خمسة مفتشين للتعليم الابتدائي، وستة معلمين يدرسون في الابتدائي، وخمسة أساتذة جامعيين، حيث يعطي للمحكم (1) للإجابة بـ "نعم"، و(0) للإجابة بـ "لا" على كل شرط، وتم تقدير معاملات الاتفاق بين المحكمين بقسمة عدد اتفاقات المحكمين على مجموع الاتفاقات والاختلافات في كل شرط من الشروط المطلوبة في المهمات، وقُدّرت معاملات الاتفاق في كل مهمة، وكشفت أن معظم هذه المعاملات تتعدى (0.80)، وفي بعض الشروط المطلوب توافرها في المهمات تم الحصول على معاملات اتفاق منخفضة تراوحت بين (0.56 - 0.75)، وعلى أساسها أجريت التعديلات المقترحة من طرف المحكمين.

أُعدت لكل مهمة من مهمات الاختبار سلم تقدير تفصيلي (Analytic Scoring Rubric)، وذلك بتجزئة نتيجة الطالب إلى عناصر منفصلة وإعطاء تقدير لكل عنصر منها (Boston, 2002;

(Gerard, 2006; Roegiers, 2000; Smit & Birri, 2014)، ويشتمل كل سلم تقدير أربعة معايير لإعطاء الدرجات، وهي:

- **الاستخدام الملائم للعمليات الحسابية:** يشير إلى استخدام الطالب للعمليات الحسابية بشكل صحيح (جمع، طرح، ضرب، قسمة، مقارنة) حتى وإن أخطأ في استخدام الأعداد المناسبة في الحل.
- **الاستخدام الملائم للأعداد:** يُعبّر عن استخدام الطالب للأعداد المناسبة التي تتضمنها المهمة في الحل حتى وإن أخطأ في اختيار العمليات المناسبة.
- **صحة الحسابات:** يشير إلى دقة الحسابات التي توصل إليها الطالب في الحل حتى وإن أخطأ في اختيار العمليات أو اختيار الأعداد المناسبة.
- **انسجام الإجابة مع السؤال:** يشير إلى ملاءمة الإجابة التي قدمها الطالب مع السؤال المطلوب في المهمة، واستخدام وحدة القياس المناسبة.

وتم التحديد الاجرائي لكل معيار بواسطة مجموعة من المؤشرات التي تدل على تحققه في كل مهمة، حيث يُعطي المقدّر الدرجة (1) إذا نجح الطالب في المعيار من خلال إجاباته على أكثر من (70%) من المؤشرات إجابة صحيحة، ويُعطي المقدّر الدرجة (0) إذا لم ينجح الطالب في الإجابة على (70%) من المؤشرات إجابة صحيحة، ويتحصل الطالب في كل مهمة على درجة كلية تتراوح بين (0-4) درجات.

وبعد التحقق من صدق محتوى الاختبار، قام الباحث بتجريب كل المهمات على عينة أولية من طلاب الرابع ابتدائي مكوّنة من (30) طالبًا بهدف التحقق من وضوح التعليمات ومناسبة الوقت، ومن خلالها تم التأكد من وضوح محتوى المهمات، وتعليماتها، وملاءمة زمن الإجابة عليها.

الإجراءات

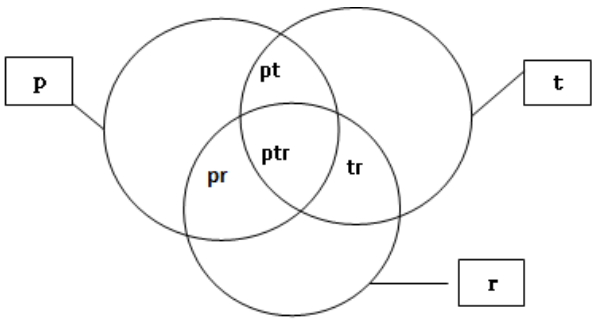
طبّق الباحث الاختبار على أفراد عينة الدراسة خلال الفترة الممتدة من 15 نيسان إلى 30 أيار 2014، وفي الفترة الصباحية في مدة (50) دقيقة بتقديم ثلاث مهمات ممزوجة (محكمة البناء، وغير محكمة البناء، وذات مهمة مشوشة) بمتوسط (15) دقيقة للمهمة الواحدة، وتم مراعاة نفس إجراءات تطبيق كل مجموعة من مهمات الاختبار.

وقبل بداية الاختبار، تم تقديم تعليمات حول الهدف من الاختبار، ومحتواه، وطريقة الإجابة عليه، والمدة الزمنية للإجابة عنه بهدف ضمان انغماس الطلاب في إنجاز المهمات، كما تمت مراعاة عدم تقديم أيّة مساعدة للطلاب لضمان ذاتية الإجابة، وتفادي التحيز لفئة معينة. وبعد استكمال تطبيق مهمات الاختبار عيّن الباحث ثلاثة مقدرين ذوي خبرة في التدريس في الطور الابتدائي لتقييم نتائج الطلاب، وبعد تعيينهم تم إخضاعهم لبرنامج تدريبي، حيث خُصّصت لكل مقدر ثلاث ساعات منفصلة لشرح إجراءات التقدير مع إعطاء أمثلة لنفس المهمات، وذلك بطريقة مفصلة مع متابعة الباحث لعملية التصحيح من أجل ضمان اتساق التقديرات.

تصميم الدراسة

تم تحليل بيانات الدراسة من خلال تصميم ثنائي البُعد (Two-facet design) متقاطع كليًا (Fully Crossed)، ويتصف هذا التصميم بأنه ثنائي البُعد لأنه يشتمل على بُعدين محتملين لتباين الخطأ، هما: المهمات والمقّدرين، وتجدر الإشارة أن الطلاب لا يمكن اعتبارهم بُعدا للقياس وإنما موضوعًا للقياس (Brennan, 1992)، والذي يشير إلى التباين الحقيقي الذي يعكس الفروق بين الطلاب في الكفاءة المقيسة، وبالتالي فهم لا يمثّلون مصدرًا محتملاً لتباين الخطأ.

