الكشف عن إسهام مصادر التباين المتعددة في ثبات اختبار في الرياضيات باستخدام نظرية التعميم

محمد أكرم أحمد العرايضة (1)، **ونضال كمال الشريفين** (2)

وزارة التربية والتعليم-الأردن & كلية التربية، جامعة اليرموك، الأردن

(قدم للنشر في 24/ 8/ 1442هـ؛ وقبل للنشر في 4/ 1/ 1443هـ)

المستخلص: هدفت الدراسة إلى الكشف عن إسهام مصادر التباين المتعددة في ثبات اختبار في الرياضيات باستخدام نظرية التعميم، وذلك بتقدير حجم تباين الخطأ المُفسر من الأبعاد (المهام، المُصححين، الفترات الزمنية) في التباين الكلي، والشروط التي يُمكن بها الوصول لمستويات أفضل لثبات الاختبار؛ حيث تكونت عينة الدراسة من (243) مفحوصًا من الصف الخامس الأساسي، طُبق عليهم اختبار في الرياضيات تكون من (12) مهمة مُركبة في مجال الأعداد والعمليات عليها، توزعت على (4) صيغ بالتساوي (التطبيق، الاستدلال، الانتقاء، الرأي). قُيّم أداء المفحوصين من قبل (3) مصححين، وطُبق عليهم الاختبار مرتين بفاصل زمني مدته أُسبوعان، استخدم الباحثان التصميم (مفحوص×مهمة×مصحح×فترة) المتقاطع كُليًا، واستخدمت برمجية (EduG) لتحليل البيانات. وأظهرت النتائج أن أكثر مصادر التباين تأثيرًا في مُعاملات التعميم تفاعل (مفحوص-مهمة)، وأن إدراج بُعد الفترة الزمنية أسهم في تحسين مُعاملات التعميم، كما أن زيادة عدد المهام وعدد الفترات الزمنية يرفع من مُعاملات التعميم، وأن زيادة عدد المصححين لا يرفع من مُعاملات التعميم بشكل جوهري.

الكلمات المفتاحية: دراسات التعميم، دراسات القرار، الثبات، تباين الخطأ.

The Detection of the contribution of multiple sources of Variance in the Reliability of a test in Mathematics by using of Generalizability Theory

Mohammad Akram Ahmad Al-araida and Nedal Kamal Alshraifin

Ministry of Education-Jordan & Yarmouk University-Jordan

(Received 6/4/2021; Accepted 12/8/2021)

Abstract: This study aimed to detect the contribution of multiple sources of the variance in the reliability of a test in mathematics by using the Generalizability Theory. The magnitude of the error variance that is explained from the facets (Tasks, Judges, and Occasions) in the total variance, and detect the conditions through which they can be better in levels of test reliability. The study sample consisted of (243) students from the fifth grade who were applied on a mathematics test that consisted of (12) Complex tasks in domain of numbers and operations on them, they were distributed on (4) formulas equally (application, inference, selection, and opinion), The performance of the persons was evaluated by (3) judges. The test was applied to them twice with a time two-week interval. The researchers used the completely crossed design (Person×Task×Judge×Occasion), and used (EduG) software to analyze the data. The results indicated that the largest sources of error variance on Generalizability coefficients were interaction (person-task), including after the Occasion contributed to improving Generalizability coefficients. The results have found that increasing the number of tasks and the number of occasions increases the Generalizability coefficients, and increasing the number of judges does not substantially increase the Generalizability coefficients.

Keywords: Generalizability Studies, Decision Studies, Reliability, Error Variance.

---

(1) دكتوراة في القياس والتقويم-وزارة التربية والتعليم-الأردن .

(1) PhD in Measurement and Education, Ministry of Education -Jordan.

البريد الإلكتروني*e-mail:mohammadaraida@yahoo.com*

(2) أستاذ دكتور في القياس والتقويم-كلية التربية-جامعة اليرموك-الأردن.

(2) Professor of Measurement and Evaluation- Faculty of Education- Yarmouk University-Jordan

البريد الإلكتروني*e-mail:nshraifin@yahoo.com*

المقدمة:

شَهد علم القياس والتقويم الحديث خلال الفترة السابقة تطورًا في نظرياته المختلفة، وتغييرات جوهرية في مفاهيمه، وطرقه وأساليبه وتقنياته، عن طريق العديد من الإصلاحات التربوية على الأنظمة التربوية في العالم، على أثر الانتقادات التي أثارتها البحوث التربوية لاستخدامات التقويم التقليدي بمُختلف صُورِه، حيث اهتم الخبراء بابتكار طرق وأساليب تُرشد وتُوجه عمليات القياس والتقويم المُعاصر.

إذْ إنَّ عملية التقويم تحتاج إلى أدوات قياس دقيقة، ذات خصائص سيكومترية عالية؛ من أجل الحصول على معلومات تُساعد الباحثين على اتخاذ القرارات المختلفة في الميادين التربوية، إلا أن التحقق من دقة أدوات القياس ليس بالأمر السهل؛ فالخصائص السيكومترية تجعل من بناء أدوات القياس أمرًا بالغ الأهمية؛ لذلك حظيت باهتمام الباحثين والعلماء في المجالات المتعددة ومنها الاختبارات بأنواعها المُختلفة التي تُعد من أهم أدوات القياس والتقويم، وأكثرها استخدامـًا (عودة، 2010م)؛ حيث ركز مُعظمهم على دراسة ثباتها (Reliability) بالاعتماد على التقويم التقليدي القائم على الاختبارات الموضوعية، والتي تأخذ بعين الاعتبار السُرعة في الإجابة، وغير قادرة على قياس عمليات التفكير العليا كالمهام، وتتلاءم نظرية القياس التقليدية (Classical Test Theory) في تقدير الخصائص السيكومترية لهذه الاختبارات لأحادية أبعادها (علام، 2004م).

من هُنا جاءت أهمية دراسة مفهوم الثبات في الاختبارات والمقاييس النفسية؛ إذ من المُؤكد البحث عن الطرق التي تساعد على تقدير أداء المفحوصين (Persons) ضمن شروط أكثر موضوعية ودقة، فتقييمات الأداء (Performance Assessment) تمتاز بأنها تتطلب معاينة أداء المفحوصين ضمن مهام (Tasks) متعددة تنتمي إلى نفس نطاق المهام، وتقدير أداء المفحوصين بواسطة مصححين (Judges) متعددين يتطلب نوعًا من الاتساق، وضمن فترات زمنية مختلفة (Occasions) (Brennan, 2001).

وفي ظلِّ تطور نظريات القياس، اتجهت العديد من الأنظمة التربوية في العالم إلى إحداث تغييرات في أبعاد العملية التربوية نحو تقييمات الأداء؛ باعتبارها حلًّا لتطوير أداء المفحوصين على حل المشكلات، وتُلامس الواقع في الحياة اليومية، حيث إن نظرية القياس التقليدية (CTT) غير قادرة على التمييز بين أخطاء القياس، لذلك من الضروري تطبيق نظرية لدراسة الثبات المُتأثر بأبعاد متعددة (Parkes, 2000).

ولذلك أثمرت جهود العلماء بالانتقال من تطبيق نظرية القياس التقليدية التي تعطي نفس القيمة

لأخطاء القياس لجميع المفحوصين، إلى تطبيق نظرية التعميم (Generalizability Theory)، التي تُعالج هذه المُشكلة بواسطة تحليل التباين (Analysis of Variance) (Crocker & Algina, 1986). وعليه، فإن نظرية التعميم(GT) وفرت نماذج وطرق تُمكن الباحث من الفصل بين مصادر الأخطاء المتعددة، كما أنها تصف طرق تقدير حجم التباين الذي تسهم فيه وكيفية تقدير ثبات القياسات المُلاحظة التي نحصل عليها من أدوات قياس مختلفة (Allen & Yen, 1979). إذ إنها تستجيب لمختلف أبعاد القياس المعقدة باعتبارها طريقة إحصائية مناسبة لتقدير ثبات القياسات السلوكية؛ لأنها تسمح بضبط مختلف أبعاد القياس ومعالجتها، وتبحث عن مختلف مصادر الخطأ التي تؤثر في ثبات تقييمات الأداء، وتُقدم للباحثين مُعاملات التعميم النسبية والمطلقة (Relative and Absolute Generalizability Coefficients)، وتُعطي طرقًا فعالة في تحسين ثبات تقييمات الأداء Brennan, 2001).

لذلك، فإن أهمية وجود أدوات القياس والتقويم المختلفة من شأنها أن تدعم عملية اتخاذ القرارات المهمة في منظومة عمل المؤسسات المختلفة لما لها من أهمية بالغة في إعطاء تقييم حقيقي وواقعي حول تقييمات المفحوصين التي تستند إلى مبادئ القياس والتقويم الحديث؛ إذ إن التطورات في النظريات السيكومترية والاحصائية تُلبي احتياجات الباحثين والمدرسين والعلماء الذين يحتاجون إلى معلومات عن صنع القرار، عن طريق استخدام بعض أدوات القياس التي تعالج العديد من المشكلات المتعلقة بخصائصها (Smith & Kullikowich, 2004).

وللكشف عن إسهام كل مصدر من مصادر التباين المتعددة (Multiple Sources of Variance)؛ من الضروري حصر الأبعاد الأكثر تأثيرًا في ثبات الاختبار، حيث إن المهام والمصححين والفترات من الأبعاد التي لفتت انتباه الباحثين إلى أنها من أكثر مصادر التباين التي تؤثر في ثبات أداء المفحوصين (Shavelson, Baxter & Geo, 1993). وتبدأ نظرية التعميم بتحديد مختلف مصادر الخطأ التي من الضروري تكميم أهميتها النسبية كمساهمات في الخطأ، فالأهمية النسبية بواسطة تقديرات مكونات التباين تستخدم في تقييم خطأ القياس ومُعاملات التعميم. كما أن نظرية التعميم تعطي أهمية لمكونات تباين التأثيرات؛ لأن مقدارها يزود الباحثين بمعلومات حول مكونات مصادر الخطأ المؤثرة في القياسات المختلفة (Marcoulides & Kyriakides, 2010)، وتكشف عن التحسينات التي يُمكن إجراؤها بتغيير، أو تعديل إجراء القياس، مثل:

تعديل عدد الفقرات، أو عدد المصححين، أو عدد فترات التطبيق، أو استبعاد مستوى، أو عدة مستويات (Shavelson & Webb, 2009).

