استخدام نظرية التعميم للكشف عن مدى مساهمة مصادر التباين المتعددة في صدق اختبار في الرياضيات

محمد العرايضة و نضال الشريفين*

Doi: //10.47015/18.3.11

تاريخ تسلم البحث: 2021/1/31 تاريخ قبوله: 2021/4/1

ملخص: هدفت الدراسة للكشف عن مدى مساهمة مصادر التباين المتعددة في صدق اختبار في الرياضيات باستخدام نظرية التعميم، وذلك من خلال تقدير حجم تباين الخطأ المفسر من الأبعاد (المهام، صيغ المهام، وطريقة التصحيح) في التباين الكلي. تكونت عينة الدراسة من (301) مفحوص من الصف الخامس الأساسي طبق عليهم اختبار في الرياضيات تكون من (12) مهمة في مجال الأعداد والعمليات عليها، توزعت على (4) صيغ بالتساوي (التطبيق، والاستدلال، والانتقاء، والرأي). استخدم الباحثان التصاميم (مفحوص×صيغة)، و(مفحوص×مهمة×صيغة)، و(مفحوص×طريقة)، و(مفحوص×مهمة×طريقة) المتقاطعة كليًا. وقد استخدمت برمجية (EduG) لتحليل البيانات. وتوصلت دراسات التعميم إلى أن أكبر مصادر تباين الخطأ في التصميم (مفحوص×صيغة) يعود إلى تفاعل (مفحوص-صيغة)، وفي التصميم (مفحوص×مهمة×صيغة) يعود إلى تفاعل (مفحوص-مهمة-صيغة)، في حين جاءت مُعاملات التعميم في التصميم (مفحوص×صيغة) أفضل منها في التصميم (مفحوص×مهمة×صيغة)، كما أن أكبر مصادر تباين الخطأ في التصميم (مفحوص×طريقة) يعود إلى تفاعل (مفحوص-طريقة)، وفي التصميم (مفحوص×مهمة×طريقة) يعود إلى تفاعل (مفحوص-مهمة-طريقة)، وجاءت مُعاملات التعميم في التصميم (مفحوص×مهمة×طريقة) أفضل منها في التصميم (مفحوص ×طريقة).

(الكلمات المفتاحية: نظرية التعميم، الصدق، مصادر التباين المتعددة، المهمة المركبة، اختبار رياضيات، تقييم الأداء)

مقدمة: يَشهد القياس والتقويم التربوي تطورًا سريعًا نتيجة تزايد الحاجة المُلحة لتطوير وبناء الاختبارات في مجال تقييم أداء (Performance Assessment) المفحوصين. لذلك فقد أصبح التقييم ضرورةً مهمةً تشمل جميع نواحي الحياة؛ فتقييمات الأداء تُعتبر شكلًا من أشكال تطور استراتيجيات التقويم، وتُعتبر من التقييمات البديلة لأداء المفحوصين. وواكبَ ذلك التطور تقدم في أسس بناء أدوات القياس المتنوعة، إضافةً إلى طُرق جمع وتحليل وتفسير البيانات المستمدة من هذه الأدوات (Urbina, 1997/2015).

وعليه، تُعد عملية التقويم عمليةً منهجيةً تتطلب جمع بياناتٍ موضوعية وصادقة من مصادر متعددة، وبأشكالٍ متنوعة، وتُحاكي عمليات التفكير العليا، مثل بلورة الأحكام واتخاذ القرارات وحل المشكلات الحياتية؛ إذ إن التقويم الذي يقيس نتاجات التعلم يعكس أداء المفحوصين ويقيسه في مواقف حقيقية، ويجعلهم ينغمسون في مهام (Tasks) ذات قيمة ومعنى بالنسبة لهم، وقادرين على معالجة المعلومات وتحليلها ونقدها (Ministry of Education, 2004). وتُعتبر الاختبارات أحد أدوات التقويم الهامة، حيث تُستخدم نتائجها في اتخاذ القرارات التربوية الهامة؛ إذ إن زيادة أهمية القرار المُتخذ تزيد من الحاجة للحصول على معلوماتٍ دقيقةٍ من الاختبار وذات صلة وثيقة بالغرض الذي أُعد لأجله، وهذا بدورهِ يقودنا لدراسة صدق الاختبار (Test Validity) للحصول على درجة عالية من الدقة بأقل خطأ ممكن (Audeh, 2010).

Using Generalizability Theory to Detect the Contribution of Multiple Sources of Variance in the Validity of a Test in Mathematics

Mohammad Al-Araideh and Nedhal Al-Sharifeen, Yarmouk *University,Jordan.*

Abstract: This study aimed to detect the contribution of multiple sources of variance in the validity of a test in mathematics by using the Generalizability Theory, through estimating the magnitude of error variance that is explained by the facets (tasks, task formulae and correction method) in the total variance. The study sample consisted of (301) students from the fifth grade who were subjected to a mathematics test that consisted of (12) tasks in the domain of numbers and operations on them. The tasks were equally distributed on (4) formulae (application, inference, selection and opinion). The researchers used the designs (person×formula), (person×task× formula), (person×method) and (person×task×method) completely crossed and used (EduG) software to analyze the data, The results of generalizability studies indicated that the largest sources of error variance were in the design (person×formula) which refers to the interaction (person-formula) and in the design (person×task×formula) which refers to the interaction (person-task-formula). The generalizability coefficients in the design (person×formula) were better than in the design (person×task×formula). The results of the study also found that the largest sources of error variance were in the design (person×method) which refers to the interaction (person-method) and in the design (person×task×method) ehich refers to the interaction (person-task-method). The generalizability coefficients in the design (person×task×method) were better than in the design (person×method).

(**Keywords**: Generalizability Theory, Validity, Multiple Sources of Variance, Complex Task, Mathematics Test, Performance Assessment)

وفي ضوء ذلك، يُعتبر الصدق من أهم المسائل والاعتبارات في عملية بناء الاختبارات، وقد تطوّر بشكل ملحوظ في الآونة الأخيرة نتيجة ظهور تقييمات الأداء التي تتميز بتعدد أبعاد تكويناتها الفرضية، وتتطلب مهامًا تدمج المعارفَ والمهارات والاتجاهات المتعددة، إضافةً إلى تقديرات مصححين (Judges) خُبراء، الأمر الذي يُسلط الضوء على إعادة النظر في مفهوم صدق تقييمات الأداء من خلال تحديد معايير جديدة، كشمولية المحتوى، وتعميم النتائج، ومقارنة التقييمات (Messick, 1995; AERA, 1999; Linn et al., 1991; Linn, 1994, Johnson et al., 2009).

* جامعة اليرموك، الأردن.

وعليهِ فإن الصدق يتأثر بمصادر خطأ متعددة، حسب ما أشار إليه ميسيك (Messick, 1995)، وهو درجة استدلالية يُعبر عنها بارتباط مهام التقييم مع مهام أخرى. وتفسير الأداء يتأثر بدرجة تعميمه عبر الفترات (Occasions)، والمصححين. كذلك فإن درجة تعميم تقييمات الأداء تعتمد على أدلةِ صدقٍ كطرائق لقياس الأداء (Kane, 1982). وفي هذا السياق، بيّن شافلسون وويب (Shavelson & Webb, 2009) أن بعض التصميمات المرتبطة بأدلة الصدق كمصادر لخطأ القياس هي طرق التقييم، ومحتوى الاختبار، وطرق التصحيح، وهذا يُساعد في جمع أدلة عن صدق قياس أداء المفحوصين. وعليهِ فإن مُكونات التباين الناتجة من الأبعاد المختلفة (المهام، وصيغ المهام، وطرق التقييم) تُعتبر أدلة لفحص صدق تقييمات الأداء؛ فزيادة تباينات الأبعاد تُعتبرُ مؤشرًا على الصدق، لذلك فإن تقييم المفحوص من خلال عينة مهام يمكن من خلاله الاستدلال على صدق المحتوى، ودرجة التقارب بين صيغ المهام (Tasks Formula)، ودرجة الاتساق بين طرائق التصحيح (Correction Methods) تُعتبر مؤشرًا على الصدق التقاربي (Convergent Validity) (Brennan, 2000; Huang, 2012).

وللحصول على أداء المفحوص، من الضروري بناء الاختبار بمادة دراسية تُحاكي حياتنا اليومية، وتحظى باهتمام العالم في وقتنا الحالي. فالرياضيات تلعب دورًا مهمًا وجوهريًا، وتعتبر بيئة خصبة لتنمية التفكير، وهي غنية بالمعلومات والمواقف التي تتطلب حل مشكلات (Abo Zinah & Ababneh, 2007).

ومن أجل التحكم بجميع مصادر أخطاء القياس التي تُؤثر في صدق قياس أداء المفحوصين، يُفضل استخدام طُرق إحصائية جديدة تَدرس مصادر التباين (Sources of Variance) في آن واحد. فنظرية الاختبار (Test Theory) غير قادرةٍ على التمييز بين مصادر الأخطاء المتعددة بتحليل واحد، لذلك جاءت نظرية التعميم (Generalizability Theory) التي تمتلك إطارًا مفاهيميًا واسعًا لأساليب إحصائيةٍ تسمح بمعالجة مختلف مشكلات القياس في الوقت نفسه (Brennan, 2001).

ولهذا فإنَّ نظرية التعميم (GT) تَستخدم طُرق تحليل التباين (Analysis of Variance) في تقدير صدق القياسات السلوكية، ولا تُعتبر بديلاً عن نظرية الاختبار؛ لأن استخدام تحليل التباين(ANOVA) ساعد في تطوير ومعالجة أبعاد القياس المعقدة، من خلال ابتكار مفهوم بديل يجمع بين طرق تحليل التباين ونظرية الاختبار يسمى نظرية التعميم. واعتبر برينان (Brennan, 2001) أن نظرية الاختبار (TT) وتحليل التباين أبوان لنظرية التعميم، حيث تقدم نظرية التعميم حلولًا في كيفية تجزئة مصادر التباين لجميع أبعاد القياس وتقدير مُكونات التباين والتفاعلات بينها في الوقت نفسه.