ووُصف التصميم بأنه متقاطع كليًا، لأن كل الطلاب يجيبون عن كل مهمة من مهمات الاختبار، ويتم تقييم كل الطلاب من طرف كل مقدّر من المقدرين، كما أن التصميم الذي استخدم في الدراسة الحالية يشتمل على الأفراد (الطلبة) Persons (p) والمهمات Tasks (t) والمقدرين Raters (r)، ويُشار إلى التصميم بالصيغة "طالب × مهمة × مقدر"، وباختصار يرمز له بالرمز (p×t×r)، بحيث يتقاطع الطلاب مع المهمات ومع المقدرين (الذي يشير إليه الرمز ×)، والشكل (1) يوضح تصميم جمع بيانات الدراسة.



الشكل (1): رسم توضيحي لتصميم الدراسة "طالب × مهمة× مقدر"

يشتمل بُعد الطلاب (p) على (331) مستوى (P_{331}-P_1)، وبُعد المهمات على (9) مستويات (t_9-t_1)، واشتمل بُعد المقدّرين على (3) مستويات (r_3-r_1)، واعتبرت عينة الطلاب، ومهمات الاختبار، والمقدّرين أبعادًا "عشوائية غير محدودة" اختيرت من نطاق واسع من الطلبة، والمهمات، والمقدّرين، لأنه تتوفر عينة واسعة من الطلاب الذين يمكن اختيارهم، وعينة واسعة من المهمات التي يمكن اختيارها أو إعدادها من طرف الباحث، وعينة واسعة من المقدرين الذين يمكنهم تقييم كفاءة الطلاب في الاختبار، وهذا يتماشى مع نظرية إمكانية التعميم التي طوّرت في

الأصل كنظرية للتأثيرات العشوائية، أي أنها تعتبر الأبعاد عشوائية اختيرت من نطاق واسع (Shavelson & Webb, 2009).

يهدف الاختبار إلى التمييز بين الطلاب (p) من حيث كفاءتهم في الأعداد والحساب باستخدام مهمات التقييم (t) والمقدرين (r)، فقد تم اختيار تصميم قياس اعتبر فيه الطلاب موضوعًا للقياس يسهم في تباين الدرجة الشاملة أو الحقيقية، أما المهمات والمقدرون والتفاعلات بينها وبين الطلاب اعتبرت مصادر للتباين تسهم في تباين خطأ القياس، ويُشار إلى تصميم القياس "طلبة/مهمات مقدّرين" الذي يُرمز له "p/tr". ومن الناحية الإحصائية يتم تقدير أحجام مصادر تباين الخطأ الراجعة إلى أبعاد القياس التي تُعرف أيضا بمكونات التباين (Variance Components) باستخدام متوسطات المربعات الناتجة عن تحليل التباين، كما يوضحه الجدول رقم (1).

جدول (1): صيغ تقدير مكونات التباين في دراسة إمكانية التعميم للتصميم ثنائي البُعد (p × t × r)

مصدر التباين	مكون التباين	تقديرات مكونات التباين
الطالب (p)	$\sigma^2 p$	$\frac{MS_p - MS_{pt} - MS_{pr} + MS_{ptr,e}}{n_t n_r}$
المهمة (t)	$\sigma^2 t$	$\frac{MS_t - MS_{tr} - MS_{pt} + MS_{ptr,e}}{n_p n_t}$
المقدر (r)	$\sigma^2 r$	$\frac{MS_r - MS_{tr} - MS_{pr} + MS_{ptr,e}}{n_p n_t}$
تفاعل طالب-مهمة (pt)	$\sigma^2 pt$	$\frac{MS_{pt} - MS_{ptr,e}}{n_r}$
تفاعل طالب-مقدر (pr)	$\sigma^2 pr$	$\frac{MS_{pr} - MS_{ptr,e}}{n_t}$
تفاعل مهمة-مقدر (tr)	$\sigma^2 tr$	$\frac{MS_{tr} - MSp_{ptr,e}}{n_p}$
الباقي: تفاعل طالب-مهمة-مقدر (ptr,e)	$\sigma^2 ptr,e$	$\sigma^2 ptr,e = MS_{ptr,e}$

المرجع: (Brennan, 1992)

حيث يشير MS إلى متوسط المربعات (Mean of Squares) المحصلة من جدول تحليل التباين، أما n فيشير إلى عدد مستويات (عينة) البُعد، حيث يشير: n_p إلى عدد الطلاب، و n_t إلى عدد المهمات، و n_r إلى عدد المقدّرين.

ومن أجل تقييم ثبات درجات الاختبار في مرحلة دراسات القرار، تم تقدير معامل إمكانية التعميم النسبي الذي يهدف إلى تحديد دقة الاختبار في تحديد المكانية النسبية للطلاب (تحديد رتبة الطالب بين زملائه)، وتقدير معامل إمكانية التعميم المطلق الذي يهدف إلى تحديد مستوى أداء الطلاب بالنسبة لمحك خارجي، وقبل تقدير معاملات إمكانية التعميم تم تقدير تباين خطأ القياس النسبي وتباين خطأ القياس المطلق باستخدام الصيغتين التاليتين: (Webb, Shavelson & Steedle, 2006)

(1) تباين الخطأ النسبي:

$$\widehat{\sigma}^2_{(\delta)} = \frac{\widehat{\sigma}^2_{\boldsymbol{pt}}}{n_t} + \frac{\widehat{\sigma}^2_{pr}}{n_r} + \frac{\widehat{\sigma}^2_{ptr,e}}{n_t n_r}$$

(2) تباين الخطأ المطلق:

$$\widehat{\sigma}^2_{(\Delta)} = \frac{\widehat{\sigma}^2_{t}}{n_t} + \frac{\widehat{\sigma}^2_{r}}{n_r} + \frac{\widehat{\sigma}^2_{tr}}{n_t n_r} + \frac{\widehat{\sigma}^2_{pt}}{n_t} + \frac{\widehat{\sigma}^2_{pr}}{n_r} + \frac{\widehat{\sigma}^2_{ptr,e}}{n_t n_r}$$

حيث يشير $\widehat{\sigma}^2_{(\delta)}$ إلى تباين الخطأ النسبي، و$\widehat{\sigma}^2_{(\Delta)}$ إلى تباين الخطأ المطلق، ويشير كل رمز من رموز البسط في الصيغتين إلى مكونات التباين، حيث يرمز $\widehat{\sigma}^2_{t}$ إلى تباين المهمة، و $\widehat{\sigma}^2_{r}$ إلى تباين المقدر، و $\widehat{\sigma}^2_{tr}$ إلى تباين تفاعل المهمة مع المقدر، و$\widehat{\sigma}^2_{pt}$ إلى تباين تفاعل الطالب مع المهمة، و $\widehat{\sigma}^2_{pr}$ إلى تباين تفاعل الطالب مع المقدر، $\widehat{\sigma}^2_{ptr,e}$ إلى تباين تفاعل الطالب مع المهمة ومع المقدر الممزوج بالأخطاء العشوائية غير المقدرة في التصميم. أما رموز المقام فتشير في الصيغتين إلى عدد مستويات (عينة) الأبعاد، حيث يشير n_t إلى عدد المهمات، و n_r إلى عدد المقدرين.