ونتيجة لذلك؛ تُعنى نظرية التعميم بنوعين من الدراسات: الأولى: دراسات التعميم (Generalizability Studies) والمعنية بجمع البيانات من نطاق الملاحظات المسموح بها، والمكونة من الأوجه التي يُعرفها الباحث، لتكون النتيجة الرئيسية مجموعة من مُكونات التباين، واهتمامه بإمكانية تعميم نتائج القياس على النطاق الشامل المراد قياسه (Allen & Yen, 1979). كما أنها تدرس فروق الخطأ من خلال تقدير مصادرها باستخدام تحليل التباين، الذي يُقدر تباين الخطأ(Error of Variance) للمفحوصين، والمهام، والمصححين، عبر الفترات الزمنية (Geo & Brennan, 2001)؛ حيث إن مُعامل التعميم يختلف باختلاف ظروف تطبيق الاختبار والعوامل المؤثرة في تباين الأداء على الاختبار. أما النوع الثاني: فهو دراسات القرار (Decision Studies) التي يُحاول الباحث فيها رسم أكثر من سيناريو؛ بهدف الحصول على أقصى ثبات وأقل خطأ بالاعتماد على دراسات التعميم (G-studies) (الحربي والحربي، 2017م)، كما يستخدم أدوات القياس من أجل صنع القرار لأغراض مختلفة، منها الفرز، أو القبول، أو غيرها.

وفي ضوء ما سبق، فإن أُولى الخطوات الضرورية التي يقوم بها الباحث في تصميم دراسة نظرية التعميم هي تحديد الشروط، أو ظروف جمع الملاحظات (Observations) (الأبعاد)؛ إذ إنه ليس هناك درجة حقيقة (True Score) واحدة للمفحوص في الأداة كما هو معروف في نظرية القياس التقليدية، وإنما تكون له درجة شاملة (Universe Score)، والتي يُقصد 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، يسمى كل واحد منّها بالوجه (Facet) أو البُعد. فالبُعد خاصية في القياس، مثل: المصححين، أو الأهداف، ويُمثل كل بُعد من هذه الأبعاد مصدرًا مهمًا من مصادر التباين التي تُعد ضرورية في عملية القياس، ويتكون كل بُعد من أبعاد القياس من مستويات تشتمل على مختلف ظروف القياس، كما تُسمى الملاحظات التي نحصل عليها من مختلف تجمعات الشروط الممثلة بالنطاق الشامل للملاحظات المقبولة (Universe of Admissible Observations) (علام، 2000م).

وعليه؛ فإنه يتم تحديد الأوجه عشوائية، أو ثابتة. فالعشوائية هي التي تُعنى بتعميم النتائج بشكل أوسع

من عينة التقنين، أما الثابتة فتكون نتائجها محصورة على عينة التقنين فقط، وبعدها يُختار التصميم. فالتصميم المتقاطع كُليًا يُجيب المفحوص فيه عن جميع المهام، وتُصحح إجاباته من جميع المصححين، ويخضع للاختبار في جميع مرات التطبيق، أما التصميم المُتداخل فيجيب كل مفحوص عن مهام مختلفة، وتُصحح إجاباته من قِبلِ مُصححين مُختلفين (الحربي والحربي، 2017م).

وضمن هذا السياق، أشار شافلسون وويب (Shavelson & Webb, 1991)، إلى أن مُعامل الثبات وفق نظرية التعميم يمكن التعبير عنه بمُعامل التعميم الذي يُعبر عنه بنسبة تباين الدرجة الشاملة إلى تباين الدرجة الملاحظة. فمُعامل التعميم النسبي يُتيح للباحثين تحديد المركز النسبي للمفحوص ورتبته بين أقرانه، بواسطة القرارات النسبية المتعلقة بمقارنة الفروق بين أداء المفحوصين، ويُستخدم في الاختبارات معيارية المرجع (Norm-Reference Tests). وفي المُقابل، فإن القرارات المطلقة المتعلقة بمقارنة أداء المفحوص بمحك خارجي، يُستخدم لها مُعامل التعميم المطلق المُستخدم في تقييم قدرة أداة القياس على مقارنة أداء المفحوصين على مستوى أداء مطلق (Brennan & Kane, 1977)، إضافةً لتفسيره البيانات المستمدة من اختبارات محكية المرجع (Criterion- Reference Tests).

وقد أُجريت العديد من الدراسات التي تُعنى بالكشف عن تغير معاينات تقييمات الأداء في ضوء نظرية التعميم، فقد أجرى شافلسون وآخرون (Shavelson et al., 1993) دراسة حول التغيرات في معاينات تقييمات الأداء في الرياضيات، شارك في الدراسة (105) مفحوص من الصف السادس بالإجابة عن ثلاث مهام، وقُدرت المهام الثلاثة من قبل مصححين اثنين، واستخدم التصميم (مفحوص×مصحح×مهمة). أظهرت النتائج أن أكبر مكون تباين خطأ قياس راجع إلى تفاعل (مفحوص-مهمة)، في حين نحتاج إلى (15) مهمة للحصول على مُعامل تعميم مقبول في دراسات القرار.

كما اهتُم في الدراسة نفسها التي أجراها شافلسون وآخرون (Shavelson et al., 1993) حول التغيرات في معاينات تقييمات الأداء في العلوم، حيث طُبّقت ثلاث مهام مستقلة على (186) مفحوصًا، وقيمهم مصححان اثنان، واستخدم التصميم (مفحوص ×مصحح ×مهمة ×فترة) بعينة حجمها (26) مفحوصًا، والتصميم (مفحوص ×مصحح ×مهمة) بعينة حجمها (50) مفحوصًا، أما التصميم (مفحوص× فترة) فاستخدم لجميع المفحوصين.

أظهرت النتائج أن أكبر مصدر لتباين الخطأ ناتج من تفاعل (مفحوص-مهمة-فترة)، بينما في التصميم الثاني كان أكبر مصدر تباين خطأ ناتج من تفاعل (مفحوص-مهمة)، ونحتاج إلى (23) مهمة للحصول على مُعامل تعميم مقبول. وتوصلت دراسة التعميم (مفحوص-مهمة-طريقة) إلى أن أكبر مكون تباين خطأ ناتج من تفاعل (مفحوص-مهمة-طريقة)؛ مما يُشير إلى أن الطرق غير متقاربة.

وفي ذات السياق، أجرى ريوز-بريمو وشافلسون (Ruiz-Primo & Shavelson, 1996) دراسة في العلوم، كان هدفها تعميم إجراءات تقييم الأداء عبر الفترات، اشتملت عينة الدراسة على (29) مفحوصًا، وطُلب من المفحوصين إنجاز ثلاث مهام، واستخدم الباحثان (8) مصححين في الفترة الزمنية الأولى، و(4) مصححين في الفترة الزمنية الثانية لتقييم الأداء، واستُخدم تصميم (مفحوص×فترة) لكل معالجة. توصلت النتائج إلى أن المعالجات الثلاث تتغير من فترة إلى أخرى، وكانت مُعاملات التعميم للقرارات المطلقة تتأثر بمصدر تباين الخطأ الناتج من تفاعل (مفحوص-فترة)، وارتفعت مُعاملات التعميم عندما جُمعت الدرجة الكلية. وأظهرت النتائج أن المفحوصين استخدموا إجراءات مُختلفة في كلا الفترتين؛ لأن مُعاملات الاتفاق جاءت منخفضة إلى متوسطة، كذلك أظهرت أن زيادة عدد مستويات الأبعاد يرفع من مُعاملات التعميم.

وفي دراسة مكبي وبارنيس (Mcbee & Barnes, 1998) هدفت إلى دراسة الاستقرار لأداء تحصيل المفحوصين في الرياضيات عبر الزمن والاتساق بين المهام. اختيرت (4) مهام، وشارك فيها (101) مفحوص من طلاب السنة الثامنة، وقدر أداؤهم بواسطة مصححين بالاعتماد على ميزان تقدير، وتم اعتماد التصميم المتقاطع كُليًا (مفحوص×مهمة×مصحح×فترة)، وحُللت البيانات بواسطة برمجية (GENOVA). وأن أكبر مصدر لتباين الخطأ يعود للمهمة وتفاعلاتها مع الأبعاد الأخرى. كما أظهرت النتائج أن زيادة عدد المهام المُتماثلة يرفع من مُعاملات التعميم.

كذلك قام كل من ويب وشليكمان وشوجر (Webb, Schlackman & Sugrue, 2000) بدراسة الاعتمادية لطرق التقييم وتعميم درجات تقييم العلوم، واشتملت عينة الدراسة على (57) مفحوصًا، وتكونت أداة الدراسة من اختبارين، في كل منهما مهمتان، وطُبق الاختبار الأول على نصف المفحوصين. وفي الوقت نفسه، طُبق الاختبار الثاني على النصف الآخر، وبعد شهر أُعيد التطبيق للاختبارين، واستخدمت التصاميم (مفحوص

×مهمة × مصحح)، و(مفحوص× مهمة× مصحح × فترة)، وقيم الأداء مصححان اثنان. أظهرت النتائج عدم وجود تباين دالٌ إحصائيـًا للفترة، والمهمة، والمُصحح، وأن أكبر مكون للتباين ناتج من تفاعل (مفحوص–مهمة)، وبيَّنت دراسات القرار أن عدم إدراج بُعد الفترة، وإدراج (3) مهام يجعل مُعامل التعميم مرتفعًا.