ومن وجهة نظر ين و شافلسون (Yin & Shavelson, 2008)، فإن ما يُميز نظرية التعميم عن نظرية الاختبار أنها تستطيع الكشف عن مصادر التباين المتعددة في الوقت نفسه، ويمكنها تقدير مُكونات تباين الأبعاد والتفاعلات بينها، إضافةً إلى مُرونتها التي تسمح بتعديل القياسات وعدد مستويات الأبعاد بهدف الاقتصاد في التكلفة والوقت. كذلك فهي تُزود الباحثين بمعلومات حول التفسيرات النسبية والمطلقة.

ولهذا، فإن أُولى الخطوات التي يقوم بها الباحث هي تحديد الشُروط أو الظُروف في جمع الملاحظات (Observations) (الأبعاد)؛ فليست هناك علامة حقيقية (True Score) واحدة للمفحوص في الأداة كما في نظرية الاختبار، وإنما تكون له علامة شاملة (Universe Score) يقصد بها القيمة المتوقعة للتقييمات الملاحظة التي يحصل عليها في مختلف المواقف التي تنتمي إلى النطاق الشامل (Universe Domain) المطلوب. وتتباين درجات المفحوص التي تنتمي إلى النطاق الشامل في أكثر من جانب؛ إذ يسمى كل واحد الوجه (Facet) أو البُعد. فالبُعد خاصية في القياس مثل المهام، ويُمثل كل بُعد من الأبعاد مصدرًا هامًا من مصادر التباين الذي من الضروري دراسته. ويتكون كل بُعد من أبعاد القياس من مستويات تشتمل على مختلف ظروف القياس، كما تُسمى الملاحظات التي نحصل عليها من مختلف الشروط الممثلة في نظرية التعميم النطاق الشامل للملاحظات المقبولة (Universe of Admissible Observations)، كالمفحوصين، ومختلف أبعاد التقييم، والمصححين (Allam, 2000). ويتم الحصول على درجة شاملة للمفحوص من خلال الوسط الحسابي لأدائه على كل المهام، تُقدر من جميع المصححين، مُتأثرةً بمصادر خطأ قياس راجعة إلى المهام، وصيغ المهام، وطرائق التصحيح.

ويُحدد ألن و ين (Allen & Yen, 1979) الغرضَ من دراساتِ التعميم (Generalizability Studies) بأنه جمع المعلومات عن المفحوصين في الاختبار في عدة مستويات من الظروف، وتحديد التباين الذي يُعزى لكل مصدر، وذلك باستخدام فكرة تحليل التباين وتقدير مُعامل التعميم (Generalizability Coefficient)، الذي يأخذ بالاعتبار اختلاف ظروف تطبيق الاختبار والعوامل المؤثرة في تباين الأداء على الاختبار، بهدف تعميم نتائج القياس على النطاق الشامل المُراد قياسه. ويُشير شافلسون و وب (Shavelson & Webb, 2009) إلى أن مُعاملَ التعميم وفقًا لنظرية التعميم يُعبَّرُ عنه من خلال نسبة تباين العلامة الشاملة إلى تباين العلامة الملاحظة. فمُعامل التعميم النسبي (Relative Generalizability Coefficient) يُستخدم في تفسير بيانات مستمدة من اختبارات معيارية المرجع (Norm-reference Tests) ليُتيحَ هذا المُعامل تحديد المركز النسبي للمفحوصين، وذلك من خلال القرارات النسبية المتعلقة بمقارنة الفروق بين أداء المفحوصين. وقد أشار الحربي والحربي

(Alharbi & Alharbi, 2017) إلى أنه بمعنى مقارنة أداء المفحوصين بعضهم ببعض. وعلى غرار ذلك، نستخدم مُعامل التعميم المطلق (Absolute Generalizability Coefficient) لتقييم قدرة أداة القياس على مقارنة أداء المفحوصين بمستوى أداء مطلق (Brennan & Kane, 1977)، ويُستخدم في تفسير بيانات مستمدة من اختبارات محكية المرجع (Criterion-reference Tests) بالاعتماد على القرارات المطلقة المتعلقة بتصنيف أداء المفحوصين وفقًا لمحكاتٍ معينةٍ (Alharbi & Alharbi, 2017). كذلك ذكر كين (Kane, 1982) أن نظرية التعميم تُعتبر ملائمة لمعالجة أبعاد الصدق؛ إذْ إنه من الممكن اعتبار درجة التقارب بين صيغ المهام وطرائق التصحيح مؤشرًا على الصدق التقاربي.

وفي هذا السياق، أُجريت العديد من الدراسات التي اهتمت بالكشف عن تغير معاينات تقييمات الأداء في ضوء نظرية التعميم. فقد أجرى شافلسون وآخرون (Shavelson et al., 1993) دراسةً حولَ التغيراتِ في معاينات تقييمات الأداء في الرياضيات، شارك فيها (105) مفحوصين من مستوى السنة السادسة بالإجابة عن ثلاث مهام قام الباحثون باعتماد تقديرها من قبل مصححين اثنين، وباستخدام التصميم (مفحوصxمصححxمهمة). وبعدها تمت مقارنة النتائجِ مع نتائجَ تقييمات برنامج تقييم كاليفورنيا للعلوم. وتوصلت النتائج إلى أن أكبر مُكون تباين خطأ قياس راجع إلى تفاعل (مفحوص-مهمة)، وبينت دراسات القرار حاجة الدراسة إلى (15) مهمة من أجل الحصول على مُعامل تعميم مقبول.

وفي دراسة أُجريت حول التغيرات في معينات تقييمات الأداء في العلوم من قِبَلِ شافلسون وآخرين (Shavelson et al., 1993)، تم تطبيق ثلاث مهام مستقلة باستخدام أربع طرق قياس. شارك في الدراسة (186) مفحوصاً من الصفين الخامس والسادس الابتدائيين، وتم التقييم من قبل مصححين اثنين، باستخدام تصميمين هما: (مفحوصxمصححxمهمةxفترة)، و(مفحوصxمصححxمهمة)، وأجريت دراسة تصميم (مفحوصx فترة) على جميع المفحوصين، كما استخدم في الدراسة نموذج كين (Kane, 1982) لفحص أدلة الصدق التقاربي للتصميم (مفحوصxمهمةxطريقة). وتوصلت الدراسة إلى أن أكبر مصدر لتباين الخطأ ناتج من تفاعل (مفحوص-مهمة-فترة) في التصميم الأول، بينما في التصميم الثاني كان أكبَر مصدر تباين خطأ ناتجًا من تفاعل (مفحوص-مهمة)، كما أن أكبر مُكون لتباين الخطأ ناتج من تفاعل (مفحوص-مهمة-طريقة) في تصميم دراسة الصدق، ونحتاج إلى (23) مهمة للحصول على مُعامل تعميم مقبول. وبينت النتائج أن طرق القياس ليست متقاربة، ويمكن أن تقيسَ جوانب مختلفةٍ من التحصيل.

وفي السياق ذاته، أجرى رويز-بريمو وشافلسون (Ruiz-Primo & Shavelson, 1996) دراسة لاستقرار التقييمات في العلوم. اشتملت عينة الدراسة على (29) مفحوصاً من طلبة الصف السادس تم اختيارهم عشوائياً من أصل (110) طلاب، بحيث يطلب من المفحوصين إنجاز ثلاث مهام استُخدم فيها (8) مصححين في الفترة الزمنية الأولى، و(4) مصححين في الفترة الزمنية الثانية، كما استُخدم تصميم (مفحوصxفترة). وبينت النتائج أن مُعاملات التعميم متدنية، وأظهرت النتائج أن المفحوصين استخدموا إجراءات مختلفة في الفترتين، وبينت دراسات القرار أن زيادة عدد المهام وعدد الفترات الزمنية من شأنها أن ترفع من مُعاملات التعميم.

كذلك قام لين وآخرون (Lane et al.,1996) بدراسة عن نظرية التعميم وصدق تقييم أداء الرياضيات، طُبقَ فيها اختبار الرياضيات على (36) مهمة من المهام ذات الإجابات المفتوحة لتقييم حل المشكلات. استخدم الباحثون تصميمين متقاطعين هما: (مفحوصx مهمة)، و(مفحوصxمهمةxمصحح)، وتصميمًا متداخلًا جزئيًا (مفحوص:(مدرسةxمهمة)). وتوصلت الدراسة إلى أن أكبر مُكون لتباين الخطأ ناتج من تفاعل (مفحوص-مهمة)، وتوصلت دراسات القرار إلى أن زيادة عدد المهام يرفع من مُعاملات التعميم، بينما كان تأثير زيادة عدد المصححين ضعيفًا، وبلغت مُعاملات التعميم بين (0.8) و (0.97) اعتمادًا على كل صيغة من صيغ المهام والمستوى الدراسي.

وأجرى مكبي وبارنز (Mcbee & Barnes, 1998) دراسة حول نظرية التعميم لقياس أداء تحصيل المفحوصين في الرياضيات، بهدف دراسة الاستقرار عبر الزمن بين المهام. تم اختيار (4) مهام، واشتملت عينة الدراسة على (101) مفحوص من طلبة السنة الثامنة، واستخدم الباحثان التصميم (مفحوصxمهمةxمصححxفترة). وقد تم تحليل البيانات بواسطة برمجية (GENOVA). وتوصلت الدراسة إلى أن أكبر مصدر لتباين الخطأ يعود للمهمة وتفاعلاتها، كما بينت نتائج دراسات القرار أثر عدد المهام الأكثر تماثلًا في تحقيق مُعاملات تعميم مقبولة.