وبناء على قيمة تباين الخطأ النسبي، وقيمة تباين الخطأ المطلق تم تقدير معامل إمكانية التعميم النسبي $\hat{\rho}^2$، ومعامل إمكانية التعميم المطلق Φ وفق الصيغة (3) والصيغة (4): (Meyer, 2010)

معامل إمكانية التعميم النسبي:

$$\hat{\rho}^2 = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

معامل إمكانية التعميم المطلق:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\Delta)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

يشير $\hat{\rho}^2$ إلى تباين الدرجة الشاملة الذي يعكس الفروق الفردية بين الطلاب في المهمات والمقدرين على حد سواء، ويشير

$\sigma^2(\delta)$ في الصيغة (3) إلى تباين الخطأ النسبي الذي يتم تقديره من الصيغة (1)، ويشير $\sigma^2(\Delta)$ في الصيغة (4) إلى تباين الخطأ المطلق الذي يتم تقديره من الصيغة (2).

وفي دراسات القرار اعتمدت الدراسة الحالية على زيادة عدد المهمات من (9-12) مهمة، وزيادة عدد المقدرين من (3-6) مقدرين، بهدف خفض أحجام مصادر تباين الخطأ لبلوغ مستوى مقبول من الثبات، وفي هذه المرحلة تكون معاملات إمكانية التعميم وتباين الأخطاء النسبية والمطلقة غير واقعية، وإنما قيم افتراضية أو متوقعة يتم الحصول عليها في حالة تغيير عدد المقدرين و/ أو عدد المهمات.

تم تحليل بيانات الدراسة باستخدام برمجية (EduG) التي تعتمد على تحليل التباين وتحليلات إمكانية التعميم، وصُمّمت من طرف كادريني (Cardinet) بمساعدة برتراند (Bertrand)، وقدمتها الجمعية السويسرية للبحث التربوي (Cardinet et al., 2010; Swiss Society for Research in Education Working Group, 2010)، حيث تسمح البرمجية من خلال دراسات إمكانية التعميم بتقدير حجم التباين الحقيقي، وحجم تباين الخطأ الذي يفسره كل بُعد من الأبعاد في التباين الكلي، كما تسمح بإجراء دراسات القرار للتعرف على أفضل الشروط التي تسمح بالحصول على مستويات أفضل من الثبات في الدراسات المستقبلية، وتتيح البرمجية استخدامات مبدأ التماثل (Symmetry) في نظرية إمكانية التعميم الذي يُعطي إمكانية لأي بُعد أن يكون موضوعًا للقياس بغض النظر عن الأفراد (Cardinet, Tourneur, & Allal, 1976).

النتائج

أولا: النتائج المتعلقة بالسؤال الأول: *ما مقدار تباين الخطأ الذي يفسّره كل من أبعاد (المهمات، والمقدرين) في التباين الكلي لدرجات طلاب الرابع ابتدائي في اختبار كفاءة الأعداد والحساب في الرياضيات*؟

للإجابة عن هذا السؤال، تم تقدير مصادر تباين الخطأ ومعاملات إمكانية التعميم النسبية والمطلقة للتصميم ثنائي البُعد: "طالب× مهمة× مقدر" كما هو موضح في الجدول رقم (2)، والذي يتضمن سبعة مكونات تباين راجعة إلى الطلاب (p) (تباين الدرجة الشاملة)، والمقدرين (r)، والمهمات (t)، وتفاعل الطالب مع المهمة (pt)، وتفاعل الطالب مع المقدر (pr)، والباقي الذي يعكس تفاعل الطالب مع المهمة ومع المقدر الممزوج بالأخطاء العشوائية غير المقدرة في التصميم (ptr,e). مع العلم أن الأخطاء العشوائية غير المقدرة في التصميم قد تكون راجعة إلى أبعاد أخرى لم يشتمل عليها تصميم الدراسة الحالية، مثلاً: فترة تطبيق الاختبار أو صيغة المهمات أو معايير تصحيح المهمات، ومصادر تباين الخطأ الراجعة إلى هذه الأبعاد وتفاعلاتها تُدمج في الباقي Residual الذي يُشار إليه (e).

جدول (2): تحليل التباين للتصميم المتقاطع كليا "طالب × مهمة × مقدر"(p × t × r)

مصدر التباين	مجموع المربعات	درجات الحرية	متوسط المربعات	المكون	النسبة (%)	الخطأ المعياري
طالب (p)	5046.504	330	15.292	0.449	22.5	0.044
مهمة (t)	2280.268	8	285.033	0.277	13.9	0.128
مقدر (r)	207.120	2	103.560	0.032	1.6	0.024
طالب- مهمة (p×t)	7659.0.46	2640	2.901	0.855	42.9	0.026
طالب- مقدر(p×r)	381.102	660	0.577	0.027	1.4	0.003
مهمة- مقدر (t×r)	103.609	16	0.475	0.018	0.9	0.006
الباقي (طالب- مهمة- مقدر) (p×t×r,e)	1760.168	5280	0.333	0.333	16.7	0.006
المجموع	17437.837	8935			%100	

وعلى اعتبار أن اهتمام الدراسة الحالية ينصب على تأثير مختلف مصادر تباين الخطأ على درجات الطلاب، فإنه تم الاعتماد على تصميم قياس يعتبر فيه الطلاب (p) كبُعد لتباين الدرجة الشاملة، والمهمة (t)، والمقدر (r) كأبعاد لتباين خطأ القياس، والجدول رقم (3) يعرض نسب تباين الخطأ النسبي، ونسب تباين الخطأ المطلق لكل بُعد من أبعاد تصميم القياس وتفاعلاتها فيما بينها.