وأجرى جيو وبرينان (Geo & Brennan, 2001) دراسةً هدفت إلى معرفة تغير مكونات التباين والإحصاءات المرتبطة بها في تقييم الأداء. طُبقت (3) صيغ من المهام في الفترة الأولى، اشتملت كل منها على (12) مهمة، وطُبقت على عينة مكونة من (50) مفحوصًا. وفي الفترة الثانية طُبقت صيغتان من المهام، اشتملت كل منهما على (6) مهام، طُبقت على عينة مكونة من (157) مفحوصًا. وفي الفترة الثالثة لدراسة القرار طُبقت صيغة واحدة من المهام، طُبقت في الفترة الأولى على عينة مكونة من (121) مفحوصًا في الإصغاء، و(130) مفحوصًا في الكتابة، وقيَّم الأداء (3) مصححين للفترتين الأولى والثانية، ومصححان اثنان للفترة الثالثة، واستخدم التصميم (مفحوص× مهمة× مصحح). توصلت النتائج إلى أن أكبر مكون لتباين الخطأ كان ناتج من تفاعل (مفحوص–مهمة) في الفترتين الأولى والثانية، في حين توصلت دراسات القرار إلى أن مُعاملات التعميم في الإصغاء متدنية، وفي الكتابة مرتفعة.

وقام الباحثان سميث وكوليكويتش (Smith & kulikowich, 2004) بدراسة مرتبطة بتطبيق نظرية التعميم ونموذج راش متعدد الأبعاد؛ بهدف تقدير صدق الأداء وثباته بواسطة تقييم مهارات حل المشكلات. قُدمت (5) مهام لحل مشكلات مُركبة على شكل اختبار لعينة تكونت من (44) مفحوصًا، طُبق الاختبار على فترتين زمنيتين بينهما شهر، وصُححت الإجابات بواسطة مصححين اثنين، واستخدم التصميم (مفحوص× مهمة× مصحح× فترة). أظهرت النتائج أن مُعامل التعميم النسبي كان مقبولًا، ومُعامل التعميم المطلق كان متدنيًا، وكان أكبر مكون لتباين الخطأ ناتج عن مكون المهمة، في حين أظهرت دراسات القرار أن التقليل من عدد مستويات الأبعاد لا يُعطي مُعاملات تعميم مقبولة.

وفي دراسة أجراها كل من تشين ونيمي ووانغ وميروكا (Chen, Niemi, Wang & Mirocha, 2007) حول فحص تعميم مهام تقييم الكتابة المباشرة، والتي هدفت إلى معرفة مستوى تعميم قياس القدرة على الكتابة وصدقه باستخدام مهام المقال. أُستخدم فيها (4) مهام، اشتملت فيها عينة الدراسة على (397) مفحوصًا، وطلب من كل مفحوص اختيار مهمتين

عشوائيـًا لإنجازها، ودُربَ (4) مصححين لتقييم الأداء، واعتُمدت فترتان زمنيتان لإنجاز المهام، واُستخدم التصميم (مفحوص ×مقال ×مصحح). أشارت النتائج إلى أن أكبر مكون لتباين الخطأ ناتج من تفاعل (مفحوص-مقال-مصحح)، أمّا دراسات القرار فتوصلت إلى أن زيادة المقالات يرفع من مُعاملات التعميم.

وفي السياق ذاته، أجرى جبريل (Gebril, 2009) دراسة حول تعميم درجة مهام الكتابة الأكاديمية، من خلال معرفة تأثير مختلف الأبعاد في درجة الكتابة، وفحص مدى تشابه المهام المدمجة مع المهام المستقلة باستخدام نظرية التعميم. اشتملت عينة الدراسة على (115) مفحوصًا من مستوى السنة الرابعة بتخصص اللغة الإنجليزية في جامعة سوهاج في مصر، اُستخدمت (4) مهام، منها (2) مستقلة، و(2) مُدمجة، وقُدمت المهام للمفحوصين خلال يومين متتاليين، بعدها صححها (3) مصححين بالاعتماد على شبكة تصحيح، واُستخدم التصميم (مفحوص× مهمة× مصحح)، وحُللت البيانات باستخدام برمجية (GENOVA). أشارت النتائج إلى أن أكبر مكونات التباين ناتج من تفاعل (مفحوص-مهمة-مصحح) الممزوج بالخطأ المتبقي، وبيّنت دراسات القرار أن زيادة عدد المهام ترفع من مُعاملات التعميم أكثر من زيادة عدد المصححين.

وأجرى هيونج (Huang, 2009) دراسةً هدفت إلى تحليل ما وراء التحليل (Meta-Analysis) لمُعظم الدراسات التي تناولت تقييمات الأداء ضمن نظرية التعميم حول مقدار تغير معاينة المهمة في تقييم الأداء. اشتملت الدراسة على (50) دراسة منشورة ما بين (1980) و (2006)، واهتمت بجمع البيانات في عدة أبعاد (طريقة التصحيح، وطريقة التقييم، ومجال الموضوع، وتصميم الدراسة، ونوع المقال، وعدد الأبعاد، ومكونات تفاعل (المفحوص – المهمة). أشارت النتائج إلى أنَّ أثر تصميم الدراسة في مُعايرة مكون تباين (المهمة)، وتفاعل (مفحوص-مهمة) دالٌ إحصائيًا، كما أن تباين تفاعل (مفحوص-مهمة) انخفض عند إدماج الفترة كبُعد، وكان متوسط حجم أثر تباين تفاعل (مفحوص-مهمة) في تصميم (مفحوص× مهمة× مصحح× فترة) مُتدنيـًا.

كذلك أجرى جيولير وجيلبال (Guler & Gelbal, 2010) دراسة حول ثبات المهام المفتوحة في الرياضيات باستخدام النظرية التقليدية في القياس ونظرية التعميم، استخدم الباحثان (24) مهمة ذات أسئلة مفتوحة معتمدة في اختبارات الدراسات الدولية في الرياضيات والعلوم (Timss-1999)، وقد طُبقت على عينة اشتملت على (203) مفحوص من

مستوى الثامنة والتاسعة في أنقرة، وصُححت المهام بواسطة شبكة تصحيح بالاعتماد على (4) مصححين. أظهرت النتائج أن الاتساق الداخلي للدرجات مرتفع بالنسبة إلى كل مصحح، ومُعامل الاتساق لتقديرات المصححين دالٌ إحصائيًا، وكانت مُعاملات الارتباط بين كل زوجين من المصححين دالة إحصائيًا، كما أظهرت المقارنات البعدية عدم وجود فرق دالٌ إحصائيـًا بين متوسط درجات كل زوجين من المصححين باستثناء المصحح الأول، أما مُعاملات التعميم فكانت مرتفعة، ومكون التباين الأكبر ناتج من تفاعل (مفحوص–مهمة). وأشارت النتائج إلى أن زيادة عدد المصححين لا يُحسن من مُعاملات التعميم، في حين زيادة عدد المهام يزيد من تأثيرها.

وفي دراسة أجراها لي (Lee, 2016) لمقارنة بين استخدام نظرية التعميم ونموذج راش متعدد الأوجه في تحليل اختبار حل المشكلات الإبداعي في الرياضيات. اشتملت عينة الدراسة على (172) مفحوصًا من الصف العاشر، وتكوّن الاختبار من (5) مهام مفتوحة في الرياضيات، وصحح المهام (4) مصححين باستخدام ميزان تصحيح، واستخدم التصميم (مفحوص×مهمة×مصحح×معيار). أظهرت النتائج أن الطريقتين تتفقان نسبيـًا في مصادر التباين، فمُكون تباين تفاعل (مفحوص×مهمة) كان كبيرًا، ومُكون تباين المصحح والتفاعلات المرتبطة معه كان منخفضًا، وأظهرت دراسات القرار أنه بزيادة عدد المهام إلى (10) يتم الحصول على مُعاملات ثبات مرتفعة.

وقام طباع (2020م) بدراسة هدفت إلى تطبيق نظرية التعميم لتقدير ثبات اختبار تقييم كفاءة الرياضيات لدى طلاب السنة الرابعة ابتدائي. طُبق فيها اختبار اشتمل على (9) مهمات مُركبة، وطُبقت المهمات على (331) مفحوصًا، واعتمد على ثلاثة مصححين لتقييم الأداء، وحُللت البيانات باستخدام برمجية (EduG)، واستخدم التصميم المُتقاطع كُليًا (مفحوص×مهمة×مصحح). أظهرت النتائج أن مصادر التباين الأكثر تأثيرًا في الثبات ناتجة من تفاعل (مفحوص–مهمة)، وتباين المهمة، أما دراسات القرار فَأظهرت أن زيادة عدد المهام أفضل من زيادة عدد المصححين في رفع مُعاملات التعميم.