وقام ويب وآخرون (Webb et al., 2000) بدراسة هدفت إلى البحث في الاعتمادية وتبديل طرق التقييم في العلوم، ودراسة تعميم تقييم العلوم. طور الباحثون اختبارين في العلوم، أحدهما لأداء المعالجة اليدوية، والثاني اختبار قلم وورقة بفقرات اختيار من متعدد. تكونت العينة من (57) مفحوصاً، واشتمل كل اختبار على مهمتين، وتم تصحيح الاختبارين من قبل مصححين اثنين. واُستُخدمت في الدراسة التصاميم (مفحوصxمهمةxمصحح)، و(مفحوصxمهمةxمصححxفترة). وتوصلت الدراسة إلى أن أكبر مُكون للتباين ناتج من تفاعل (مفحوص-مهمة). أما دراسات القرار، فقد بينت أنه دون تضمين بُعد الفترة كان مُعامل التعميم مرتفعًا في اختبار القلم والورقة، ومقبولاً في اختبار المعالجة اليدوية في الفترة الأولى ومرتفعًا في الفترة الثانية. وللحصول على مُعاملات تعميم مرتفعة، يُنصح بإدراج (3) مهام مع عدم تضمين بُعد الفترة.

وأجرى سميث و كوليكويتش (Smith & kulikowich, 2004) دراسةً مرتبطةً بتطبيق نظرية التعميم ونموذج راش متعدد الأبعاد على تقييم مهارات حل المشكلات، لتقدير صدق وثبات تقييم أداء مهارات حل المشكلات، ولتحقيق أهداف الدراسة قام الباحثان بتقديم (5) مهام لحل مشكلات مركبة لعينة تكونت من (44) مفحوصاً، طُبقَ الاختبار على فترتين، وصُحِّحَ بواسطة مصححين اثنين، وتم استخدام التصميم (مفحوصxمهمةxمصححxفترة). وتوصلت الدراسة إلى أن مُعامل التعميم النسبي كان مقبولاً، ومُعامل التعميم المطلق كان متدنياً، وكان أكبر مُكون لتباين الخطأ ناتجًا من مُكون المهمة، وبينت دراسات القرار أن التقليل من عدد مستويات الأبعاد لا يعطي مُعاملات تعميم مقبولة.

وهدفت دراسة ني وآخرين (Nie et al., 2007) إلى تطبيق نظرية التعميم في فحص جودة التقييم البديل في الرياضيات، وشارك فيها (29) مفحوصاً، وذلك من خلال قيامهم بإنجاز مهمتين. وتم تقدير أدائهم بواسطة مصححين اثنين، وباستخدام التصميم ثنائي الأبعاد (مفحوصxمهمةxمصحح). وبينت نتائج الدراسة أن مُعاملات التعميم متدنية، وأن أكبر مصادر تباين الخطأ ناتج من تفاعل (مفحوص-مهمة)، وتوصلت دراسات القرار إلى أن الزيادة في عدد المهام أفضل من الزيادة في عدد المصححين.

وأجرى تشين وآخرون (Chen et al., 2007) دراسةً بهدف فحص تعميم مهام تقييم الكتابة المباشرة، وصدق قياس القدرة على الكتابة، وذلك من خلال استخدام (4) مهام. اشتملت عينة الدراسة على (397) مفحوصاً، وطُلبَ من المفحوص اختيار مهمتين عشوائياً لإنجازهما. وقد تم تدريب (4) مصححين لتقدير أداء المفحوصين، بالاعتماد على فترتين زمنيتين لإنجاز المهام، واستُخدم التصميم المتقاطع كلياً (مفحوصxمقالxمصحح). وبينت نتائج الدراسة أن أعلى مُكون لتباين الخطأ ناتج من تفاعل (مفحوص-مقال-مصحح)، وتوصلت دراسات القرار إلى أنه بزيادة عدد المقالات، ترتفع مُعاملات التعميم.

وأجرى تانيلون وآخرون (Tanilon et al., 2009) دراسة هدفت إلى التحقق من صدق اختبار القبول المصمم لتقييم عينات الأداء في المهام الأكاديمية وتطويره. ولتحقيق ذلك، تم اعتماد (9) مهام في الفهم، والتطبيق، والاستدلال لتقييم عينات الأداء في التعليم ودراسات الطفل. وقد تم اختيار عينة من المفحوصين الحاصلين على درجة البكالوريوس في التربية بواقع (108) مفحوصين، واستُخدم التصميم (مفحوصxمهمةxمصحح)، وتم تقييم الأداء من خلال مصححين اثنين، وحُللت البيانات باستخدام برنامج (EduG)، وتم التحقق من مؤشرات الصدق التنبؤي. وأسفرت النتائج عن أن مُعاملَ الاعتمادية كان مقبولًا، ومرتفعاً في ضوء علامة القطع. وفي المقابل، توصلت دراسات القرار إلى أن زيادة عدد المهام إلى (20) مهمة ترفع مُعامل الاعتمادية ضمن المدى المقبول.

وأجرى هوانج (Huang, 2009) دراسة هدفت لتحليل ما وراء التحليل (Meta-Analysis) لمُعظم الدراسات التي تناولت تقييمات الأداء ضمن نظرية التعميم حول مقدار تغير معاينة المهمة في تقييم الأداء، حيث أُدرجت (50) دراسة اشتملت على (130) مجموعة بيانات مستقلة. وقد تضمنت الدراسة المنشورة ما بين عامي (1980) و (2006)، وتم ترميز مجموعة البيانات في عدة أبعاد (طريقة التصحيح، وطريقة التقييم، ومجال الموضوع، وتصميم الدراسة، ونوع المقال، وعدد الأبعاد، ومُكونات تفاعل المفحوص–المهمة). وأشارت مجموعة الدراسات إلى أن نسبة التباين للمهمة كانت مرتفعة بينما كان تباين (مفحوص-مهمة) مرتفعًا بشكل أكبر. وتوصلت الدراسة إلى أن أثر طريقة التصحيح وأثر طريقة التقييم غير دالَين إحصائياً، وأن أثر طبعة النشر غير دال إحصائياً. أما تباين (مفحوص-مهمة)، فقد انخفض بشكل ملموس عند إدماج الفترة كبعد.

وأجرت مابي (Mabe, 2014) دراسة استخدمت فيها نظرية التعميم بهدف تطوير أداة قياس في المهارات الاجتماعية. ولتحقيق الهدف، قام (3) مصححين بتقييم المفحوصين (4) مرات على (6) مهارات اجتماعية، واشتملت العينة على (20) مفحوصًا من الصف السادس. وأظهرت النتائج أن مُعامل التعميم كان مرتفعاً، كما أظهرت نتائج دارسات القرار أنه للوصول إلى مُعاملات تعميم مرتفعة لأداة القياس، تجب زيادة عدد المصححين إلى (10)، وتطبيق المقياس (10) مرات.

وقام الحربي والحربي (Alharbi & Alharbi, 2017) بإجراء دراسة حول مؤشرات الثبات باستخدام نظرية التعميم ومؤشرات صدق البناء لمقياس موهبة الابداع. وقد اشتملت العينة على (4368) مفحوصاً، من الصف الثالث الابتدائي إلى الصف الثالث الثانوي. وتكون المقياس من (20) فقرة، واستخدم مُعامل كرونباخ ألفا لدراسة ثبات المقياس ونظرية التعميم للحصول على جودة تقويم المصححين. واستخدم الباحثان التحليل العاملي التوكيدي لدراسة الصدق. وبينت النتائج أن المقياس يتمتع بجودة عالية، وأن مُعاملات التعميم كانت مقبولة، وأن صدق البناء يتضمن خمسة عوامل رئيسة مستقلة.

وأجرى طباع (Tebaa, 2020) دراسة هدفت إلى تطبيق نظرية التعميم لتقدير ثبات اختبار تقييم كفاءة الرياضيات لدى طلبة الصف الرابع الابتدائي. وقام الباحث بإعداد اختبار اشتمل على (9) مهمات، واشتملت العينة على (331) مفحوصاً. وتم تقدير الأداء من قبل (3) مصححين باستخدام التصميم المُتقاطع كُليًا (مفحوصxمهمةxمصحح)، وحُللت البيانات بواسطة برمجية (EduG). وبيّنت النتائج أن مصدر التباين الأكثر تأثيرًا على ثبات الأداء هو تفاعل (مفحوص–مهمة)، كما بينت دراسات القرار أن زيادة عدد المهام أفضل من زيادة عدد المُصححين لرفع مُعاملات التعميم.

وفي حدود علم الباحثين، تبينت نُدرة الدراسات التي استخدمت هذه النظرية المهمة، وخاصةً في الدراسات العربية في مجال تقدير صدق الاختبارات، كما أن مُعظم الدراسات ركزت على دراسة اختبارات تقييم الأداء التي تتكون من مهام مُركبة (Complex Tasks). وتتميز هذه الدراسة باستخدام مهام تنوعت إلى مهام التطبيق، والاستدلال، والانتقاء، والرأي التي تم الاعتماد في بنائها على فئات مهام دويل (Doyle, 1983). أما بعض الدراسات السابقة، فقد اهتمت بتطبيق نظرية التعميم على برمجية (GENOVA) كدراسة (Mcbee & Barens, 1998)، وبعضها استخدم برمجية (EduG). وهي من البرمجيات الحديثة التي تعمل على تقدير مُكونات التباين وحساب مُعاملات التعميم، كدراستي (Tebaa, 2020; Tanilon et al., 2009). ومن حيث العينة، تباينت أحجام العينات في الدراسات السابقة ما بين (20) مفحوصًا كدراسة (Mabe, 2014) و (4368) مفحوصًا كدراسة (Alharbi & Alharbi, 2017)، أما الدراسة الحالية فبلغ عدد أفراد عينتها (301) مفحوص، وهي كبيرة نسبياً. وقد أثبتت العديد من الدراسات (Shavelson et al., 1993; Mcbee & Barens, 1998; Smith & Kulikowicl, 2004; Nie et al., 2007) أن معظم مُعاملات التعميم لتقييمات الأداء كانت مُتدنية في مُعظم التخصصات مثل الرياضيات والعلوم والتخصصات الأخرى، لذلك يُفضل إعادة النظر في دراستها. واهتمت الدراسة الحالية بدراسة الصدق عبر صيغ التقييم المختلفة وطرائق التصحيح المعتمدة، وذلك من خلال الحصول على أدلة الصدق التقاربي بالاعتماد على نموذج (Kane, 1982) كما في دراسة (Shavelson et al., 1993). وتميزت الدراسة الحالية باختلاف ضبط تأثير التباين في صدق الاختبار في الأبعاد المُستخدمة، وباستخدام التصميم المتقاطع بدلًا من التصاميم المتداخلة لأنه يعتبر أقوى (Hung, 2009). ومما يُؤكد أهمية هذه الدراسة تطبيقها نظرية التعميم التي تُعتبر من أكثر الأساليب دقةً ومرونةً في تحديد مصادر الخطأ.