جدول (3): تحليل إمكانية التعميم لتصميم القياس "طلاب/مهمات مقدرين" (p/tr)

مصدر التباين	تباين التمييز	تباين الخطأ النسبي	نسبة تباين الخطأ النسبي	تباين الخطأ المطلق	نسبة تباين الخطأ المطلق
طالب (p)	0.449	.....	.....	.....	.....
مهمة (t)	.....	.....	.....	0.0308	19.4
مقدر (r)	.....	.....	.....	0.0108	6.8
طالب- مهمة (pt)	.....	0.095	81.9	0.095	59.9
طالب- مقدر (pr)	.....	0.0090	7.8	0.0090	5.7
مهمة- مقدر (tr)	.....	.....	.....	0.0006	0.38
طالب- مهمة-مقدر (ptr,e)	.....	0.0123	10.6	0.0123	7.8
مجموع التباينات	0.449	0.116	100	0.158	100
الانحراف المعياري	0.670	الخطأ المعياري النسبي	0.341	الخطأ المعياري المطلق	0.398
معامل إمكانية التعميم النسبي			0.79		
معامل إمكانية التعميم المطلق			0.74		

يبيّن الجدول رقم (3) أن معامل إمكانية التعميم النسبي (0.79) ومعامل إمكانية التعميم المطلق (0.74) لم يصلا إلى الحد الأدنى المطلوب (0.80) (Bain & Pini, 1996)، وذلك رغم أنها قريبة من الحد الأدنى لمعامل إمكانية التعميم النسبي، وتجدر الاشارة إلى أن تقييم الكفاءة لأغراض اتخاذ قرارات نهائية حول الطلاب تتطلب مستويات مقبولة من الثبات أو إمكانية التعميم سواء لاتخاذ قرارات نسبية (مقارنة أداء الطالب بزملائه) أو قرارات مطلقة (مقارنة أداء الطالب بالأهداف المرجوة أو بمحك خارجي).

ويُلاحظ من الجدول (3) أن أكبر مصدر رئيسي للخطأ (النسبي والمطلق) راجع إلى التفاعل بين الطالب والمهمة بنسبة (81.9%) من تباين الخطأ النسبي، و(59.9%) من تباين الخطأ المطلق، والذي يعكس التغيّر في متوسط أداء الطلاب من مهمة إلى أخرى، بمعنى أن بعض الطلاب حصلوا على درجة عالية في مهمة، في حين حصلوا على درجات منخفضة في مهمة أخرى. وأهم مصدر للخطأ (المطلق) الذي يأتي بعد تفاعل الطالب مع المهمة هو التأثير الرئيسي لمهمات الاختبار (19.4%) من تباين الخطأ المطلق، والراجع إلى التغيّر أو الاختلاف في مستوى صعوبة المهمات بالنسبة لكل الطلاب.

ويتضح من الجدول رقم (3) أن بعض مهمات التقييم تبدو أكثر صعوبة من الأخرى، فعلى سبيل المثال يتبيّن أن مدى تباين صعوبة المهمات بين المهمة الأكثر صعوبة المهمة 6 = (0.279) والمهمة الأكثر سهولة المهمة 7 = (1.99) كان كبيرًا (1.711) درجة علمًا بأن أقصى مدى ممكن في كل مهمة (4.00).

جدول (4): الإحصاءات الوصفية لمهمات الاختبار

نوع المهمات	أرقام المهمات	المتوسط الحسابي	الانحراف المعياري	التباين
مهمات محكمة البناء	المهمة 1	0.961	1.395	1.947
	المهمة 2	0.725	1.283	1.646
	المهمة 3	0.309	0.655	0.430
مهمات غير محكمة البناء	المهمة 4	0.750	1.135	1.290
	المهمة 5	1.207	1.682	2.831
	المهمة 6	0.279	0.668	0.447
مهمات ذات معلومات مشوّشة	المهمة 7	1.99	1.640	2.690
	المهمة 8	1.188	1.501	2.254
	المهمة 9	1.379	1.508	2.275

بالنسبة للأثر الرئيسي للمقدر (6.8%) من تباين الخطأ المطلق والتفاعل بين الطالب والمهمة والمقدر الممزوج بالأخطاء العشوائية غير المقاسة في التصميم (10.6%) من تباين الخطأ النسبي و (7.8%) من تباين الخطأ المطلق، والتفاعل بين الطالب والمقدر (7.8%) من تباين الخطأ النسبي و (5.7%) من تباين الخطأ المطلق منخفضة مقارنة بمصدر تباين تفاعل الطالب مع المهمة والأثر الرئيسي للمهمات، ولكنها جديرة بالاهتمام.

ويعكس تفاعل الطلاب مع المهمات والمقدّرين الممزوج بالأخطاء العشوائية غير المقيسة في التصميم التغير في أداء الطلاب وترتيبهم عبر مختلف المهمات ومختلف المقدرين، ويعكس التأثير الرئيسي للمقدر التغير أو الاختلاف في درجة تساهل أو تشدّد المقدرين في تصحيح المهمات، كما يعكس تفاعل الطلاب مع المهمات التغير في درجات الطلاب من مقدر إلى آخر. أما تفاعل المقدر مع المهمة (0.38%) من تباين الخطأ المطلق منخفض جدًا وجدير بالإهمال.

ويتضح من نتائج الجدول رقم (4) بأن تقديرات المصححين متقاربة نسبيًا، بحيث لوحظ فارق صغير نسبيًا (0.398) درجة بين المقدر الأكثر تشدّدًا المقدر3 = (0.822) والمقدر الأقل تشدّدا المقدر1 = (1.22) علمًا بأن أقصى مدى ممكن (4.00).

جدول (5): الإحصاءات الوصفية للمقدرين

أرقام المقدرين	المتوسط الحسابي	الانحراف المعياري	التباين
المقدر 1	1.220	1.480	2.191
المقدر 2	0.888	1.391	1.936
المقدر 3	0.822	1.353	1.831

ثانيًا: النتائج المتعلقة بالسؤال الثاني: *ما شروط تحقيق اختبار كفاءة طلاب الرابع ابتدائي في الأعداد والحساب في الرياضيات لأفضل مستويات الثبات؟*

للإجابة عن السؤال الثاني، تم زيادة عدد المهمات وعدد المقدرين بهدف خفض مصادر تباين الخطأ ورفع معاملات إمكانية تعميم درجات الاختبار، وذلك عن طريق تجريب زيادة عدد المهمات من (9-12) وعدد المقدرين (3-6)، وكما في الجدول رقم (6) الذي يوضح معاملات إمكانية التعميم النسبية والمطلقة وتباينات الخطأ النسبية والمطلقة لكل خيارات دراسات القرار بناء على دراسات إمكانية التعميم الأولية.