وحسب علم الباحثين، هناك نُدرة في الدراسات العربية التي اهتمت بتطبيق نظرية التعميم، والتي تتناول تصميمات تجريبية اعتمدت على المفحوصين، والمصححين، والمهام، والفترات الزمنية كأبعاد أساسية تعتمد على تصميمات ثلاثية البُعد في تقييم الأداء. إضافة إلى أن جميع الدراسات السابقة ركزت على دراسة اختبارات تقييم الأداء التي تتكون من

مهام، وكان عدد المهام المُستخدمة فيها ما بين مهمتين و(24) مهمة، في حين عدد المهام في الدراسة الحالية (12) مهمة كدراسة (Geo & Brennan, 2001)، وتميزت بتناولها مهام مُركبة (Complex Tasks) تنوعت لـ مهام في التطبيق، والاستدلال، والانتقاء، والرأي، والتي اعتُمد في بنائها على فئات مهام دويل (Doyle, 1983)، والدراسات السابقة اعتمدت في تقييم الأداء على عدد من المُصححين تراوح ما بين (2-5)، في حين اعتُمد على (3) مُصححين في الدراسة الحالية. أما بالنسبة لإدراج بُعد الفترة فمُعظم الدراسات أظهرت أن له تأثيرًا في معاملات التعميم باستثناء الدراسات (Webb et al., 2000; Smith & Kulikowich, 2004)، والدراسات (Mcbee & Barnes, 1998; Gebril, 2009)؛ فإنها استخدمت برمجية (GENOVA). في المُقابل، هذه الدراسة استخدمت في تحليلاتها برمجية (EduG) كدراسة طباع (2020)، وتباينت أحجام العينات في الدراسات السابقة؛ إذ انحصر عدد أفراد العينات من (29) مفحوصًا في دراسة (Ruiz-Primo & Shavelson, 1996) إلى (397) مفحوصًا كما في دراسة (Chen et al., 2007). أما الدراسة الحالية فبلغ عدد أفراد عينتها (243) مفحوصًا، كما استُخدم التصميم المتقاطع في الدراسة الحالية كمُعظم الدراسات السابقة عِوضـًا عن التصاميم المتداخلة؛ باعتبارها أكثر كفاءةً حسب تعبير هيونج (Hung, 2009). وأثبتت العديد من الدراسات (Ruiz-Primo & Shavelson, 1996; Webb et al. 2000; Geo & Brennan, 2001; Smith & Kulikowich, 2004; Lee, 2016) أن معظم تقييمات الأداء كان ثباتها متدنيًا في معظم التخصصات، مثل: الرياضيات والعلوم وغيرها؛ من هُنا تشكلت لدى الباحثين فكرة إعادة النظر في دراسة مصادر التباين التي تؤثر في ثبات اختبار في الرياضيات.

مُشكلة الدراسة وأسئلتها:

تهدف الدراسة الحالية إلى الكشف عن إسهام مصادر تباين الخطأ المتعددة في ثبات اختبار في الرياضيات باستخدام نظرية التعميم؛ حيث إن نظرية القياس التقليدية تعنى بمصدر خطأ واحد في القياس، ولا تُميز بين مصادر تباين الخطأ التي تدرسها من أجل تخفيض خطأ القياس، ففي تقييمات الأداء تُعد مصادر تباين الخطأ مؤثرة بشكل مباشر في اتساق النتائج الناجمة عن عينة من أداء المفحوص المُنتقى من نطاق مُعقد يتحدد بترابط أبعاد متعددة في الوقت نفسه، كالمهام، والمصححين، والفترات الزمنية. وعليه؛ فإنها من أكثر مصادر عدم ثبات تقييمات أداء المفحوصين، فإعطاء المفحوصين مهام متعددة تختلف من حيث الصعوبة تبعًا لقدراتهم؛ بهدف دراسة التباين في أدائهم أمر بالغ الأهمية، ودراسة التباين بين تقديرات

المُصححين التي تكون غالبًا غير مُتطابقة لأداء المفحوصين في مهام أشبه ما تكون بالأسئلة الإنشائية التي تُراعي المعرفة الجزئية للمفحوصين، يستدعي الأهمية لدراستها، وأنَّ تغير أداء المفحوص من فترة إلى أخرى، يُسلط الضوء على الاهتمام بدراستها أيضًا (Shavelson et al., 1993)، ولا يوجد نظرية تسمح بتقدير تباين الأفراد والمصادر المتعددة لتباين الخطأ في الوقت نفسه إلا نظرية التعميم؛ فهي تُعد من أفضل الطرق الإحصائية المستخدمة في تقدير جودة تقييمات الأداء للمفحوصين، ومن هُنا تشكلت إعادة النظر في دراسة ضعف ثبات أداء المفحوصين، الناتج من عدم اتساق المهام أو المصححين عبر الفترات الزمنية. وفي هذا الإطار تُحاول الدراسة الحالية الإجابة عن السؤالين الآتيين:

1-ما نسبة إسهام تباين الخطأ الذي يُفسره كل من أبعاد (المهام، المصححين، الفترات الزمنية) في التباين الكلي في ثبات اختبار في الرياضيات؟

2-ما شروط تحقيق أفضل مستويات ثبات اختبار في الرياضيات ضمن أبعاد (المهام، المصححين، الفترات الزمنية)؟

أهداف الدراسة:

الدراسة الحالية تُحقق هدفين رئيسين، هما الكشف عن:

1-حجم تباين الخطأ الذي يُفسره كُل من الأبعاد (المهام، والمصححين، والفترات الزمنية) في التباين الكُلي في ثبات اختبار في الرياضيات.

2-الشُروط التي يُمكن بها الحصول على أفضل مستويات ثبات اختبار في الرياضيات.

أهمية الدراسة:

تظهر الأهمية النظرية للدراسة الحالية في موضوعها الذي يتناول مؤشرات الخصائص السيكومترية لتقييمات الأداء للمفحوصين في اختبار رياضيات مُعد وفق المهام المُركبة باستخدام نظرية التعميم، وهذا الاختبار يكشف عن مهارات التفكير العليا لدى المفحوصين من خلال مجموعة من المهام التي تعمل على تنمية طرق التفكير لدى المفحوصين، وحل المُشكلات المعقدة في الرياضيات؛ حيث أسهمت هذه الدراسة في الكشف عن حجم تباين الخطأ الذي تُفسره أبعاد: المهام، والمصححين، والفترات الزمنية في التباين الكُلي في ثبات اختبار في الرياضيات، وكذلك الكشف عن الشُروط التي يُمكن من خلالها الحصول على أفضل مستويات ثبات اختبار في الرياضيات.

ومن حيث الأهمية التطبيقية؛ يُتوقع أن تعمل هذه الدراسة على توفير إطار نظري حول نظرية التعميم وأهميتها في تحقيق التقييم الحقيقي للأداء؛ إذْ إنَّ هناك

نـدرة في الـدراســـات العربيـة التي تطرقـت لهـذا الأســلوب، كما تُفيد نتائج هذه الدراســة العاملين في ميدان التدريس، ومركز الاختبارات في وزارة التربية والتعليم، والبـاحثين في ميـدان القيـاس والتقويم، بمعرفتهم لمصـــادر تباين الخطأ في أداء المفحوصـــين؛ بهدف تحسين الخصـائص السيكومترية للاختبارات، وربما تمثل إضـــافة جديدة لمنهجية يمكن توظيفها في بنـاء الاختبـارات المقننـة في مراكز القيـاس والتقويم المتخصصة.

محددات الدراسة:

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

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

3-اقتصر محتوى الاختبار على مجال (الأعداد والعمليات عليها) من مجالات كتاب الرياضيات المُقرر تدريسه من وزارة التربية والتعليم الأُردنية للصف الخامس الأساسي.

4-اقتصرت الدراسة على إدراج الأبعاد (المهام، المُصححين، الفترات الزمنية)، ولم تُدرج أبعاد أُخرى، كصيغ المهام وطرق التصحيح أو غيرها.

مصطلحات الدراسة:

1-**نظرية التعميم**: مجموعة الطرق الإحصائية الأكثر مرونة مع مختلف أبعاد تقدير ثبات القياسات السلوكية، وتعتمد في تقديراتها على مؤشرات دقيقة تعالج مختلف الأبعاد والتفاعلات الناتجة عنها، وتستخدم تصميمات متقاطعة لمختلف حالات القياس في دراسات التعميم، وتسمح بتقديم إجراءات تحسين القياسات في دراسات القرار.

وتعرف **إجرائيًا**: بأنها إطار إحصائي متعدد الأبعاد يُعبر عن تقديرات تباين الخطأ المُفسر، ومعاملات التعميم النسبية والمطلقة (الموضحة في معادلة (1)، ومعادلة (2)).

2-**الثبات**: تعميم الدرجة الملاحظة في اختبار الرياضيات من خلال أداء المفحوص على المهام المُركبة، مقارنة بالدرجة الشاملة التي حصل عليها ضمن الأبعاد (مهام، مصححين، فترات زمنية).

ويعرف **إجرائيًا**: بأنه معامل التعميم النسبي، ومعامل التعميم المطلق، الناتج من درجات المفحوصين ($P_1 - P_{243}$).

3-**تقييم الأداء**: درجة أداء المفحوصين (P_1 – P_{243}) على اختبار في الرياضيات الذي يشمل (12) مهمة موزعة بالتساوي على صيغ المهام (التطبيق، الاستدلال، الانتقاء، الرأي).

ويعرف **إجرائيًا**: بأنه درجة المفحوص على مهمة، أو صيغة، أو على الاختبار الذي يتكون من (12) مهمة في الرياضيات موزعة بالتساوي على صيغ (التطبيق، الاستدلال، الانتقاء، والرأي).

4-**مصادر التباين**: تقديرات تُعبر عن اختلاف الوسط الحسابي للدرجات التي يحصل عليها المفحوص في اختبار الرياضيات تحت مختلف أبعاد القياس (المهام($T_1 - T_{12}$)، والمصححين($J_1 - J_3$)، والفترات الزمنية($O_1 - O_2$)، والتفاعلات بينهما)، وتُقدر مصادر التباين عن طريق وسط المربعات وعدد المستويات.

وتعرف **إجرائيًا**: بأنه تباين الخطأ النسبي، وتباين الخطأ المطلق الناتج من درجات المفحوصين (P_1 – P_{243}) ضمن أبعاد الدراسة المُستخدمة.

5-**المهمة المُركبة**: مجموعة من المعلومات في مجال الأعداد والعمليات عليها في الرياضيات وضعت في سياق حياتي، وتتطلب من المفحوص استخدام جميع المعارف والمهارات والاتجاهات التي اكتسبها لحلها خلال فترة زمنية، وضمن معايير محددة.

وتعرف إجرائيًا: بأنها مهارة رياضية تتطلب استخدام جميع المعارف والمهارات والاتجاهات التي اكتسبها المفحوص لإيجاد حل خلال فترة زمنية، وضمن مستويات أداء محددة، يُستدل عليها بدرجات المفحوصين ($P_1 - P_{243}$) ضمن أبعاد الدراسة المُستخدمة.