مشكلة الدراسة وأسئلتها

هدفت الدراسة إلى الكشف عن مصادر تباين الخطأ المتعددة الأكثر تأثيرًا في صدق اختبار في الرياضيات؛ فاتخاذ القرارات حول أداء المفحوصين بمستويات صدق منخفضة يؤدي إلى قرارات وتفسيرات غير صائبة، ويعود ذلك لوجود صعوبة في تحقيق أدلة تفسيرات الصدق، لعدم القدرة على تحديد مكونات تباين الخطأ ضمن بُعد واحد، كما في نظرية الاختبار التي تسمح بتقدير خطأ واحد فقط في الموقف الاختباري. ولهذا من الضروري تسليط الضوء على مُكونات تباين تقييمات الأداء للمفحوصين التي تجعلها ثابتة أو قابلة للتعميم، وذلك من خلال البحث عن طرقٍ كفيلةٍ بأن تُساهم في تحقيق أدلة صدق مُتقاربة، لذلك من المهم إعطاء المفحوص مهام متعددة لدراسة التباين في الأداء؛ لمعرفة حجم التغير بين أداء المفحوصين الناتج من صيغ المهام، وطرق التصحيح. ولهذا أشارت الدراسات إلى أن وجود مصادر تباين كالمهام، وصيغ المهام، وطرق التقييم، يؤثر في صدق قياس أداء المفحوصين. ومن أجل ذلك، استخدمت الدراسة الحالية نظرية التعميم لقُدرتها على تحديد مساهمة تباين الخطأ من كل بُعد من أبعاد الدراسة في تحليل واحد، كما أن مسألة ضعف الطرق في دراسة صدق قياس أداء المفحوصين في الاختبارات تستحق أن تنال مزيدًا من الاهتمام والبحث؛ لأنها أصبحت تُشكل مشكلة حقيقية، وفي هذا الإطار، تُحاول الدراسة الحالية الإجابة عن السؤالين التاليين ضمن دراسات التعميم:

1- ما مدى مساهمة تباين الخطأ الذي يفسره كل من بعدي (المهام، وصيغ المهام) في التباين الكلي في صدق اختبار في الرياضيات؟

2- ما مدى مساهمة تباين الخطأ الذي يفسره كل من بعدي (المهام، وطرق التصحيح) في التباين الكلي في صدق اختبار في الرياضيات؟

أهمية الدراسة

تظهر أهمية الدراسة من خلال تناولها صدق تقييمات أداء المفحوصين في اختبار رياضيات مُعد وفق المهام المركبة باستخدام نظرية التعميم، وهذا الاختبار يكشف عن مهارات التفكير العليا لدى المفحوصين من خلال مجموعة من المهام التي تعمل على تنمية طرق التفكير وحل المشكلات المركبة للمفحوصين في الرياضيات. وتُبين كيفية تقدير مصادر تباين الأخطاء ومُعاملات التعميم وتفسيرها، وتوفر إطارًا نظريًا حول نظرية التعميم وأهميتها في تحقيق التقييم الحقيقي للأداء. ومما يُعلي من شأن هذه الدراسة نُدرة الدراسات التي تناولت هذه المُشكلة، ويُتوّقع من نتائج هذه الدراسة أن تُفيد الباحثين في إجراء دراسات مُشابهة.

محددات الدراسة

1- اقتصرت عينة الدراسة على طلبة الصف الخامس الأساسي في المدارس الحكومية التابعة لمديرية التربية والتعليم للواءي الطيبة والوسطية/محافظة إربد في الفصل الدراسي الأول من العام الدراسي 2020/2019م.

2- اقتصرت أداة الدراسة المُستخدمة على (12) مهمةً مُركبة وزعت بالتساوي على (4) صيغ (التطبيق، والاستدلال، والانتقاء، والرأي).

3- اقتصر محتوى الاختبار على مجال (الأعداد والعمليات عليها) من مجالات كتاب الرياضيات المُقرر تدريسه من وزارة التربية والتعليم الأُردنية للصف الخامس الأساسي في العام الدراسي 2020/2019م.

4- اقتصرت الدراسة على إدراج الأبعاد (المهام، وصيغ المهام، وطرق التصحيح)، ولم تُدرج أبعاد أُخرى كالفترات والمصححين أو غيرها.

التعريفات الإجرائية

نظرية التعميم: مجموعة الطرق الإحصائية الأكثر مرونة مع مختلف أبعاد تقدير صدق القياسات السلوكية، المُستخدمة في هذه الدراسة لتقدير مصادر التباين المتعددة ومعاملات التعميم النسبية والمطلقة (الموضحة في المعادلات من المعادلة رقم (1) إلى المعادلة رقم (8)) تبعًا للتصميم المُستخدم.

الصدق: يُشير إلى دقة الاستدلالات وتفسير أداء المفحوصين $(P_1 - P_{301})$ في اختبار الرياضيات المُعد وفق المهام المُركبة، ويُعبر عنه بدرجة التقارب بين أربع صيغ مهام (التطبيق، والاستدلال، والانتقاء، والرأي)، ودرجة التقارب بين طرق التصحيح (التحليلي، والشمولي)، ويُعتبر مؤشرًا على الصدق التقاربي.

تقييم الأداء: يُعبر عنه بالعلامة على مهمة، أو صيغة، أو على الاختبار الذي يشتمل على (12) مهمة موزعة بالتساوي على صيغ (التطبيق، والاستدلال، والانتقاء، والرأي)، بحسب طريقة التصحيح المُستخدمة.

مصادر التباين: تُشير إلى التقديرات التي تُعبر عن اختلاف الوسط الحسابي للعلامات التي يحصل عليها المفحوص في اختبار الرياضيات تحت مختلف شروط أبعاد القياس (المهام $(T_1 - T_{12})$، صيغ المهام $(F_1 - F_4)$، طرق التصحيح$(M_1 - M_2)$)، وتقدر عن طريق متوسط المربعات وعدد المستويات، ويُعبر عنها في هذه الدراسة من خلال تباين الخطأ النسبي، وتباين الخطأ المطلق، تبعًا للتصميم المُستخدم.

المهمة المركبة: هي مجموعة من المعلومات في مجال الأعداد والعمليات عليها في الرياضيات، وضعت في سياق حياتي، تتميز بالواقعية والتركيب، وتتطلب من المفحوص استخدام جميع المعارف والمهارات والاتجاهات التي اكتسبها لإيجاد حل خلال فترة زمنية معينة ضمن معايير محددة.

منهج الدراسة

استُخدم المنهج الوصفي في إجراء هذه الدراسة، وهو شكل من أشكال التحليل والتفسير العلمي المُنظم لوصف الظاهرة المدروسة، أو مُشكلة مُحددة يتم التعبير عنها كميًا عن طريق جمع بيانات ومعلومات عنها، ويتم تصنيفها وتحليلها (Melhem, 2002). وهذه الدراسة تهتم باستكشاف مصادر تباين الخطأ المتعددة الأكثر تأثيرًا في صدق قياس أداء المفحوصين.

مجتمع الدراسة

تكون مجتمع الدراسة من جميع طلاب وطالبات الصف الخامس الأساسي للعام 2020/2019م، البالغ عددهم (1385) طالبًا وطالبة في مدارس مديرية التربية والتعليم للواءي الطيبة والوسطية/ إربد في الأردن.

عينة الدراسة

تم تطبيق أداة الدراسة على عينة مكونة من (301) مفحوص من الصف الخامس الأساسي بواقع (12) مدرسةً حكوميةً تابعةً لمديرية لواءي الطيبة والوسطية/اربد، تم اختيارها بالطريقة العشوائية.

أداة الدراسة

قام الباحثان باستخدام اختبار في الرياضيات تكوّن من (12) مهمة موزعة بالتساوي على (4) صيغ من المهام (التطبيق، والاستدلال، والانتقاء، والرأي) المتعلقة بحل المشكلات التي يواجهها المفحوص في حياته اليومية، من منهاج الصف الخامس الأساسي المقرر تدريسه في مدارس وزارة التربية والتعليم الأردنية.

وللتحقق من صدق أداة الدراسة، قام الباحثان بالتحقق من الصدق الظاهري (صدق المحتوى) من خلال عرض أداة الدراسة على (18) مُحكمًا من ذوي الاختصاص، وتم الاعتماد على نسبة الاتفاق بين تقديرات المحكمين على كل معيار من معايير شبكة تحكيم أداة الدراسة التي تراوحت قيمها بين (70%) إلى (100%)، مما يدل على أنها تتمتع بصدق محتوى مُرتفع. ولحساب الصدق التجريبي، قام الباحثان بتطبيق الاختبار على عينة استطلاعية مكونة من (40) مفحوصًا، وكانت جميع معاملات الارتباط لا تقل قيمها عن (0.7)، مما يعتبر مؤشرًا جيدًا على صدق الاختبار. وللتأكد من ثبات أداة الدراسة، تم حساب مُعامل ثبات الاستقرار من خلال تطبيق الاختبار على العينة الاستطلاعية مرتين بفاصل زمني مدته أسبوعان، ويُعبر عنه بمعامل ارتباط بيرسون بين علامات المفحوصين على الاختبار في مرتي التطبيق وكان يساوي تقريبًا (0.89)، وهذا مؤشر على أن الاختبار يتمتع بثباتٍ مرتفعٍ. وتم حساب ثبات الاتساق الداخلي بين مهام الاختبار من خلال تطبيق معادلة كرونباخ ألفا (Alpha Cronbach)، وكان يساوي تقريبًا (0.91)، مما يؤكد أن أداة الدراسة تتمتع بمستوى ثباتٍ مرتفعٍ.