جدول (6): دراسات القرار للتصميم ثنائي البُعد "طالب × مهمة × مقدر"

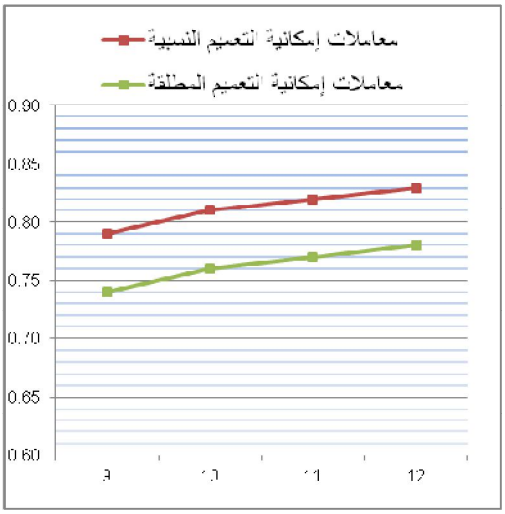
عدد المهمات	9	10	11	12	9	9	9
عدد المقدرين	3	3	3	3	4	5	6
$\mathbf{Ep\hat{}^2}$	0.79	**0.81**	0.82	0.83	**0.80**	0.81	0.81
$\boldsymbol{\hat{\sigma}^2(\delta)}$	0.117	0.106	0.097	0.089	0.111	0.108	0.106
$\Phi\hat{}$	0.74	0.76	0.77	0.78	0.75	0.76	0.76
$\boldsymbol{\hat{\sigma}^2(\Delta)}$	0.156	0.145	0.137	0.124	0.151	0.146	0.142

ملاحظة: $E\hat{\rho}^2$: معامل إمكانية التعميم النسبي المتوقع، $\hat{\sigma}^2(\delta)$: تباين الخطأ النسبي المتوقع، $\hat{\sigma}^2(\Delta)$: تباين الخطأ المطلق المتوقع، $\widehat{\Phi}$: معامل إمكانية التعميم المطلق المتوقع.

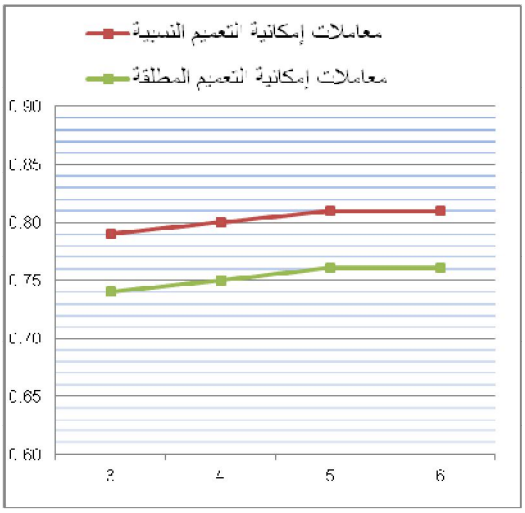
من خلال متابعة دراسة إمكانية التعميم بدراسات القرار عن طريق زيادة عدد المهمات وعدد المقدّرين، أظهرت النتائج أنه يجب زيادة عدد مهمات التقييم إلى (10) على الأقل لضمان ثبات القياس النسبي (معامل إمكانية تعميم نسبي = 0.81)، ورفع عدد المهمات إلى أكثر من 12 مهمة (14 مهمة) باستخدام (4) مقدرين لضمان ثبات القياس المطلق (معامل إمكانية التعميم المطلق = 0.80).

ومن ناحية أخرى، أظهرت نتائج تحليلات زيادة عدد المقدرين أنه يجب استخدام (4) مقدرين على الأقل في (9) مهمات لضمان معامل إمكانية تعميم نسبي مقبول (0.80)، في حين لا يمكن بلوغ المستوى المطلوب من معامل إمكانية التعميم المطلق حتى عند زيادة عدد المقدرين إلى أقصى حد ممكن، فرفع عدد المقدرين إلى 14 (وهو عدد هائل بالمقارنة بعدد المقدرين المستخدمين في تصميم الدراسة) يبقى معامل إمكانية التعميم المطلق عند مستوى (0.77).

ويوضح الشكلان (2) و(3) أن معاملات إمكانية التعميم في حال زيادة عدد المهمات من (9-12) أعلى من معاملات إمكانية التعميم في حال زيادة (عدد المقدرين من 3 - 6)، بحيث بلغ معامل إمكانية التعميم النسبي عند زيادة عدد المهمات إلى 12 (0.83) ومعامل إمكانية التعميم المطلق (0.78). في حين بلغ معامل إمكانية التعميم النسبي عند زيادة المقدرين إلى 6 (0.81) ومعامل إمكانية التعميم المطلق (0.76). وفي حال زيادة عدد المقدرين من (5 - 6) استقرت معاملات إمكانية التعميم النسبية والمطلقة عند (0.81 و0.76) على التوالي.



الشكل (2): معاملات إمكانية التعميم بعد زيادة عدد المهمات من (9-12)



الشكل (3): معاملات إمكانية التعميم بعد زيادة عدد المقدرين (3-6)

وهكذا تسهم زيادة عدد المهمات بشكل أفضل في رفع إمكانية تعميم وموثوقية درجات الاختبار من زيادة عدد المقدرين، وتصبح زيادة المهمات أفضل طريقة لتحقيق مستويات ثبات اختبار كفاءة الأعداد والحساب في الرياضيات لدى طلاب الرابع ابتدائي، فإذا تم الأخذ بعين الاعتبار ممارسات التقييم في المدرسة الجزائرية فإن السؤال المطروح: ما مستوى ثبات الاختبار إذا استخدم مقدر واحد فقط وليس ثلاثة مقدرين لتصحيح كل المهمات؟، فإنّه يُسجّل انخفاضًا في معامل إمكانية التعميم النسبي إلى (0.74)، ومعامل إمكانية التعميم المطلق إلى (0.67)، وما مستوى ثبات الاختبار إذا استخدمت ثلاث مهمات فقط وليس تسع مهمات وثلاث مقدرين لاعتبارات متعلقة بالجهد والتكلفة والوقت؟ فإنّه يُسجّل انخفاضًا أيضًا في معامل إمكانية التعميم النسبي إلى (0.58) وفي معامل إمكانية التعميم المطلق إلى (0.51).