منهج الدراسة وإجراءاتها:

منهج الدراسة:

استُخدم المنهج الوصفي، الذي يُعنى بدراسة الظاهرة ووصفها وصفًا دقيقًا، ويُعبر عنها كيفيـًا بوصف الظاهرة وخصائصها، أو كميـًا بتوضيح مقدار حجم الظاهرة، ودرجة ارتباطها مع الظواهر الأُخرى (عبيدات، عدس، وعبد الحق، 2005)، فالمنهج المُستخدم في الدراسة الحالية يهتم بالتعرف على مصادر تباين الخطأ المتعددة الأكثر تأثيرًا في ثبات أداء المفحوصين.

مجتمع الدراسة:

تكون مجتمع الدراسة من جميع طلاب وطالبات الصف الخامس الأساسي للعام الدراسي 2019/ 2020م، والبالغ عددهم (1385) طالباً وطالبة، وذلك حسب بيانات التقرير الإحصائي لأعداد الطلبة الصادر عن قسم التخطيط التربوي في

مديرية التربية والتعليم للواءي الطيبة والوسطية للعام الدراسي 2019/ 2020م.

عينة الدراسة:

طُبقت أداة الدراسة على عينة من طلاب الصف الخامس الأساسي في المدارس التابعة لمديرية التربية والتعليم للواءي الطيبة والوسطية/ محافظة اربد للعام الدراسي 2019/2020م، وتكونت من (243) طالبًا وطالبة بواقع (12) مدرسة حكومية، واختيرت بالطريقة العشوائية العنقودية؛ حيث وحدة الاختيار هي المدرسة، ومثل حجم العينة ما نسبته (18٪) من حجم مجتمع الدراسة، ويُعد مناسبًا لأغراض هذه الدراسة.

أداة الدراسة:

استخدم الباحثان اختبارًا في الرياضيات تكوّن من (12) مهمة، موزعة بالتساوي على (4) صيغ من المهام (التطبيق، الاستدلال، الانتقاء، الرأي) المتعلقة بحل المشكلات التي قد تواجه المفحوص في حياته اليومية. وللتحقق من صدق أداة الدراسة؛ تحقق الباحثان من الصدق الظاهري (صدق المحتوى) بشبكة تحكيمٍ تحتوي على معايير، عُرضت على (18) مُحكمًا من ذوي الاختصاص والخبرة (معلمي الرياضيات للصف الخامس الأساسي، ومشرفي الرياضيات، وأساتذة جامعيين، ومتخصصي القياس والتقويم)؛ حيث تم التحقق من تمثيل مهام الاختبار للمجال الدراسي المطلوب، وتم الأخذ بمُلاحظاتهم حول المهام والعمل على تعديلها بالاعتماد على نسبة الاتفاق بين تقديرات المحكمين التي لا تقل عن (70٪) للمهمة، وحُسب الصدق التجريبي (الصدق المرتبط بمحك) بالاعتماد على بيانات المفحوصين في العينة الاستطلاعية التي كان حجمها (40) مفحوصًا؛ إذ كانت قيم مُعاملات الارتباط مرتفعة ولا تقل عن (0.7)؛ مما يُعد مؤشرًا كافيًا إلى درجة صدق الاختبار. كما قام الباحثان بتقدير مُعامل ثبات الاستقرار (Stability coefficient) للاختبار أو ما يُسمى بمعامل ثبات الإعادة (Test R-Test)؛ حيث بلغت قيمته (0.89)، وقُدّر ثبات الاتساق الداخلي للاختبار باستخدام معادلة كرونباخ ألفا، وبلغت قيمته (0.91)؛ مما يدل على أن الاختبار يتمتع بثبات مقبول.

إجراءات الدراسة:

1- اختيرت المدارس التي جرى على طُلابها تطبيق الاختبار مرتين بفاصل زمني مدته أسبوعان بالطريقة العشوائية العنقودية، باعتبار أن وحدة الاختيار هي المجموعة.

2- حُسب الوسط الحسابي لزمن الاختبار على العينة الاستطلاعية، وتبين أنهم يحتاجون إلى (3)

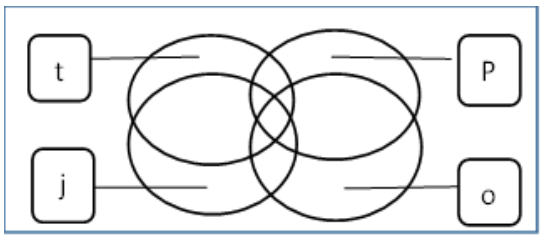
جلسات منفصلة بزمن (45) دقيقة لكل جلسة؛ حيث يعرض في كل جلسة على المفحوص (4) مهام مختلفة؛ وذلك بهدف التنويع في الإستراتيجيات التي يستخدمها المفحوص.

3- طُبق الاختبار بمساعدة المعلمين بعد تقديم تعليمات الاختبار لهم بشكل موحد وموضوعي؛ بهدف الحصول على نفس ظروف مرتي تطبيق الاختبار.

4- دُرّب (3) مصححين على ميزان التصحيح التحليلي الذي أُعدَّ للاختبار، حيث حُددت مجموعة من مستويات الأداء (Level of Performance) وعددها (4) لتقييم أداء المفحوص على المهام المُركبة، واشتمل كل مستوى على مجموعة من المحكات (Criteria)؛ بهدف تقدير أداء المفحوص على كل مستوى أداء من خلال إتقانه لهذه المحكات، واعتُمد أداء المفحوص الذي يحقق النسبة (50٪) من إتقان المحكات ضمن المستوى يُعطى علامة مستوى الأداء؛ فالمفحوص الذي يتقن مستوى الأداء المطلوب يعطى العلامة (1)، وغير المتقن لمستوى الأداء يعطى العلامة (0)، ومن ثم تجمع علامات جميع مستويات الأداء لتشكل تقديرًا لأداء المفحوص على المهمة، ومن ثم فإن القيمة العُظمى لأداء أي مفحوص على أي مهمة (4) علامات، والقيمة الصُغرى لأداء أي مفحوص على أي مهمة (صفر).

تصميم الدراسة (p×t×j×o):

يُلاحظ من شكل (1) الذي يُبين رسمًا توضيحيًا للتصميم (مفحوص× مهمة× مصحح× فترة)، ولا يُعد المفحوصون مصدرًا من مصادر الأخطاء؛ لأنهم موضوع القياس، واتخاذ القرار متعلق بهم (Guler & Gelbal, 2010)، وأن كل رمز يُعبر عن مصدر تباين، بحيث (p) المفحوصون، و(t) المهام، و(j) المصححون، و(o) الفترات الزمنية، و(pt) تفاعل (مفحوص-مهمة)، و(pj) تفاعل (مفحوص-مصحح)، و(po) تفاعل (مفحوص-فترة)، و(tj) تفاعل (مهمة-مصحح)، و(to) تفاعل (مهمة-فترة)، و(jo) تفاعل (مصحح-فترة)، و(ptj) تفاعل (مفحوص-مهمة-مصحح)، و(pto) تفاعل (مفحوص-مهمة-فترة)، و(tjo) تفاعل (مهمة-مصحح-فترة)، و(ptjo) تفاعل (مفحوص-مهمة-مصحح-فترة).



الشكل (1): رسم توضيحي لتصميم الدراسة (p×t×j×o)

ويمكن حساب مُعامل التعميم النسبي (ρ^2) بمعادلة (1)، وحساب مُعامل التعميم المطلق (ϕ) بمعادلة (2) للتصميم السابق؛ حيث إن كل رمز يُعبر عن تباين: (σ^2_p) الدرجة الشاملة، و(σ^2_t) المهام، و(σ^2_j) المصححون، و(σ^2_o) الفترات الزمنية، و(σ^2_{pt}) تفاعل (مفحوص–مهمة)، و(σ^2_{pj}) تفاعل (مفحوص–مصحح)، و(σ^2_{po}) تفاعل (مفحوص–فترة)، و(σ^2_{tj}) تفاعل (مهمة–مصحح)، و(σ^2_{to}) تفاعل (مهمة–فترة)، و(σ^2_{jo}) تفاعل (مصحح–فترة)، و(σ^2_{ptj}) تفاعل (مفحوص–مهمة–مصحح)، و(σ^2_{pto}) تفاعل (مفحوص–مهمة–فترة)، و(σ^2_{pjo}) تفاعل (مفحوص–مصحح–فترة)، و(σ^2_{tjo}) تفاعل (مهمة–مصحح–فترة)، و(σ^2_{ptjo}) تفاعل (مفحوص–مهمة–مصحح–فترة)، و(n'_t) تُعبر عن عدد المهام، و(n'_j) تُعبر عن عدد المصححين، و(n'_o) تُعبر عن الفترات الزمنية (Brennan, 2001).

$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pj}}{n'_j} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{ptj}}{n'_t n'_j} + \frac{\sigma^2_{pjo}}{n'_o n'_j} + \frac{\sigma^2_{pto}}{n'_t n'_o} + \frac{\sigma^2_{ptjo}}{n'_t n'_j n'_o}} \quad \ldots\ldots (1$$

$$\phi = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_t + \sigma^2_{pt}}{n'_t} + \frac{\sigma^2_j + \sigma^2_{pj}}{n'_j} + \frac{\sigma^2_o + \sigma^2_{po}}{n'_o} + \frac{\sigma^2_{tj} + \sigma^2_{to+}\sigma^2_{jo+}\sigma^2_{ptj}}{n'_t n'_j} + \frac{\sigma^2_{pto} + \sigma^2_{pjo} + \sigma^2_{joo}}{n'_t n'_o} + \frac{\sigma^2_{ptjo}}{n'_t n'_j n'_o}} \quad \ldots (2)$$

أساليب تحليل البيانات:

استُخدم برنامج (EXCEL) في حساب مُعاملات الاتفاق للمحكمين، والحصول على الرسوم البيانية لدراسات القرار، واستخدمت برمجية (SPSS.21) لحساب مُعاملات الارتباط الخاصة بالخصائص السيكومترية لأداة الدراسة، واستخدمت برمجية (EduG) لتحليل البيانات؛ لسهولة استخدامها (SSREWG, 2010) للإجابة عن أسئلة الدراسة لحساب (الإحصاءات الوصفية، وتحليل التباين، وتحليل التعميم والقرار لتصميم الدراسة).