إجراءات الدراسة

1- قام الباحثان بتحديد عينة الدراسة بالطريقة العشوائية، من خلال تحديد المدارس التي تم التطبيق فيها.

2- قام الباحثان ببناء اختبار رياضيات تكون من (12) مهمة موزعة بالتساوي على (4) صيغ مهام في مجال من مجالات الرياضيات لطلبة الصف الخامس الأساسي. وتم التحقق من الخصائص

السيكومترية للأداة، كما تم حساب الوسط الحسابي لإكمال المفحوصين الاختبار في العينة الاستطلاعية، وتبين أنهم يحتاجون إلى (3) جلساتٍ منفصلة لتطبيق الاختبار بزمن (45) دقيقةً لكل جلسة.

3- تم تطبيق الاختبار بمساعدة المعلمين الذين قاموا بتقديم تعليمات الاختبار بشكل موحد وموضوعي بهدف العدالة بين المفحوصين، ومن أجل ألا يكون هناك مصدر لتباين أدائهم ناتج من تقديم التعليمات أو طريقة التعامل مع المفحوصين، حيث تم تدريبهم من خلال ورشة تدريبية مُدتها (3) ساعات على مهارات التعامل مع المفحوصين، وتحفيزهم بشكل يثير دافعيتهم لإنجاز المهام بشكل نموذجي.

4- قام الباحثان بتحويل علامات المفحوصين التي حصلوا عليها من طريقة التصحيح التحليلية بالاعتماد على ميزان التصحيح، وذلك من خلال تصنيف أدائهم في فئتين، حيث تصبح علامة المفحوص الحاصل على (1) فما دون تُصبح (0)، وتصبح علامة المفحوص الحاصل على (2) فأكثر مساوية للعلامة الكلية (4) على المهمة، للحصول على درجة تقارب أداء المفحوصين عبر طرق التصحيح (التحليلية، والشمولية).

تصاميم الدراسة

استخدم الباحثان مجموعة من التصاميم المتقاطعة كُليًا (Fully Crossed Designs)، حيث إن الأبعاد والتفاعلات الحادثة بينها يمكن التعبير عنها من خلال الرموز التالية: (p) المفحوصون، (t) المهام، (f) صيغ المهام، (m) طرق التصحيح، (pf) تفاعل (مفحوص-صيغة)، (pt) تفاعل (مفحوص-مهمة)، (tf) تفاعل (مهمة-صيغة)، (pm) تفاعل (مفحوص-طريقة)، (tm) تفاعل (مهمة-طريقة)، (ptf) تفاعل (مفحوص-مهمة-صيغة)، (ptm) تفاعل (مفحوص-مهمة-طريقة).

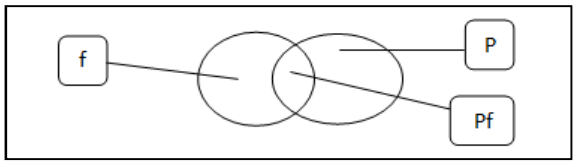
وقد أدرج الباحثان معادلات لحساب مُعامل التعميم النسبي (E_P^2) ومُعامل التعميم المطلق(ϕ) لكل تصميم من تصميمات الدراسة

التصميم الأول (p×f)

يبين الشكل (1) رسمًا توضيحيًا للتصميم الأحادي (مفحوص×صيغة)، ويُعتبر أُحاديًا لأن المفحوصين هم موضوع القياس ولا يُعتبرون مصدرًا من مصادر الأخطاء.

الشكل (1)

رسم توضيحي للتصميم الأول (p×f)



ومن خلال المعادلة (1)، والمعادلة (2) أدناه، يُحسب مُعامل التصميم النسبي ومعامل التصميم المطلق للتصميم السابق.

$$E_P^2 = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_{Pf}^2}{n'_f}} \dots\dots\dots\dots\dots (1)$$

$$\phi = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_f^2}{n'_f}+\frac{\sigma_{Pf}^2}{n'_f}} \dots\dots\dots\dots\dots (2)$$

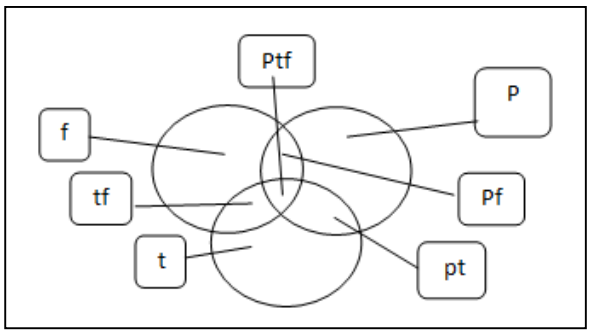
حيث (σ_P^2) تُعبر عن تباين الدرجة الشاملة، و(σ_f^2) تُعبر عن تباين صيغ المهام، و(σ_{Pf}^2) تُعبر عن تباين تفاعل (مفحوص-صيغة)، و(n'_f) تُعبر عن عدد صيغ المهام.

التصميم الثاني (p×t×f)

يبين الشكل (2) رسمًا توضيحيًا للتصميم الثنائي (مفحوص×مهمة×صيغة).

الشكل (2)

رسم توضيحي للتصميم الثاني (p×t×f)



ومن خلال المعادلة (3)، والمعادلة (4) أدناه، يُحسب مُعامل التصميم النسبي ومعامل التصميم المطلق للتصميم السابق.

$$E_P^2 = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_{Pt}^2}{n'_t}+\frac{\sigma_{Pf}^2}{n'_f}+\frac{\sigma_{Ptf}^2}{n'_t n'_f}} \dots\dots\dots\dots\dots (3)$$

$$\phi = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_t^2+\sigma_{pt}^2}{n'_t}+\frac{\sigma_f^2+\sigma_{Pf}^2}{n'_f}+\frac{\sigma_{tf}^2+\sigma_{ptf}^2}{n'_t n'_f}} \dots\dots\dots\dots\dots (4)$$

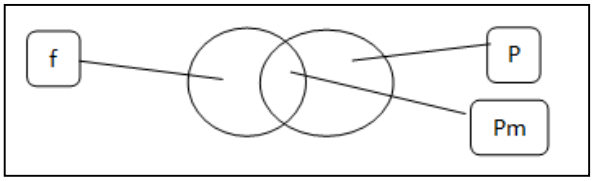
حيث (σ_P^2) تُعبر عن تباين الدرجة الشاملة، و(σ_t^2) تُعبر عن تباين المهام، و(σ_f^2) تُعبر عن تباين صيغ المهام، و(σ_{Pt}^2) تُعبر عن تباين تفاعل (مفحوص-مهمة)، و(σ_{Pf}^2) تُعبر عن تباين تفاعل (مفحوص-صيغة)، و(σ_{Ptf}^2) تُعبر عن تباين تفاعل (مفحوص-مهمة-صيغة)، و(n'_t) تُعبر عن عدد المهام، و(n'_f) تُعبر عن عدد صيغ المهام.

التصميم الثالث (p×m)

يُبين الشكل (3) رسمًا توضيحيًا للتصميم الأحادي (مفحوص×طريقة).

الشكل (3)

رسم توضيحي للتصميم الثالث (p×m)



ومن المعادلة (5)، والمعادلة (6) أدناه، يُحسب مُعامل التصميم النسبي ومعامل التصميم المطلق للتصميم السابق.

$$E_P^2 = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_{Pm}^2}{n'_m}} \dots\dots\dots\dots\dots (5)$$

$$\phi = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_m^2}{n'_m}+\frac{\sigma_{Pm}^2}{n'_m}} \dots\dots\dots\dots\dots (6)$$

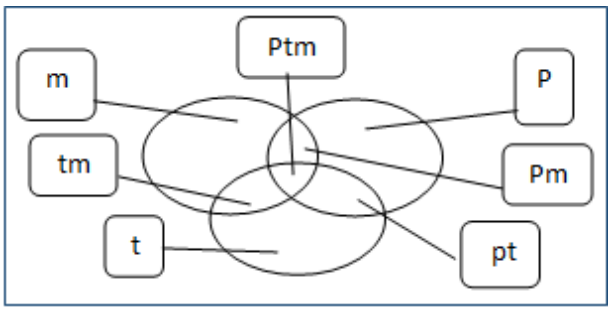
حيث أن (σ_P^2) تُعبر عن تباين الدرجة الشاملة، و(σ_m^2) تُعبر عن تباين طرق التصحيح، و(σ_{Pm}^2) تُعبر عن تباين تفاعل (مفحوص-مهمة)، و(n'_m) تُعبر عن عدد صيغ المهام.

التصميم الرابع (p×t×m)

يُبين الشكل (4) رسمًا توضيحيًا للتصميم الثنائي (مفحوص×مهمة×طريقة).

الشكل (4)

رسم توضيحي للتصميم الرابع (p×t×m)



ومن المعادلة (7)، والمعادلة (8) أدناه، يُحسب مُعامل التعميم النسبي ومعامل التعميم المطلق للتصميم السابق.

$$E_P^2 = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_{Pt}^2}{n'_t}+\frac{\sigma_{Pm}^2}{n'_m}+\frac{\sigma_{Ptm}^2}{n'_t n'_m}} \dots\dots\dots\dots\dots (7)$$

$$\phi = \frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_t^2+\sigma_{pt}^2}{n'_t}+\frac{\sigma_m^2+\sigma_{Pm}^2}{n'_m}+\frac{\sigma_{tm}^2+\sigma_{ptm}^2}{n'_t n'_m}} \dots\dots\dots\dots\dots (8)$$

حيث (σ_P^2) تُعبر عن تباين الدرجة الشاملة، و(σ_t^2) تُعبر عن تباين المهام، (σ_m^2) تُعبر عن تباين طرق التصحيح، و(σ_{Pt}^2) تُعبر عن تباين تفاعل (مفحوص-مهمة)، و(σ_{Pm}^2) تُعبر عن تباين تفاعل (مفحوص-طريقة)، و(σ_{Ptm}^2) تُعبر عن تباين تفاعل (مفحوص-مهمة-طريقة)، و(n'_t) تُعبر عن عدد المهام، و(n'_m) تُعبر عن عدد طرق التصحيح.