وأكدت تحليلات إمكانية التعميم بأن درجات الطلاب في الاختبار باستخدام (9) مهمات و(3) مقدّرين غير قابلة للتعميم كفاية على مهمات تقييم أخرى وعلى مقدرين آخرين، وإن كانت النتائج مشجعة باعتبار أن قيمة معامل إمكانية التعميم النسبي (0.79) وقيمة معامل إمكانية التعميم النسبي (0.74)، إلا أنها لم تصل إلى الحدّ الأدنى المطلوب (0.80)، كما أن زيادة عدد المهمات ترفع من معاملات إمكانية التعميم النسبية والمطلقة أفضل من زيادة عدد المقدرين، وهذا كما أشرنا راجع إلى أثر تفاعل الطالب مع المهمة، والتأثير الرئيسي للمهمة على ثبات درجات الاختبار.

مناقشة النتائج

توصلت الدراسة إلى ثلاث نتائج رئيسية:

1) وجود مصدرين رئيسين للخطأ راجعين إلى تأثير تفاعل الطالب مع المهمة، وإلى التأثير الرئيسي للمهمة على ثبات درجات الاختبار.

2) عدم بلوغ ثبات درجات اختبار تقييم كفاءة الطلاب في الرياضيات (الأعداد والحساب) مستوى إمكانية التعميم (الثبات) المطلوب،

3) للحصول على معاملات إمكانية تعميم نسبية ومطلقة مقبولة نحتاج إلى 10 و14 مهمة على التوالي، ونحتاج إلى 4 مقدرين للحصول على معامل إمكانية تعميم نسبي مقبول، في حين لا يمكن بلوغ معامل إمكانية تعميم مطلق مقبول حتى عند زيادة عدد المقدرين إلى حد أكبر.

وأظهرت نتائج الدراسة بأن مصادر الخطأ التي أثّرت بشكل أكبر على ثبات درجات الاختبار راجعة إلى تأثير تفاعل الطالب مع المهمة، والتأثير الرئيسي للمهمة، وقد قدّرت نسبة تفاعل الطالب مع المهمة (81.9%) من تباين الخطأ النسبي و (59.9%) من تباين الخطأ المطلق، ويرجع ارتفاع تفاعل الطالب مع المهمة إلى اختلاف أو تغيّر في متوسط أداء الطالب من مهمة إلى أخرى، بمعنى حدوث تذبذب في أداء نفس الطلاب للمهمات نفسها التي تختلف فيما بينها من حيث صياغتها، والذي يمكن أن يرجع بدوره إلى تغيير الطالب لاستراتيجيات المعالجة من مهمة لأخرى سواء بسبب الاختلاف في سياق المهمة أو في طريقة بنائها (محكمة البناء، غير محكمة البناء، ذات معلومات مشوّشة) أو في مستوى صعوبتها. وتتفق هذه النتيجة مع نتائج العديد من الدراسات السابقة التي أكدت ارتفاع في حجم تفاعل الطالب مع المهمة في مختلف المجالات الدراسية (Gao & Brennan, 2001; Güler & Gelbal, 2010 ; Hébert & al., 2014 ; Huang, 2009 ; Lane et al., 1996 ; McBee & Barnes, 1998 ; Nie et al., 2007; Shavelson et al., 1993 ; Webb et al., 2000). في حين تتعارض مع نتائج بعض الدراسات (Chen et al., 2007; Gebril, 2009; Lee & Kantor, 2007) التي توصلت إلى أن أكبر مصدر لتباين الخطأ راجع إلى تفاعل الطالب مع المهمة مع المقدر الممزوج بالأخطاء العشوائية غير المقدرة في التصميم،

وتتعارض أيضا مع النتيجة التي توصلت إليها دراسة (Casanova & Demeuse, 2011) بأن أكبر مصدر تباين للخطأ راجع إلى المقدرين وتفاعل المقدر مع المهمة.

ويمكن أن يفسّر ارتفاع مكوّن تباين تفاعل الطالب مع المهمة من خلال مشكلة معرفية ناتجة عن انتقال أثر التعلم لدى الطلاب من مهمات معيّنة إلى أخرى (Parkes, 2001)، وذلك بسبب عدم قدرة الطلاب على نقل معارفهم من مهمة إلى أخرى، ومن السياق المدرسي إلى السياق العملي. وقد أدى ارتفاع تباين تفاعل الطالب مع المهمة إلى خفض إمكانية تعميم درجات أداء الطلاب في هذا النوع من التقييمات (Miller & Linn, 2000)، كما حاول بعض الباحثين تجريب بعض الاستراتيجيات المعرفية في التعلم اعتمدت على خرائط المفاهيم بهدف خفض تباين تفاعل الطالب مع المهمة (Parkes, Zimmaro, Zappe & Suen, 2000).

وفيما يتعلق بالتأثير الرئيسي للمهمة، فقد تبيّن أنه ثاني أكبر مصدر لتباين الخطأ، وقد بلغت نسبة تأثيره (19.4%) من التباين الكلي، ويرجع ارتفاع مصدر تباين المهمة بالأساس إلى الاختلاف في درجة صعوبة المهمات لدى جميع الطلاب، بحيث يمكن أن تكون بعض المهمات أكثر صعوبة من المهمات الأخرى لدى نفس الطلاب بسبب الاختلاف في سياق بعض المهمات عن المهمات الأخرى أو بسبب الاختلاف في السندات المقدمة أو في طريقة صياغة تعليمات المهمات، وهذا ما أدى إلى صعوبة بعض المهمات عن المهمات الأخرى. وقد توصلت بعض الدراسات إلى هذه النتيجة، فقد أكدت مراجعة الباحثين هوانج (Huang, 2009) وإنّامي وكويزومي (In'nami & Koizumi, 2015) إلى أن تأثيرات المهمة قد فسّرت نسبة كبيرة من تباين الخطأ في درجات أداء الطلاب.