النتائج:

الإحصاءات الوصفية لفترتي تطبيق الاختبار:

اعتمد الباحثان على فترتين لتطبيق الاختبار، وكانت الأوساط الحسابية والانحرافات المعيارية لأداء المفحوصين خلال فترتي تطبيق الاختبار كما يأتي:

جدول (1): الإحصاءات الوصفية لفترتي تطبيق الاختبار

الفترة الزمنية	الأولى	الثانية
الوسط الحسابي	1.301	1.447
الانحراف المعياري	1.453	1.529

يُلاحظ من النتائج في جدول (1) أن الأوساط الحسابية لفترتي تطبيق الاختبار متقاربة بفارق لا يتجاوز (0.15)، وهذا يدل على الاتساق في نتائج الاختبار عبر الفترتين الزمنيتين، أما الانحرافات المعيارية بين فترتي تطبيق الاختبار فكان مقدارها على

الترتيب تبعـًا لفترة التطبيق (1.453)، (1.529)؛ مما يُشير إلى عدم وجود تباين كبير، وهذا يدل على استقرار أداء المفحوصين عبر الفترات.

نتائج السؤال الأول:

للإجابة عن السؤال الأول؛ تم الاعتماد على دراسة تحليل التباين، وتحليل التعميم للتصميم (مفحوص× مهمة× مصحح× فترة) للمهام الكلية للاختبار، وكانت النتائج ما يأتي:

جدول (2): نتائج تحليل التباين للتصميم (مفحوص×مهمة×مصحح×فترة) للمهام الكلية للاختبار

مصدر التباين	مجموع المربعات	درجات الحرية	وسط المربعات	المكونات				
				العشوائية	المختلطة	المصححة	٪	الخطأ
مفحوص (P)	9987.087	242	41.269	0.495	0.495	0.495	%24.9	0.052
مصحح (J)	401.321	2	200.66	0.031	0.031	0.031	%1.6	0.024
مهمة(T)	4265.359	11	387.76	0.251	0.251	0.251	%12.6	0.104
فترة(O)	43.953	1	43.953	0.003	0.003	0.003	%0.1	0.004
مفحوص-مصحح	353.654	484	0.731	0.009	0.009	0.009	%0.4	0.002
مفحوص-مهمة	10106.5	2662	3.797	0.356	0.356	0.356	%17.9	0.019
مفحوص-فترة	774.942	242	3.202	0.04	0.04	0.04	%2	0.008
مصحح-مهمة	264.053	22	12.002	0.023	0.023	0.023	%1.1	0.007
مصحح-فترة	16.702	2	8.351	0.003	0.003	0.003	%0.1	0.002
مهمة-فترة	102.945	11	9.359	0.01	0.01	0.01	%0.5	0.005
مفحوص-مهمة-مصحح	1827.764	5324	0.343	0.032	0.032	0.032	%1.6	0.004
مفحوص-مصحح-فترة	220.05	484	0.455	0.015	0.015	0.015	%0.7	0.003
مفحوص-مهمة-فترة	4256.087	2662	1.599	0.44	0.44	0.44	%22.1	0.015
مصحح-مهمة-فترة	21.558	22	0.98	0.003	0.003	0.003	%0.1	0.001
مفحوص-مهمة-مصحح-فترة	1487.812	5324	0.28	0.28	0.28	0.28	%14.1	0.005
المجموع	34129.79	17495					%100	

تُشير النتائج في جدول (2) إلى أن أكبر مكون كان مصدره المفحوص ومقداره (24.9٪)، وأقل مكون كان مصدره كلًا من الفترة وتفاعل (مصحح-فترة)، وتفاعل (مهمة-مصحح-فترة) ومقدار كل منها

يساوي (0.1٪). أما أكبر خطأ معياري فناتج من (0.001). وكانت نتائج تحليل التعميم في جدول المفحوص ويساوي (0.052)، وأقل خطأ معياري (3) ما يأتي:
ناتج من تفاعل (مهمة-مصحح-فترة) ويساوي

جدول (3): نتائج تحليل التعميم للتصميم (مفحوص×مهمة×مصحح×فترة) للمهام الكلية للاختبار.

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية	تباين الخطأ المطلق	النسبة المئوية
مفحوص (P)	0.495	.....	.....	.....	.....
مصحح (J)	.....	.....	.....	0.01	٪9.2
مهمة(T)	.....	.....	.....	0.021	٪18.6
فترة(O)	.....	.....	.....	0.002	٪1.3
مفحوص-مصحح	.....	0.003	٪3.8	0.003	٪2.6
مفحوص-مهمة	.....	0.03	٪38	0.03	٪26.4
مفحوص-فترة	.....	0.02	٪25.4	0.02	٪17.7
مصحح-مهمة	.....	.....	.....	0.001	٪0.6
مصحح-فترة	.....	.....	.....	0	٪0.4
مهمة-فترة	.....	.....	.....	0	٪0.4
مفحوص-مهمة-مصحح	.....	0.001	٪1.1	0.001	٪0.8
مفحوص-مصحح-فترة	.....	0.002	٪3.1	0.002	٪2.2
مفحوص-مهمة-فترة	.....	0.018	٪23.5	0.018	٪16.3
مصحح-مهمة-فترة	.....	.....	.....	0	٪0
مفحوص-مهمة-مصحح-فترة	.....	0.004	٪5	0.004	٪3.5
مجموع التباينات	0.495	0.078	100٪	0.112	٪100
الانحراف المعياري	.0704	الخطأ المعياري النسبي	0.279	الخطأ المعياري المطلق	0.335
معامل التعميم النسبي	0.86				
معامل التعميم المطلق	0.82				

*(.....) تشير إلى عدم وجود مصدر للتباين في الخلية.

تُبيّن النتائج في جدول (3) أن أكبر مكون لتباين (مفحوص-مهمة) ومقداره على التوالي (38٪)، الخطأ في القياس النسبي أو المطلق راجع إلى تفاعل (26.4٪)؛ مما يدل على أن أداء المفحوصين كان

متباينـًا تبعـًا لنوع المهمة المعطاة لهم، وكما نلاحظ أن ثاني أكبر مكون للتباين راجع إلى المهمة ومقداره (18.6٪)؛ مما يدل على أن هناك اختلافًا في مستوى صعوبة المهمة المعطاة للمفحوصين. ومن ثم يأتي مكون لتباين الخطأ في القياس النسبي والمطلق راجع إلى تفاعل (مفحوص-فترة) ومقداره على التوالي (25.4٪)، (17.7٪)؛ مما يدل على اختلاف أداء المفحوص باختلاف الفترات الزمنية، والذي يقل عنه مكون التباين الراجع إلى تفاعل (مفحوص-مهمة-فترة) ومقداره (23.5٪) في القياس النسبي، وفي المطلق مقداره (16.3٪). يليه مكون التباين للمصحح مقداره (9.2٪)، وباقي مكونات التباين لم تتجاوز (5٪). كذلك نلاحظ أن الخطأ المعياري النسبي ومقداره (0.279) كان أقل من الخطأ المعياري المطلق ومقداره (0.335)؛ مما يشير إلى أن مُعامل التعميم النسبي (0.86) أعلى من مُعامل التعميم المطلق (0.82).

نتائج السؤال الثاني:

للإجابة عن السؤال الثاني، طُرح سيناريوهان اثنان لدراسات القرار لزيادة مستويات الأبعاد؛ بهدف رفع تباين الدرجة الشاملة من أجل الحصول على مُعاملات تعميم مُرتفعة، ومعرفة أي الأبعاد ذات التأثير الأكبر في ثبات أداء المفحوصين؛ إذ كانت النتائج على النحو الآتي:

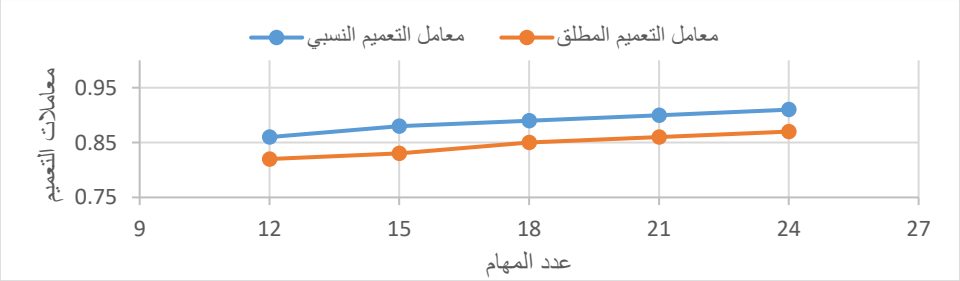
جدول (4): نتائج دراسات القرار للتصميم (مفحوص×مهمة×مصحح×فترة) بتثبيت عدد الفترات الزمنية.