قام الباحثان باستخدام التصميمين الأول والثاني للإجابة عن السؤال الأول، والتصميمين الثالث والرابع للإجابة عن السؤال الثاني لتقدير دلالات الصدق التقاربي بين صيغ المهام وطرق التصحيح. ولم يتم إدراج دراسات القرار للتصاميم؛ لأن ذلك يتطلب تعديل مهام الاختبار أو الصيغ أو إجراءات التطبيق.

التحليلات الإحصائية

تم استخدام برنامج (EXCEL) في حساب مُعاملات الاتفاق والنسب المئوية لها، واستخدمت برمجية (SPSS) لحساب مُعاملات الارتباط، كما تم استخدام برمجية (EduG) لتحليل البيانات، وتم من خلالها حساب الإحصاءات الوصفية، وتحليل التباين والتعميم لكل تصميم من تصاميم الدراسة.

النتائج

الإحصاءات الوصفية لصيغ مهام الاختبار

تم حساب المتوسطات الحسابية والانحرافات المعيارية لأداء المفحوصين على صيغ المهام، كما هو مبين في الجدول (1).

الجدول (1)

الإحصاءات الوصفية لصيغ مهام الاختبار

صيغة المهمة	التطبيق	الاستدلال	الانتقاء	الرأي
الوسط الحسابي	5.425	4.167	3.996	2.105
الانحراف المعياري	5.416	4.998	4.322	2.382

يُلاحظ من الجدول (1) أن المتوسطات الحسابية لصيغ الاختبار كانت مُتباينة وتراوحت بين (5.425) و (2.105)، حيث كان أداء المفحوصين أعلى صيغة مهام التطبيق من غيرها، وعلى صيغ مهام الرأي الأقل. وكانت الانحرافات المعيارية مختلفة؛ ففي مهام التطبيق كان الانحراف المعياري (5.416) مما يدل على تباين الأداء، وعلى مهام الرأي كان الانحراف المعياري (2.382) مما يدل على تجانس الأداء.

نتائج السؤال الأول

للإجابة عن السؤال الأول، كانت نتائج تحليل التباين والتعميم كما يلي:

الجدول (2)

نتائج تحليل التباين للتصميم (مفحوص×صيغة) للمهام الكلية للاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	المكونات				
				العشوائية	المختلطة	المصححة	النسبة المئوية	الخطأ المعياري
مفحوص (P)	6736.45	300	22.455	4.766	4.766	4.766	%53.9	0.459
صيغة (F)	625.592	3	208.531	0.682	0.682	0.682	%7.7	0.438
مفحوص-صيغة	3050.679	900	3.39	3.39	3.39	3.39	%38.4	0.16
المجموع	10412.72	1203					%100	

تُشير النتائج في الجدول (2) إلى أن أكبر مُكون تباين كان مصدره تفاعل (مفحوص-صيغة) ويساوي (53.9%) من مكونات التباين الكلي، وأن أقل مُكون تباين كان مصدره الصيغة ويساوي (7.7%) من مكونات التباين الكلي، وأن أكبر خطأ معياري ناتج من الصيغة ويساوي (0.438)، في حين أن أقل خطأ معياري ناتج من تفاعل (مفحوص-صيغة) ويساوي (0.16).

الجدول (3)

نتائج تحليل التعميم للتصميم (مفحوص×صيغة) للمهام الكلية للاختبار

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية لتباين الخطأ النسبي	تباين الخطأ المطلق	النسبة المئوية لتباين الخطأ المطلق
مفحوص (P)	4.766	.....	.....	.....	.....
صيغة (F)	.....	.....	.....	0.17	%16.7
مفحوص-صيغة	.....	0.847	%100	0.847	%83.3
مجموع التباينات	4.766	0.847	%100	1.018	%100
الانحراف المعياري	2.183	الخطأ المعياري النسبي	0.921	الخطأ المعياري المطلق	1.009
معامل التعميم النسبي		0.85			
معامل التعميم المطلق		0.82			

تُبين النتائج في الجدول (3) أن أكبر مُكون لتباين الخطأ في القياس النسبي والمطلق راجع إلى تفاعل (مفحوص-صيغة) الممزوج بالأخطاء العشوائية ويُساوي كل منهما (100%) و (83.3%) على التوالي من مكونات التباين الكلي، مما يدل على أن أداء المفحوصين كان متباينًا تبعًا لنوع الصيغة. ويتضح أن ثاني أكبر مُكون للتباين راجع إلى الصيغة ويُساوي (16.7%) من مكونات التباين الكلي، مما يدل على أن هناك اختلافًا في مستوى صعوبة صيغ المهام، وأن مُكون تباين الصيغة كان أقل من مُكون تباين تفاعل (مفحوص-صيغة) نتيجة وجود أخطاء عشوائية ممزوجة معه. وفي المُقابل، كان مُعامل التعميم النسبي (0.85) أكبر من مُعامل التعميم المطلق (0.82)، وهذا يُشير إلى أن مُعاملات التعميم جاءت مقبولة، وإلى أنه يُمكن الوصول لمُعاملات تعميم مقبولة والاستدلال على الصدق التقاربي من خلال صيغ المهام.

الجدول (4)

نتائج تحليل التباين للتصميم (مفحوص×مهمة×صيغة) للمهام الكلية للاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	المكونات				
				العشوائية	المختلطة	المصححة	النسبة المئوية	الخطأ المعياري
مفحوص (P)	2537.797	300	8.459	0.487	0.487	0.487	%18.5	0.059
مهمة (T)	621.606	2	310.80	0.222	0.222	0.222	%8.4	0.183
صيغة (F)	316.04	3	105.34	0.069	0.069	0.069	%2.6	0.078
مفحوص-مهمة	1324.521	600	2.208	0.207	0.207	0.207	%7.8	0.034
مفحوص-صيغة	1607.683	900	1.786	0.135	0.135	0.135	%5.1	0.032
مهمة-صيغة	255.779	6	42.63	0.137	0.137	0.137	%5.2	0.071
مفحوص-مهمة-صيغة	2486.722	1800	1.382	1.382	1.382	1.382	%52.4	0.046
المجموع	9150.148	3611					%100	

وتبين النتائج في الجدول (4) أن أكبر مُكون كان مصدره تفاعل (مفحوص×مهمة×صيغة) الممزوج بالأخطاء العشوائية ويساوي (52.4%) من مكونات التباين الكلي، وأن أقل مُكون كان مصدره الصيغة ويساوي (2.6%) من مكونات التباين الكلي. ويتضح أن أكبر خطأ معياري ناتج من المهمة ويساوي (0.183)، في حين أن أقل خطأ معياري ناتج من تفاعل (مفحوص-صيغة) ويساوي (0.032).

الجدول (5)

نتائج تحليل التعميم للتصميم (مفحوص×مهمة×صيغة) للمهام الكلية للاختبار

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية لتباين الخطأ النسبي	تباين الخطأ المطلق	النسبة المئوية لتباين الخطأ المطلق
مفحوص (P)	0.487	.....	.....	.....	.....
مهمة (T)	.....	.....	.....	0.074	%23.1
صيغة (F)	.....	.....	.....	0.017	%5.4
مفحوص-مهمة	.....	0.069	%31.6	0.069	%21.5
مفحوص-صيغة	.....	0.034	%15.5	0.034	%10.5
مهمة-صيغة	.....	.....	.....	0.011	%3.6
مفحوص-مهمة-صيغة	.....	0.115	%52.9	0.115	%35.9
مجموع التباينات	0.487	0.218	%100	0.320	%100
الانحراف المعياري	0.698	الخطأ المعياري النسبي	0.467	الخطأ المعياري المطلق	0.566
معامل التعميم النسبي		0.69			
معامل التعميم المطلق		0.60			

تُشير النتائج في الجدول (5) أن أكبر مُكون لتباين الخطأ في القياس النسبي والمطلق راجع إلى تفاعل (مفحوص-مهمة-صيغة) الممزوج بالأخطاء العشوائية ويساوي كل منهما (52.9%) و (35.9%) على التوالي من مكونات التباين الكلي، مما يدل على أن أداء المفحوصين كان متباينًا تبعًا لنوع المهمة والصيغة المعطاة لهم. كذلك يُلاحظ أن ثاني أكبر مُكون للتباين راجع إلى المهمة ومقداره (23.1%)، مما يدل على أن هناك اختلافًا في مستوى صعوبة المهمة المعطاة للمفحوصين. ومن ثم يأتي مُكون تباين تفاعل (مفحوص-مهمة) ومقداره في القياس النسبي والمطلق على التوالي (31.6%) و(21.5%)، ومن ثم يأتي مُكون تباين تفاعل (مفحوص-صيغة) ومقداره في القياس النسبي والمطلق على التوالي (15.5%) و(10.5%). ويلاحظ أن باقي مُكونات التباين كانت قليلة نوعًا ما. وفي المُقابل، فإن مُعامل التعميم النسبي (0.69) أكبر من المطلق (0.60)، كما أن مُعاملات التعميم جاءت متدنية، وهذه المُعاملات تدل على نقص في درجة تقارب أداء المفحوصين في صيغ الاختبار نتيجة إدماج بُعد المهمة.

الإحصاءات الوصفية لطرق تصحيح مهام الاختبار

اعتمد الباحثان طريقتين لتصحيح مهام الاختبار، وكانت المتوسطات الحسابية والانحرافات المعيارية لأداء المفحوصين في الاختبار باستخدام هذه الطرق كما في الجدول (6).