وبالنظر إلى تقييم الكفاءة، فان تقديم عدد كبير من المهمات في الاختبار تنتج عنها تكاليف وجهود إضافية مرتفعة، وتتطلب توفير المزيد من الوسائل أثناء تمرير المهمات وتقدير أداء الطلاب (Parkes, 2000)، وتؤدي أيضًا إلى تأثيرات غير مرغوبة على الطلاب كالتعب والإرهاق، لذا من الضروري تقديم مهمات متجانسة لخفض مصدر تباين المهمة، فقد أشار مسيبي وبارنز (McBee & Barnes, 1998) إلى أن تجانس المهمات يسهم في خفض تباين الخطأ الراجع إلى المهمة وتفاعل الطالب مع المهمة، إلا أن إعداد مهمات متجانسة كذلك عملية صعبة للغاية نظرًا لاختلاف مهمات تقييم الكفاءة من حيث خصائصها ومستوى تعقيدها ودرجة صعوبتها.

ومن الممكن أن ينطبق هذا التفسير مع المهمات التي أدرجت في الاختبار، فكل مجموعة مهمات من المجموعات الثلاثة صيغت بطريقة مختلفة عن الأخرى، وهذا ما جعل مستوى صعوبة كل صيغة يختلف عن الأخرى، حيث كانت المهمات ذات المعلومات المشوّشة أقل صعوبة من المهمات غير المحكمة البناء، والمهمات غير محكمة البناء بدورها أقل صعوبة من المهمات محكمة البناء.

ويمكن أن يفسّر ذلك بأن المعلومات المشوّشة المقدّمة في سياق المشكلة لم تضف للمهمات ذات المعلومات المشوّشة مستوى تعقيد أكبر، وكذلك عدم وضوح طريقة الحل في المهمات غير محكمة البناء لم تكن كافية من حيث التعقيد بالمقارنة مع المهمات المحكمة البناء التي تم تقديم المعلومات الضرورية للحل ووضوح المطلوب من الطلاب.

وبيّنت دراسات القرار انخفاضًا في معاملات إمكانية التعميم النسبية والمطلقة بالمقارنة مع الحد الأدنى المطلوب، والذي يتطلب أن يصل أو يفوق "0.80" (Bain & Pini, 1996)، فإذا كانت نسبة تباين الدرجات الملاحظة إلى تباين الدرجة الشاملة الناتجة عن مصادر التباين تصل أو تفوق (80%) تعدّ مقبولة، وإذا كانت أقل من (80%) يجب البحث عن طرق لتحسين أداة القياس (Cardinet, 1988)، بحيث أن أظهرت النتائج أن قيمة معامل إمكانية التعميم النسبي (0.79) وقيمة معامل إمكانية التعميم المطلق (0.74) لم يصلا إلى المستوى المطلوب (0.80)، رغم أن هذه المعاملات مشجعة، وقريبة من الحدّ الأدنى، إلا أن تعميم درجات الطلاب في الاختبار على نطاق أوسع من المهمات والمقدرين يمكن أن ينتج عنه تحفظات.

وتتفق هذه النتائج مع نتائج العديد من الدراسات التي أكدت عدم بلوغ التقييمات مستويات ثبات مقبولة (Gao & Brennan, 2001; Hébert et al., 2014; McBee & Barnes, 1998; Nie et al., 2007; Shavelson et al., 1993 ; Taylor & Pastor, 2013)، في حين تتعارض نتائج هذه الدراسة مع ما توصلت إليها عدد من دراسات (Bain, 2008 ; Güler & Gelbal, 2010; Lee & Kantor, 2007 ; Webb et al., 2000) فرغم كفاية عدد المهمات المتضمنة في الاختبار، إلا أن مستوى الثبات غير كاف، ويعود بالأساس كما أشرنا إلى تأثير مصادر تباين الخطأ الراجعة إلى تفاعل الطالب مع المهمة والتأثير الرئيسي للمهمة على ثبات درجات الاختبار. وقد اعتبرت مشكلة ضعف الثبات خاصية من خصائص تقييمات الكفاءة، وبالتالي تواجه هذه التقييمات مشكلتين أساسيتين، تتعلّق الأولى بانخفاض في معاملات ثباتها وتتعلق الثانية بارتفاع تكاليفها.

كما بينت دراسات القرار ضرورة رفع عدد المهمات إلى (14) مهمة لبلوغ معامل إمكانية تعميم نسبي (0.80) ورفع المهمات إلى (14) مهمة لبلوغ معامل إمكانية التعميم مطلق (0.80) باستخدام 3 مقدرين، بالإضافة إلى ضرورة رفع المقدرين إلى (4) باستخدام (9) مهمات للحصول على مستوى إمكانية التعميم النسبي (0.80)، كما أن رفع عدد المقدرين إلى أقصى حدّ ممكن لا يمكن بلوغ مستوى مقبول من معامل إمكانية التعميم المطلق.

ويمكن أن نستخلص أن زيادة عدد المهمات كانت أكثر فعالية من زيادة عدد المقدرين، ويمكن أن يفسّر ذلك بمصادر تباين الخطأ الأكثر تأثيرًا على ثبات درجات الطلاب، والتي كانت راجعة إلى الارتفاع في تغير معاينة المهمة (تفاعل الطالب مع المهمة، والتأثير

الرئيسي للمهمة)، ولم تكن راجعة إلى تغيّر معاينة المقدر وتفاعلاته مع الأبعاد الأخرى (التأثير الرئيسي للمقدر، وتفاعل الطالب مع المقدر، وتفاعل المهمة مع المقدر، وتفاعل الطالب مع المهمة ومقدر)، وقد أكدت العديد من الدراسات السابقة هذه النتيجة (Chen et al., 2007; Gao et al., 1994 ; Gebril, 2009; Lane et al., 1996; McBee & Barnes, 1998 ; Nie et al., 2007; Shavelson et al., 1993 ; Taylor & Pastor, 2013)، ويمكن أن يفسر ذلك بالعناية الشديدة بعملية التصحيح، والتدريب الكافي للمقدرين، والاعداد المحكم لشبكات التصحيح، كل هذه العوامل ساهمت بشكل كبير في خفض مصادر تباين الخطأ الراجعة إلى المقدرين، وتفاعلات المقدرين مع الطلاب ومع المهمات.

وعلى ضوء ما تقدم يبدو من المهم جدا العناية بفحص أداء الطلاب في مواقف أكثر ملاءمة، وضمان انغماس الطلاب في انجاز المهمات، وضرورة إعداد موازين تقدير واضحة، والحرص على إعداد مهمات واقعية متجانسة للحصول على تقييمات ذات جودة فنية عالية.