الأبعاد				معامل التعميم النسبي	معامل التعميم المطلق
مفحوص	مهمة	مصحح	فترة		
243	12	3	2	0.86	0.82
243	12	4	2	0.87	0.82
243	12	5	2	0.87	0.83
243	15	3	2	0.88	0.84
243	15	4	2	0.88	0.84
243	15	5	2	0.89	0.85
243	18	3	2	0.89	0.85
243	18	4	2	0.89	0.86
243	18	5	2	0.90	0.86
243	21	3	2	0.90	0.86
243	21	4	2	0.90	0.87

تابع/ جدول (4):

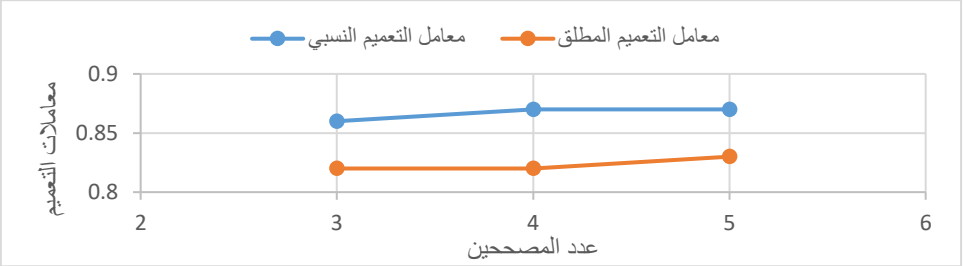
الأبعاد				معامل التعميم النسبي	معامل التعميم المطلق
مفحوص	مهمة	مصحح	فترة		
243	21	5	2	0.90	0.87
243	24	3	2	0.91	0.87
243	24	4	2	0.91	0.88
243	24	5	2	0.91	0.88

تُشير النتائج الواردة في جدول (4) إلى أنه تم التدرج في زيادة عدد المهام بدءًا من (12) إلى (24) بزيادة مقدارها (3) مهام في كل مرة، وتم عرض التدرج في زيادة عدد المصححين بدءًا من (3) إلى (5) بزيادة مصحح واحد في كل مرة. كما يُلاحظ الحالة (24) مهمة و(3) مصححين سيكون مُعامل التعميم النسبي والمطلق على التوالي (0.91)، (0.87)، حيث كان كل منها على التوالي (0.86)، (0.82)، في حين أن زيادة عدد المصححين من (3) إلى (5) لم يظهر فرقًا كبيرًا عند الحالة (12) مهمة؛ إذ ارتفع مُعامل التعميم النسبي والمطلق على التوالي من (0.86)، (0.82) إلى (0.87) إلى (0.83)



الشكل (2): تمثيل بياني لمُعاملات التعميم في حال زيادة عدد المهام(12-24)، المصححين(3)، الفترات الزمنية (2)

يُلاحظ من الشكل (2) أن زيادة عدد المهام يسهم -بشكل ملحوظ- في تحقيق مستويات ثبات أفضل، وقيم مُعاملات التعميم النسبية والمطلقة عندما يكون عدد المهام (12)، وعدد المصححين (3)، وعدد الفترات الزمنية (2)، فإنها تساوي على التوالي (0.86، 0.82)؛ مما يدل على القدرة على تعميم نتائج الاختبار بوضع الاختبار الحالي دون اللجوء إلى زيادة أي من مستويات الأبعاد؛ وهذا يؤكد الأهمية الجوهرية لإدراج بُعد الفترة الزمنية ضمن الدراسة. واستخدمت (12) مهمة فقط في الاختبار لاعتبارات من أجل تقليل الجهد والتكلفة والوقت، فهي كافية لتحقيق مُعاملات تعميم ذات قيم مقبولة.



الشكل (3): تمثيل بياني لمُعاملات التعميم في حال زيادة عدد المصححين(3-5)، المهام(12)، الفترات الزمنية (2)

يُلاحظ من الشكل (3) أن زيادة عدد المصححين لا يسهم -بشكل ملحوظ- في تحقيق مستويات ثبات مقبولة، وبالاعتماد على مُعاملات التعميم النسبية والمطلقة في حال كان عدد كل من المصححين (5)، والمهام (12)، فإنها تساوي على التوالي (0.87، 0.83)؛ مما يدل على قدرة تعميم نتائج الاختبار دون اللجوء إلى زيادة عدد المصححين، ومما يدل على مناسبة اختيار الباحثين (3) مصححين في الاختبار. ومستويات الثبات تكون متماثلة عنها عند اختيار (5) مصححين، وتصل مُعاملات التعميم النسبية والمطلقة لمستويات أعلى بما لا يتجاوز (0.01) من قيمة مُعاملات التعميم.

ومما تجدر الإشارة إليه، أنه في إجراءات التحسين لمُعاملات الثبات، سواءً كانت خاصة بزيادة عدد مهام الاختبار الكلية، أو عدد المصححين، أو كليهما من أجل الحصول على مُعاملات ثبات أكبر لا تعطي نفعـًا كبيرًا؛ بما أنه تحققت المستويات المطلوبة لتعميم النتائج، وهذا راجع إلى إدراج بُعد الفترة الزمنية. وفي جدول (5) كانت نتائج دراسات القرار للسيناريو الثاني ما يأتي:

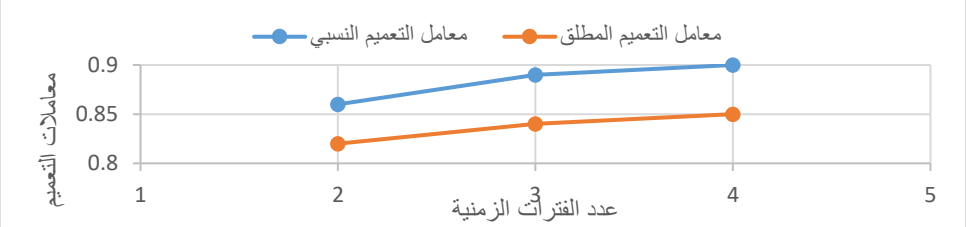
جدول (5): نتائج دراسات القرار للتصميم (مفحوص×مهمة×مصحح×فترة) بتثبيت عدد المصححين

الأبعاد				معامل التعميم النسبي	معامل التعميم المطلق
مفحوص	مهمة	مصحح	فترة		
243	12	3	2	0.86	0.82
243	12	3	3	0.89	0.84
243	12	3	4	0.90	0.85
243	15	3	2	0.88	0.84
243	15	3	3	0.90	0.86
243	15	3	4	0.91	0.87
243	18	3	2	0.89	0.85

تابع/ جدول (5):

الأبعاد				معامل التعميم النسبي	معامل التعميم المطلق
مفحوص	مهمة	مصحح	فترة		
243	18	3	3	0.91	0.87
243	18	3	4	0.92	0.88
243	21	3	2	0.90	0.86
243	21	3	3	0.92	0.88
243	21	3	4	0.93	0.89
243	24	3	2	0.91	0.87
243	24	3	3	0.92	0.89
243	24	3	4	0.93	0.90

يتضح من النتائج في جدول (5) أنه تم تثبيت عدد المصححين (3) لجميع الحالات، وتم عرض التدرج في زيادة عدد المهام بدءًا من (12) إلى (24) بزيادة مقدارها (3) مهام في كل مرة، وتم عرض التدرج في زيادة عدد الفترات الزمنية بدءًا من (2) إلى (4) بزيادة فترة زمنية في كل مرة. وعند الحالة (12) مهمة، و(4) فترات زمنية، يرتفع مُعامل التعميم النسبي من (0.86) إلى (0.90)، ومُعامل التعميم المطلق أيضًا ارتفع من (0.82) إلى (0.85)، في حين أن زيادة عدد المهام من (12) إلى (15) أظهر فرقًا أقل من حالة زيادة عدد الفترات الزمنية من (2) إلى (4) عند الحالة (12) مهمة؛ حيث ارتفعت مُعاملات التعميم النسبية والمطلقة على التوالي من (0.86)، (0.82) إلى (0.88)، (0.84). كذلك يلاحظ أن زيادة عدد المهام وعدد الفترات الزمنية يرفع من مُعاملات التعميم بشكل أفضل، كما أن زيادة عدد الفترات يعد أفضل من زيادة عدد المهام من أجل تحسين مُعاملات التعميم، ولكن يترتب على ذلك زيادة التكاليف والجهود المبذولة بشكل أكبر مما يترتب على زيادة عدد المهام.



الشكل (4): تمثيل بياني لمُعاملات التعميم في حال زيادة عدد الفترات الزمنية (2-4)، المهام (12)، المصححين (3)

يُلاحظ من الشكل (4) أن زيادة عدد الفترات الزمنية يسهم -بشكل ملحوظ- في تحقيق مستويات ثبات مقبولة؛ إذ إن مُعاملات التعميم النسبية والمطلقة تُساوي على التوالي (0.86، 0.82) في حال كان عدد كل من الفترات الزمنية (2)، والمهام (12)، والمصححين (3)؛ وهذا يدل على تعميم نتائج الاختبار دون اللجوء إلى زيادة عدد المصححين أو عدد الفترات الزمنية. وإذا ما قورنت مستويات الثبات تكون متماثلة نوعـًا ما عنها عند اختيار (4) فترات زمنية، وتصل مُعاملات التعميم النسبية والمطلقة لمستويات أعلى بما لا يتجاوز (0.04) من قيمة مُعاملات التعميم. وإذا نُظر إلى شكل (3) وشكل (4) يُلاحظ أن زيادة عدد مستويات الفترة الزمنية تُحقق تحسنـًا أكبر في قيمة مُعاملات التعميم النسبية والمطلقة بدرجة أكبر عنه عند زيادة عدد المصححين.

مناقشة النتائج

مناقشة نتائج السؤال الأول:

أظهرت النتائج أن أكثر مصادر التباين تأثيرًا على مُعاملات التعميم هو مصدر تباين تفاعل (مفحوص-مهمة)، وتفاعل (مفحوص-مهمة-فترة)، والمهمة، وتتفق هذه النتائج مع الدراسات (Shavelson et al., 1993; Mcbee & Barens, 1998; Smith & Kulikowich, 2004؛ طباع، 2020)؛ فمُعظم نتائج هذه الدراسات جاء فيها تباين المهمة منخفضًا باستثناء دراسات (Mcbee & Barens, 1998; Smith & Kulikowich, 2004)؛ حيث كان تباين المهمة فيها مرتفعًا، وهذا عائدٌ إلى اختلاف في درجة صعوبة المهام لجميع المفحوصين، فمهام الرأي كان أداء المفحوصين عليها متدنيًا؛ مما يُشير إلى أنها كانت صعبة لمُعظم المفحوصين، ومهام التطبيق كان أداء المفحوصين عليه مرتفعًا؛ مما يُشير إلى أنها كانت سهلة لمُعظم المفحوصين، وفيما يخص مصدر تباين الفترة وتفاعلاتها مع الأبعاد فكانت قليلة.