الجدول (6)

الإحصاءات الوصفية لطرق تصحيح مهام الاختبار

طريقة التصحيح	التحليلية	الشمولية
الوسط الحسابي	1.308	1.683
الانحراف المعياري	1.519	1.799

يُلاحظ من الجدول (6) أن المتوسطات الحسابية لطرق تصحيح مهام الاختبار كانت مُتباينة؛ ففي طريقة التصحيح التحليلية كان الوسط الحسابي (1.308) في الطريقة الشمولية (1.683). وكانت الانحرافات المعيارية مُتباينة أيضًا تبعًا لطريقة التصحيح؛ ففي الطريقة التحليلية كان الانحراف المعياري (1.519)، وهو أقل من (1.799) في الطريقة الشمولية، مما يدل على تجانس أداء المفحوصين في الطريقة التحليلية كان بشكل أكبر منه في الطريقة الشمولية.

نتائج السؤال الثاني

للإجابة عن السؤال الثاني، كانت نتائج تحليل التباين وتحليل التعميم كالاتي:

الجدول (7)

نتائج تحليل التباين للتصميم (مفحوص×طريقة) للمهام الكلية للاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	متوسط المربعات	المكونات				
				العشوائية	المختلطة	المصححة	النسبة المئوية	الخطأ المعياري
مفحوص (P)	5582.249	300	18.608	6.032	6.032	6.032	%42.6	0.803
طريقة (M)	485.671	1	485.671	1.592	1.592	1.592	%11.2	1.317
مفحوص-طريقة	1963.284	300	6.544	6.544	6.544	6.544	%46.2	0.533
المجموع	8031.204	601					%100	

تُوضح النتائج في الجدول (7) أن أكبر مُكون للتباين كان مصدره تفاعل (مفحوص-طريقة) ويساوي (46.2%) من مكونات التباين الكلي، وأقل مُكون كان مصدره الطريقة ويساوي (11.2%) من مكونات التباين الكلي. وتبين أن أكبر خطأ معياري ناتج من الطريقة ويساوي (1.317). في المُقابل، فإن أقل خطأ معياري ناتج من تفاعل (مفحوص-طريقة) ويساوي (0.533).

الجدول (8)

نتائج تحليل التعميم للتصميم (مفحوص×طريقة) للمهام الكلية للاختبار

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية لتباين الخطأ النسبي	تباين الخطأ المطلق	النسبة المئوية لتباين الخطأ المطلق
مفحوص (P)	6.032	.....	.....	.....	.....
طريقة (M)	.....	.....	.....	0.796	%19.6
مفحوص-طريقة	.....	3.272	%100	3.272	%80.4
مجموع التباينات	6.032	3.272	%100	4.068	%100
الانحراف المعياري	2.456	الخطأ المعياري النسبي	3.272	الخطأ المعياري المطلق	4.068
معامل التعميم النسبي		0.65			
معامل التعميم المطلق		0.60			

تُبين النتائج في الجدول (8) أن أكبر مُكون لتباين الخطأ في القياس النسبي والمطلق راجع إلى تفاعل (مفحوص-طريقة) الممزوج بالأخطاء العشوائية ويساوي كل منهما (100%) و (80.4%) على التوالي من مكونات التباين الكلي، وهي أقل من مُكون تباين الطريقة (19.6%)، مما يدل على تباين أداء المفحوصين تبعًا لطريقة التصحيح ووجود أخطاء عشوائية ممزوجة معه. ويُلاحظ أن قيمة كل من مُعامل التعميم النسبي ومعامل التعميم المطلق كانت على التوالي (0.65) و (0.60) وهي متدنية، ولا يُمكن من خلال طرق التصحيح وحدها الوصول لمُعاملات تعميم مقبولة.

الجدول (9)

نتائج تحليل التباين للتصميم (مفحوص×مهمة×طريقة) للمهام الكلية للاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	المكونات				
				العشوائية	المختلطة	المصححة	النسبة المئوية	الخطأ المعياري
مفحوص (P)	3245.601	300	10.819	0.360	0.360	0.360	%16.8	0.037
مهمة (T)	2702.492	11	245.681	0.344	0.344	0.344	%16.1	0.162
طريقة(M)	155.825	1	155.825	0.032	0.032	0.032	%1.5	0.036
مفحوص-مهمة	3996.367	3300	1.211	0.011	0.011	0.011	%0.5	0.021
مفحوص-طريقة	647.212	300	2.157	0.081	0.081	0.081	%3.8	0.015
مهمة-طريقة	425.983	11	38.726	0.125	0.125	0.125	%5.8	0.051
مفحوص-مهمة-طريقة	3925.878	3300	1.190	1.190	1.190	1.190	%55.6	0.029
المجموع	15099.35	7223					%100	

تُشير النتائج في الجدول (9) أن أكبر مُكون للتباين كان مصدره تفاعل (مفحوص×مهمة×طريقة) الممزوج بالأخطاء العشوائية ويساوي (55.6%) من مكونات التباين الكلي، وأقل مُكون كان مصدره تفاعل (مفحوص-مهمة) ويساوي (0.5%) من مكونات التباين الكلي، وتبين أن أكبر خطأ معياري ناتج من المهمة ويساوي (0.162)، وأقل خطأ معياري ناتج من تفاعل (مفحوص-طريقة) ويساوي (0.015).

الجدول (10)

نتائج تحليل التعميم للتصميم (مفحوص×مهمة×طريقة) للمهام الكلية للاختبار

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية لتباين الخطأ النسبي	تباين الخطأ المطلق	النسبة المئوية لتباين الخطأ المطلق
مفحوص (P)	0.36	.....	.....	.....	.....
مهمة (T)	.....	.....	.....	0.029	%20.4
طريقة (M)	.....	.....	.....	0.016	%11.4
مفحوص-مهمة	.....	0.001	%1	0.001	%0.6
مفحوص-طريقة	.....	0.040	%44.4	0.04	%28.7
مهمة-طريقة	.....	.....	.....	0.005	%3.7
مفحوص-مهمة-طريقة	.....	0.050	%54.6	0.05	%35.2
مجموع التباينات	0.36	0.091	%100	0.141	%100
الانحراف المعياري	0.6	الخطأ المعياري النسبي	0.301	الخطأ المعياري المطلق	0.375
		معامل التعميم النسبي	0.80		
		معامل التعميم المطلق	0.72		

وأخيرًا، تُشير النتائج في الجدول (10) إلى أن أكبر مُكون لتباين الخطأ في القياس النسبي والمطلق راجع إلى تفاعل (مفحوص-مهمة-طريقة) الممزوج بالأخطاء العشوائية ويساوي كل منهما (54.6%) و(35.2%) على التوالي من مكونات التباين الكلي، مما يدل على أن أداء المفحوصين كان متباينًا تبعا لنوع المهمة والطريقة المطبقة للحصول على تقديراته. ويُلاحظ أن ثاني أكبر مُكون للتباين راجع إلى تفاعل (مفحوص-طريقة) ومقداره في القياس النسبي والمطلق على التوالي (44.4%) و(28.7%)، مما يدل على أن هناك اختلافًا في تقديرات المفحوصين تبعًا لطريقة التصحيح. ومن ثم يأتي مُكون تباين المهمة ومقداره (20.4%) في القياس المطلق، يليه مُكون تباين الطريقة في القياس نفسه (11.4%)، بينما كانت باقي مُكونات التباين قليلة نوعاً ما. وكانت مُعامل التعميم النسبي (0.80) أعلى من مُعامل التعميم المطلق (0.72)، مما يُشير إلى أن مُعاملات التعميم جاءت مقبولة، وهذه المُعاملات تدل على درجة تقارب أداء المفحوصين في طرق تصحيح الاختبار. وقد رفع إدماج بُعد المهمة من مُعاملات التعميم.

مناقشة نتائج السؤال الأول

بينت نتائج الدراسة أن أكبر مصدر للتباين في التصميم (مفحوص×صيغة) يعود إلى تفاعل (مفحوص-صيغة) الممزوج بالأخطاء العشوائية، أما مصدر تباين الصيغة فقد كان ضعيفًا مقارنةً بمصدر تباين تفاعل (مفحوص-صيغة)، مما يدل على أن أداء المفحوص يتباين من صيغة إلى أخرى؛ إذ إن أداء المفحوص كان مرتفعًا في صيغة ومنخفضًا في غيرها، مما يُشير إلى أن مصدر التباين الأكثر تأثيرًا في الدرجة الشاملة هو تباين (مفحوص-صيغة). كذلك توصلت الدراسة إلى أن أكبر مصدر للتباين في التصميم (مفحوص×مهمة×صيغة) يعود إلى تفاعل (مفحوص-مهمة-صيغة) الممزوج بالأخطاء العشوائية، وهذا يدل على أن الوسط الحسابي لأداء المفحوصين يختلف باختلاف المهمة والصيغة، وذلك لاستخدام المفحوص طُرقًا مختلفة لحل المهام بغض النظر عن الصيغة التي تنتمي لها المهمة. وتتفق نتائج الدراسة الحالية مع دراسة (Shavelson et al., 1993) في أن أكبر مصدر للتباين هو تفاعل (مفحوص-مهمة-صيغة). وقد أكدت دراسات (Lane et al., 1996; Mcbee & Barnes, 1998; Webb et al., 2000; Tabaa., 2020) وجود مصادر للتباين ناتجة من المهمة، وذلك لاختلاف درجة صعوبة المهام، وتباين تفاعل (مفحوص-مهمة).

هذا في حين جاءت مُعاملات التعميم مقبولة ضمن دراسة التعميم لتصميم (مفحوص×صيغة)؛ إذ كانت أفضل من مُعاملات التعميم لتصميم (مفحوص×مهمة×صيغة)، وذلك بسبب ارتفاع مصدر تباين المهمة في التصميم (مفحوص×مهمة×صيغة)، لأنه تم إدماج بعد المهمة في دراسات التعميم، مما جعلها تؤثر بشكل أكبر في تباين الدرجة الشاملة. ويعود ذلك لتباين طرق الحل للصيغ والمهام. وتتفق هذه النتيجة مع دراسة (Shavelson et al., 1993) في أن مُعاملات التعميم في تصميم (مفحوص×مهمة×صيغة) جاءت متدنية، بسبب مصادر التباين (مفحوص-مهمة-صيغة) الممزوجة بالأخطاء العشوائية، و(مفحوص-مهمة) و (مهمة)، في حين أن مُعاملات التعميم في تصميم (مفحوص×صيغة) جاءت ضمن المدى المقبول وذلك بسبب عدم إدخال بعد المهمة. وتختلف هذه الدراسة مع دراستي (Ruiz-Primo & Shavelson, 1996; Smith & Kulikowich, 2004) في عدم الحصول على معاملات تعميم مقبولة؛ لاستخدام المفحوصين طرقًا مختلفةً في حل المهام.