التوصيات

بناء على نتائج الدراسة الحالية يمكن اقتراح إجراء دراسات مستقبلية تسهم في إثراء أدبيات القياس والتقييم من خلال استخدام نظرية إمكانية التعميم حول:

- تجريب مدى فعالية تغيير تصميمات الدراسة، واستخدام مهمات التقييم الأكثر تماثلاً، واستخدام استراتيجية خرائط المفاهيم من أجل خفض مصدر تباين تفاعل الطالب مع المهمة.
- إدماج بعض أبعاد القياس الأخرى في تصميمات إمكانية التعميم أثناء تقدير ثبات وصدق اختبارات تقييم الكفاءة، مثل: فترات التقييم، ومجالات المحتوى، وصيغ الاختبار، ومعايير التقدير.
- قياس صدق وثبات أنظمة تقييم الكفاءة الأخرى على غرار الاختبارات، وأساليب الملاحظة، والتقييم الذاتي، وتقييم الأقران، وملفات الانجاز.
- فحص أدلة الصدق الحديثة لتقييمات الكفاءة وفق الاطار الفكري الذي قدمه ميسيك (Messick, 1995) لأدلة المحتوى، والعمليات والتأصيل النظري، والأدلة البنائية، وإمكانية التعميم، والأدلة الخارجية، والعواقب التربوية.

References

Allam, S. (2000). *Psychological and educational measurement and assessment: Principles, practices and modern perspectives*. Cairo: Dar Al-Fikr Al-Arabi.

Allam, S. (2004). *Alternative educational assessment: Theoritical and methodological basics and pratical applications*. Cairo: Dar Al-Fikr Al-Arabi.

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(2), 153-177.

Baartman, L., Prins, F., Kirschner, P., & Van der Vleuten, C. P. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation*, 33(3-4), 258-281.

Bain, D. (2008). *Radiographie d'une épreuve commune de mathématiques au moyen du modèle de la généralisabilité. Actes du 20ème Colloque.* Genève: ADMEE-Europe. Availible online at: https://plone.unige.ch/sites/admee08/symposiums.

Bain, D. (2014). Généralisabilité et évaluation des compétences: Pistes et fausses pistes. In C. Dierendonck (Ed.), *L'évaluation des compétences en milieu scolaire et en milieu professionnel*. Bruxelles: De Boeck Supérieur.

Bain, D., & Pini, G. (1996). *Pour évaluer vos évaluations. La généralisabilité : Mode d'emploi*. Genève: Centre de Recherches Psychopédagogiques.

Boston, C. (2002). *Understanding scoring rubrics: A guide for teachers.* University of Maryland: ERIC Clearinghouse.

Brennan, R. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.

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.

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. (1988). *Evaluation scolaire et pratique*. Bruxelles: De Boeck-Wesmael.

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.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Application to educational measurement. *Journal of Educational Measurement,* 13(2), 119-135.

Casanova, D., & Demeuse, M. (2011). Analyse de différentes facettes influant sur la fiabilité de l'épreuve d'expression écrite d'un test de français langue étrangère. *Mesure et Évaluation en Éducation,* 34(1), 25-53.

Chen, E., Niemi, D., Wang, H., & Mirocha, J. (2007). *Examining the generalizability of direct writing assessment tasks.* University of California. Los Angeles: CRESST (CSE Technical Report N° 718). Available online at: cresst.org/publications/cresst-publication-3089

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.

Cronbach, L., Rajaratnam, N., & Gelser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, 16(2), 137–163.

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. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and ue of performance assessments. *Applied Measurement in Education,* 4(4), 289-303.

Feuer, M., & Fulton, K. (1993). The many faces of performance assessment. *The Phi Delta Kappan*, 74(6), 478.

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.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing,* 26(4), 507–531.

Gerard, F. (2006). L'évaluation des acquis des élèves dans le cadre de la réforme éducative en Algérie. In N. Toualbi-Thaâlibi (Ed.), *Réforme de l'éducation et innovation pédagogique en Algérie*. UNESCO: ONPS.

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., 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, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement,* 69(6), 887-912.

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.

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.

McBee, M., & Barnes, L. (1998). The generalizability of a performance assessment measureing achievement in eight-grade mathe-matics. *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.

Meyer, J., (2010). *Reliability.* New York: Oxford University Press.

Miller, M., & Linn, R. (2000). Validation of performance-based assessments. *Applied Psychological Measurement,* 24(4), 367–378.

Ministry of National Education. (2011*). Fourth year primary school curriculum.* Algiers: The National Authority for School Publications.

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,* 8(16), 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.

Roegiers, X. (2000). *Pour une pédagogie de l'intégration.* Bruxelles: De Boeck Supérieur.

Roegiers, X. (2004). *L'école et l'évaluation.* Bruxelles: De Boeck Supérieur.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences.* Bruxelles: De Boeck Supérieur.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing new modes of assessment: In search for qualities and standards.* Boston: Kluwer Academic.

Shavelson, R., Baxter, G., & 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 assessents. *Journal of Educational Measurement,* 30(3), 215-232.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* California: Sage Publications.

Shavelson, R., & Webb, N. (2009). Generalizability theory and its contribution to the discussion of the generalizability of research findings. In K. Erickan., & W. Roth (Eds.), *Generalizability from educational research* (pp. 13-32). New York: Routledge.

Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation, 43*, 5-13.

Swiss Society for Research in Education Working Group. (2010). *EduG user guide.* Neuchatel: IRDP. Available online at: http://www.irdp.ch /edumetrie/logiciels.html

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.

Tebaa, F. (2017). Critics of using competences concept in educational practices related to assessment. *Social Sciences Journal*, 24, 162-177.

Tebaa, F., & Lifa, N. (2015). Competences assessment from the perspective of generalizability theory. *Psychological and Educational Studies Review*, 12, 206-227.

Webb, N., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education*, 13(3), 277–301.

Webb, N., Shavelson, R., & Steedle, J. (2012). Generalizability theory in assessment contexts. In C. Secolsky., & B. Denison (Eds.), *Handbook on mesurement, assessment and evaluation in higher education* (pp. 132-149). New York: Routledge.

Yin, Y., & Shavelson, R. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education*, 21(3), 273-291.