وربما يُعزى ارتفاع مصدر تباين تفاعل (مفحوص-مهمة-فترة) إلى تغير أداء المفحوصين عبر المهام المختلفة والفترات الزمنية المختلفة؛ إذ إن أداء بعض المفحوصين على بعض المهام كان أفضل في الفترة الأولى من الفترة الثانية، والعكس صحيح؛ مما يدل على أن المهام غير متجانسة، وإن المفحوصين قاموا باستخدام طرق لحل المهام تختلف باختلاف المهمة أو الفترة، ويُعدُّ هذا المصدر ثاني أكبر مصادر التباين تأثيرًا على ثبات أداء المفحوصين. وأكدت العديد من الدراسات أن بُعد الفترة يعد مخفيـًا؛ لأنه قلل من تباين تفاعل (مفحوص-مهمة) وامتزج به (Shavelson et al., 1993; Ruzi-Primo et al., 1993; Mcbee & Barens, 1998; Webb et al., 2000; Lee,

2016)، وأن إدراج بُعد الفترة الزمنية ضمن فترات متعددة يُنتج عنه مزيجًا من التفاعل ما بين المهمة والفترة ليُشكلا مصدرًا رئيسيـًا للخطأ، في حين إذا لم يُدرج بُعد الفترة الزمنية، فإن تغير المهمة يُعدّ مصدرًا رئيسًا من مصادر الأخطاء.

وجاءت مُعاملات التعميم النسبية والمطلقة ضمن المدى المقبول، وتؤيد نتائج هذه الدراسة نتائج الدراسات السابقة التي حققت مُعاملات التعميم المقبولة (Webb et al., 2000; Smith & Kulikowich, 2004؛ طباع، 2020). في المُقابل، فإن بعض الدراسات لم تتوصل إلى مُعاملات التعميم ضمن المدى المقبول، ومنها (Shavelson et al., 1993; Mcbee & Barens, 1998; Lee, 2016)، كما أظهرت أن إدراج بُعد الفترة الزمنية يسهم في تحسين مُعاملات التعميم؛ لأن ذلك يُحسن من تباين الدرجة الشاملة، مما يُشير إلى تحسن في ثبات أداء المفحوصين عبر الفترات الزمنية.

مناقشة نتائج السؤال الثاني:

توصلت النتائج في دراسات القرار إلى أن زيادة عدد المهام وعدد الفترات يسهم في ارتفاع مُعامل التعميم النسبي ومُعامل التعميم المطلق، وبزيادة عدد المصححين لا يرتفع مُعامل التعميم النسبي ومُعامل التعميم المطلق بشكل جوهري. وتتفق النتائج مع الدراسات (Ruzi-Primo et al., 1993; Webb et al., 2000; Huang, 2009؛ طباع، 2020)، على أن أكبر مصادر التباين تأثيرًا في ثبات أداء المفحوصين هو تفاعل(مفحوص-مهمة-فترة) وتفاعل (مفحوص-مهمة)، والمهمة، وربما يعود ارتفاعها لأسباب تتعلق بزيادة تجانس المهام، كما أن التأثير البسيط للمصححين وتفاعلاته مع الأبعاد الأخرى يُعزى إلى التدريب الجيد للمصححين على آلية التصحيح، بوجود ميزان تصحيح واضح المعايير والمحكات التي يجب أن يجتازها المفحوص للحصول على أداء جيد.

الاستنتاجات والتوصيات:

أظهرت نتائج الدراسة أن بُعد المهام يُشكل مصدرًا أكبر لتباين الخطأ من بُعد المصححين، ويعود ذلك لوجود مهام متنوعة تتناسب مع قدرات المفحوصين من حيث الصعوبة؛ فبعضها تكون سهلة وبعضها الآخر صعبة، وذلك بحسب قدرة المفحوص ونوع المهمة التي يُتقنها، إضافةً إلى أن إعداد ميزان تصحيح تحليلي، وتدريب المُصححين بشكل جيد، ومتابعة عملية التصحيح بشكل منتظم، جعل هناك اتساقًا بين تقديرات المصححين؛ مما قلل من تباين الخطأ الناتج من المصححين. كما أن إدراج بعد الفترة الزمنية قد أسهم -بشكل ملحوظ- في رفع مُعاملات التعميم، وانعكس على تحسن مؤشرات

الثبات وجعلها مُرتفعة. وأخيرًا، فإن زيادة تمثيل أبعاد المهام والفترات الزمنية يعمل على تخفيض الخطأ ورفع الثبات، في حين أن زيادة عدد المصححين لم يُقلل من الخطأ؛ لأنه لم يُسهم -بشكل جوهري- في رفع مُعاملات التعميم. وبشكل عام، فإن زيادة شروط القياس يُقلل من تباين الخطأ ويعمل على زيادة دقة القياس، شريطة زيادة شروط القياس الأكثر تأثيرًا في خفض تباين الخطأ، آخذين بالاعتبار أعباء التكلفة والجهد والزمن في ذلك.

بناءً على ما سبق من النتائج؛ فإن الدراسة الحالية توصي بإجراء بعض الدراسات المستقبلية التي تثري أدبيات القياس والتقويم بتطبيقات على نظرية التعميم، وخاصة العربية منها، وهي الآتي:

1- معرفة أثر المنفعة من دراسات القرار في ضوء الجهد والتكلفة والوقت الذي تحتاجه الدراسات المستقبلية بهدف رفع معاملات التعميم.

2- الكشف عن أثر مصادر تباين الخطأ من أدوات قياس متحررة من اللغة.

3- المُقارنة بين مُعاملات الثبات في ظل نظريات القياس الثلاث: (الاستجابة للفقرة، والتعميم، والتقليدية).

4- المُقارنة بين طرق التعامل مع مصادر التباين المتعددة في نظرية الاستجابة للفقرة متعددة الأبعاد ونظرية التعميم.

5- استخدام نظرية التعميم في معالجة البيانات المفقودة؛ بهدف استقرار الثبات عن طريق مكونات التباين.

6- المُقارنة بين مؤشرات الثبات باستخدام نظرية التعميم في برمجية (EduG)، وبرمجية (GENOVA).

7- المُقارنة بين مؤشرات الاتفاق للمصححين ضمن نظريات القياس الثلاث.

❋ ❋❋

قائمة المصادر والمراجع

أولًا: المراجع العربية:

الحربي، خليل؛ والحربي، عيد (2017). مؤشرات الثبات باستخدام نظرية التعميم ومؤشرات صدق البناء لمقياس موهبة الإبداع. *مجلة جامعة طيبة للعلوم التربوية. 12* (3)، 425-441.

عبيدات، ذوقان؛ وعدس، عبد الرحمن؛ وعبد الحق، كايد. (2005). *البحث العلمي: مفهومه، أدواته، أساليبه*. عمان: دار الفكر ناشرون وموزعون.

علام، صلاح الدين محمود. (2000). *القياس والتقويم التربوي والنفسي*. القاهرة: دار الفكر العربي.

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

عودة، أحمد (2010). *القياس والتقويم في العملية التدريسية*. اربد: دار الأمل للنشر والتوزيع.

طباع، فاروق (2020). استخدام نظرية إمكانية التعميم لتقدير ثبات اختبار تقييم كفاءة لدى طلاب السنة الرابعة ابتدائي. *المجلة الأردنية في العلوم التربوية، 16* (1)، 1-18.

ثانيًا: المراجع الأجنبية:

Al - Harbi, K. & Al - Harbi, E. (2017). Reliability indicators of using generalization theory and construct validity evidences for mawhiba creativity test (in Arabic). *Taibah University Journal for Educational Sciences. 12* (3). 425-441.

Allen, J., & Yen, W. (1979). *Introduction to measurement theory*. Monterey California: Brooks/Cole Publishing Company.

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

Brennan, R., & Kane, M. (1977). An Index of Dependability for Mastery Tests. *Journal of Educational Measurement*, *14*(3), 277-289.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern theory*. New York: Holt. Rinehart and Winston.

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

Huang, C. (2009). Magnitude of Task-Sampling variability in performance assessments: a meta-analysis. *Educational and Psychological Measurement*, *69* (6), 887-912.

Gebril, A. (2009). Score generalizability of academic writing tasks: dose one test method fit it all. *Language testing*, *26* (4), 507-531.

Geo, X., & Brennan, L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*(2), 191-203.

Lee, M. (2016). A Comparison of Generalizability Theory and Many Facet Rash Measurements in an Analysis of Mathematics Creative Problem Solving Test. *The Journal of Curriculum and Evaluation*, *19* (2), 251-279.

Marcoulides, E., & Kyriakides, L. (2010). Using generalizability theory. In B.M.B. Creamers., L. Kyriakides & P. Sammons (Eds.), *Methodological advance educational effectiveness in research* (pp.221-245). New York: Routledge

Guler, N., & Gelbal, S. (2010). Studying reliability of open-ended mathematics items according to the generalizability theory. *Educational Sciences: Theory & Practice*, *10* (2), 1011-1019.

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.

Parkes, J. (2000). The Relationship between the reliability and cost of performance assessments. *Education Policy Analysis Archives*, *8* (16), 1-14.

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. (1991). *Generalizability Theory: A primer*. California: Sage Publications.

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

Smith, E., & kulikowich, J. (2004). An Application of Generalizability theory and many facet Measurement Using Complex Problem-Solving Skills Assessment, *Educational and Psychological Measurement***,** *64* (4), 617-639.

Swiss Society for Research in Education Working Group. (2010). EduG User Guide. IRDP: Neuchatel, Switzerland. http://www.irdp.ch/edumetrie/logiciels.html

Tebaa, F. (2020). Using Generalizability Theory in Estimating Reliability of a Mathematical Competence Assessment Test of Fourth Year Primary School Students (in Arabic*). Jordan Journal of Educational Sciences*. *16* (1). 1-18.

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

❋ ❋❋