مناقشة نتائج السؤال الثاني

توصلت الدراسة إلى أن أكبر مصدر للتباين في التصميم (مفحوص×طريقة) يعود إلى تفاعل (مفحوص-طريقة) الممزوج بالأخطاء العشوائية، أما مصدر تباين الطريقة فقد كان ضعيفًا مقارنةً بمصدر تباين تفاعل (مفحوص-طريقة)، مما يدل على وجود تباين لأداء المفحوص من طريقة إلى أخرى؛ إذ إن أداء المفحوص كان مرتفعًا في طريقة تصحيح ومنخفضًا في الطريقة الأخرى، مما يُشير إلى أن مصدر التباين الأكثر تأثيرًا في الدرجة الشاملة هو تباين (مفحوص-طريقة). كذلك توصلت الدراسة إلى أن أكبر مصدر للتباين في التصميم (مفحوص×مهمة×طريقة) يعود إلى تفاعل (مفحوص-مهمة-طريقة) الممزوج بالأخطاء العشوائية، وهذا يدل على أن الوسط الحسابي لأداء المفحوصين يختلف باختلاف المهمة والطريقة التي يتم بها تصحيح المهام. وقد أكدت الدراسات السابقة وجود مصادر للتباين ناتجة من المهمة وذلك لاختلاف درجة صعوبة المهام لجميع المفحوصين، ووجود مصادر للتباين ناتجة من تفاعل (مفحوص-مهمة). وذلك لاستخدام المفحوص طُرقًا مختلفة لحل المهام بغض النظر عن الطريقة التي صُححت بها المهَمة. وتتفق نتائج الدراسة الحالية مع دراستي (Huang, 2009; Shavelson et al., 1993) في أن أكبر مصدر للتباين هو تفاعل (مفحوص-مهمة-طريقة)، في حين أن دراسة (Huang, 2009) وجدت أن طريقة التصحيح أو التقييم غير دالة إحصائيًا مما يجعلها تختلف مع هذه الدراسة في مدى فعالية طريقة التصحيح في صدق الاختبار.

في المُقابل، جاءت مُعاملات التعميم مقبولة ضمن دراسة التعميم لتصميم (مفحوص×مهمة×طريقة) وأفضل من مُعاملات التعميم لتصميم (مفحوص×طريقة)، بسبب إدماج بُعد المهمة، مما جعلها تؤثر بشكل أكبر في تباين الدرجة الشاملة، بسبب مصادر التباين (مفحوص-مهمة-طريقة) الممزوج بالأخطاء العشوائية، و(مفحوص-مهمة)، و(مهمة)، ويعود ذلك لتباين طرق الحل وطرق التصحيح وإدراج بُعد المهمة. واختلفت نتائج الدراسة الحالية مع دراسات (Lane et al.,1996; Webb et al., 2000; Nie et al., 2007; Tabaa, 2020) التي بينت أن أكبر مصدر للتباين هو تفاعل (مفحوص-مهمة)، وذلك نتيجة إدخال أبعاد مختلفة عن الأبعاد المُستخدمة في الدراسة الحالية.

الاستنتاجات والمُقترحات

بناءً على نتائج الدراسة المتعلقة بسؤالي الدراسة، نستنتج أن التصميم (مفحوص×صيغة) قدم مؤشرات صدق مرتفعة مقارنةً مع التصميم (مفحوص×مهمة×صيغة)، ويعود ذلك إلى إدراج بُعد المهمة. كذلك فإن التصميم (مفحوص×مهمة×طريقة) قدم مؤشرات صدق مقبولة مقارنةً مع التصميم (مفحوص×طريقة)، ويعود ذلك إلى تباين طُرق الحل والتصحيح وإدراج بُعد المهمة أيضًا، ويُمكن استنتاج أن اختلاف النتائج في السؤالين الأول والثاني يعود إلى أن عدد صيغ المهام المُدرجة ضمن السؤال الأول (4) صيغ، بينما كان عدد طرق التصحيح المُدرجة ضمن السؤال الثاني طريقتين فقط، مما يعني أن عدد طرق التصحيح أقل من عدد صيغ المهام. وتجدر الإشارة إلى أن صيغ المهام مستقلة عن بعضها البعض نوعًا ما من حيث افتراضاتها وطُرق بنائها، بينما طرق التصحيح مرتبطة مع بعضها البعض؛ إذ إن طريقة التصحيح الشمولية تعتمد على تقديرات طريقة التصحيح التحليلية، وعليه فإن مُعاملات التعميم في التصميمين (مفحوص×صيغة)، و(مفحوص×مهمة×طريقة) توفر لنا إمكانية تماثل النتائج عبر مستويات مختلفة من الأبعاد، وبالتالي

تُعتبر دليلاً على الصدق التقاربي في كل تصميم على حدة. أما اختلاف مُعاملات التعميم باختلاف الأبعاد، فهو يُعد دليلًا على الصدق التمييزي.

بناءً على ما سبق من نتائج توصلت إليها الدراسة، تُوصي الدراسة الحالية بإجراء بعض الدراسات المستقبلية التي من شأنها أن تُثري أدبيات القياس والتقويم باستخدام نظرية التعميم، وخاصة باللغة، العربية وهي كالآتي:

1- الكشف عن مدى مساهمة مصادر تباين الخطأ من أدوات قياس أخرى (الاستبانات، اختبارات الاستعداد،...).

2- إجراء دراسات في مجالات أخرى كالزراعة والصناعة والطب، مثل الكشف عن فعالية الأدوية.

3- إجراء دراسات تهدف إلى محاولة تقليل الخطأ التجريبي والتحكم بالآثار التجريبية.

4- مقارنة مؤشرات الصدق من خلال برمجيات (EduG) و (GNOVA) في ظروف اختبارية مختلفة.

5-إجراء دراسات ذات أبعاد مُختلفة كمؤشرات استدلالية على الصدق التقاربي أو الصدق التميزي.

References

Abo Zinah, F. & Ababneh, A. (2007). *Mathematics teaching curricula for the first grades*. Dar Al-Maserah for Publishing and Distribution.

Alharbi, K. & Alharbi, E. (2017). Reliability indicators of using generalization theory and construct-validity evidences for Mawhiba creativity test. *Taibah University Journal for Educational Sciences*, 12(3). 425-441. http://search.shamaa.org/FullRecord?ID=267259

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

Allen, J. & Yen, W. (1979). *Introduction to measurement theory*. Brooks 1 Cole Publishing Co.

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

Audeh, A. (2010). *Measurement and evaluation in the teaching process*. Dar Al-Amal for Publishing and Distribution.

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

Brennan, L. (2001). *Generalizability theory*. Springer-Verlag.

Brennan, L. & Kane, T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277-259.

Chen, E., Niemi, D., Wang, J., Wang, H. & Mirocha, J. (2007). *Examining the generalizability of direct writing assessment tasks*. Technical Report 718. CRESST: University of California, Los Angeles.

Doyle, W. (1983). Academic work. *Review of Educational Research*, 53(2), 159-199.

Huang, Ch. (2009). Magnitude of task-sampling variability in performance assessments: 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, 123-139.

Johnson, R. L., Penny, J. A. & Gordan, B. (2009). *Assessing performance: Designing, scoring and validating performance tasks*. Guilford Press.

Kane, T. (1982). A sampling model of validity. *Applied Psychological Measurement***,** 6, 125-160.

Lane, S., Liu, M., Ankenmann, D. & Stone, A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, 33(1), 71-92.

Linn, L. (1994). Evaluating the technical quality of proposed national examination systems. *American Journal of Education*, 102(4), 565-580.

Linn, L., Baker, L. & Dunbar, B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 15-21.

Mabe, M. (2014). *Validation of social skills construct using multitrait multimethod and generalizability approaches.* Unpublished Doctoral Dissertation, University of Rhode Island

McBee, M. & Barnes, B. (1998). The generalizability of a performance assessment measuring achievement in eight-grade mathematics. *Applied Measurement in Education*, 11(2), 179-194.

Melhem, S. (2002). Research methods in educational and psychological scieces. Dar Al-Maserah for Publishing and Distribution.

Messick, S. (1995). Validity of psychological-assessment validation of inferences from person's responses and performances as a scientific inquiry into score meaning. *American Psychologist***,** 50(9), 741-749.

Ministry of education. (2004). *Evaluation strategies and tools: Theoretical framework.* Amman: Jordan.

Nie, Y., Yeo, M. & Lau, S. (2007). Application of generalizability theory in the investigation of the quality of journal writing in mathematics. *Studies in Educational Evaluation*, 33, 371-383.

Ruiz-Primo, A. & Shavelson, J. (1996). Rhetoric and reality in science performance assessments: An update**.** *Journal of Research in Science Teaching*, 33(10), 1045-1063.

Shavelson, J., Baxter, P. & Geo, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215-232.

Shavelson, J. & Webb, M. (2009), *Generalizability theory*: *A primer*. Sage Publications.

Smith, E. & Kullikowich, J. (2004). An application of g-theory and many rash measurement using complex problem-solving skills assessment, *Educational and facet Measurement***,** 64(4), 617-639.

Tebaa, F. (2020). Using generalizability theory in estimating reliability of a mathematical competence assessment test of fourth-year primary-school students. *Jordan Journal of Educational Sciences*. 16(1). 1-18.

Tanilon, J., Segers, M., Vedder, P. & Tillema, H. (2009). Development and validation of an admission test designed to assess samples of performance on academic tasks. *Studies in Educational Evaluation*, 35, 168–173.

Urbina, A. (2015). Psychological testing (S. Allam, Trans.). Dar Al-Fikr Al-Arabi. 1997.

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

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