جامعة سوهاج
كلية التربية
المجلة التربوية

أثر عدد المهام الأدائيَّة وطريقة التقدير والمقيِّمين في ثبات سلالم التقدير لأنشطة تقييم الأداء لمقرَّر العلوم باستخدام نظريَّة إمكانيَّة التعميم

إعداد

د/ عائشة احمد محمد الحجاجي

حاصلة على الدكتوراه تخصص قياس وتقويم

جامعة ام القرى

معلمة بوزارة التعليم بمنطقة مكة المكرمة

تاريخ استلام البحث : ١٨ أغسطس ٢٠٢٣م – تاريخ قبول النشر: ١٧ سبتمبر ٢٠٢٣م

DOI: 10.12816/EDUSOHAG.2023.

مستخلص:

هدَفَ البحثُ إلى الكشف عن الأهميَّة النسبيَّة لجميع مصادر التباين الراجعة للمهام الأدائيَّة وطريقة التقدير والمقيِّمين والتفاعلات بينها بالنسبة للتباين الكلي، وتحديد عدد مستويات الأوجه للوصول إلى معاملات إمكانيَّة تعميم موثوقة. وللتحقُّق من الخصائص السيكومتريَّة لأداة البحث طُبّقت مهام الأداء على عيّنة من (٢٧) طالبة من طالبات الصفّ الثاني المتوسط، وتم تقديرها بطريقة كليَّة ثم تحليليَّة من قِبل ثلاثة مُقيِّمين. وأظهرت النتائج أن قِيَم مكونات التباين صغيرة بالنسبة لمصادر التباين التي ترجع إلى الوجهين طريقة التقدير والمهام والتفاعلات فيما بينها، والتي تُعطي مؤشرًا على أن تقديرات الأداء في المهام باستخدام طريقتَي التقدير متقاربة، بقيم معاملات إمكانيَّة التعميم النسبيَّة والمُطلَقة (٠,٩٢ – ٠,٨٩). وتكوّنَت عيّنة البحث من (١١٥) طالبة من طالبات الصفّ الثاني المتوسط بمدينة مكَّة المكرَّمة، نفّذْن مهام الأداء، وتم تقدير درجاتهن باستخدام سُلَّم التقدير الوصفي الكلي والتحليلي من قِبل مُقيِّمين، وتحليل البيانات بواسطة تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x طريقة التقدير x المُقيِّمين). وأظهرت نتائج دراسة التعميم: أن أكبر قيمة للتباين تَعُود إلى موضوع القياس بنسبة بلغت (٦٠,٨%)، ثم مصدر التباين الراجع للتفاعل بين (الطلَّاب x المهام) بنسبة (١٩,٨)، وبلغت قيم معاملات إمكانيَّة التعميم النسبي والمطلق (٠.٨٨–٠.٨٥). وبيَّنَت النتائج أيضًا أن زيادة مستويات وجه المهام تؤثر في قيم معاملات إمكانيَّة التعميم، وللوصول إلى قيم مرتفعة لمعاملات إمكانيَّة التعميم يُتطلَّب زيادة عدد المهام من ثلاث مهام أدائيَّة أو أكثر.

الكلمات المفتاحية: مهام الأداء، طريقة التقدير، المقيِّمين، معاملات إمكانيَّة التعميم، نظريَّة إمكانيَّة التعميم.

The Effect of The Number of Performance Tasks and Method assessment and Raters on The Reliability of Rating Scales to Evaluate Performance in Science Course Using Generalizability Theory

Abstract

The study aimed to determine the relative of all source's variance due the tasks, Method, raters and the interactions between them, and determine the conditions numbers of facets that can be relied upon to reach Generalizability coefficients to estimate student performance, the psychometric characteristics of the tools by applying the tasks on (27) Students, estimated in a holistic and analytic methods by three raters. Results were verified, and the values of the variance components were small for the sources of variance that are due estimation method, tasks and interactions between them, which give that the performance estimates in the tasks relied upon, the values coefficients relative and absolute generalizations (0,92 - 0,89). sample of 115 students, performed three performance tasks, estimated by tow raters. using fully crossed three facets design (Persons x Tasks x Method x Rater(. results showed that object of measurement ratio (60.8%), then variance due to the interaction (Students x Tasks) rate (19.8%). the relative and absolute generalizability coefficients were (0.88-0.85). The results showed that increasing the facet levels of the tasks affects the values of the generalizability coefficients, in order to reach high values, it is necessary to increase the number of tasks to three.

Keywords: Performance tasks, Method assessment, Raters, generalizability coefficients, Generalizability theory.

مقدمة:

يُعتبر نظام تقييم الأداء من الأنظمة التي تُقدِّم مزيجًا استراتيجيًّا من المعلومات؛ وذلك لأنه يوفِّر صورة أكثر شموليّة لما يمكن للأفراد القيام به ومدى جودة هذه الأداءات. وما يُميِّز التقييم القائم على الأداء إسهامه في عمليَّة تحسين التعلم لدى الطلَّاب، كما أنه آليّة لجمع المعلومات حول تمكُّن الطلَّاب وقدرتهم على توظيف المعرفة في مواقف مختلفة، كما يسهم في التشخيص الفردي للطلَّاب وتقديم تغذية راجعة فورية.

ويعتمد التقييم القائم على الأداء على النتاجات والأداءات، ومدى جودة هذا النتاج وإتقان المهارات، غير أن التركيز على هذا الجانب في الأبحاث العلميَّة منخفض مقارنةً بالاختبارات التحصيليَّة نظرًا لوجود بعض المشكلات السيكومتريَّة من الموضوعيَّة والصدق والثبات والعدالة، بالإضافة إلى مشكلات عمليَّة منها تحديد مستويات الأداء والمعايير التي ينبغي التركيز عليها، والتي تتطلَّب وقتًا وجهدًا كبيرَيْن (Dunbar, Kortez & Hoover, 1991; Burke, 2009).

ويَقُوم المُقيّمون بتقييم الأداء باستخدام قوائم الشطب أو سلالم التقدير وغيرها من أدوات التقويم البديل؛ حيث يتم جمع أدلة من خلال محكات تقدير الأداء التي تدل على مستويات كفايات الطلَّاب ومهاراتهم الأدائيَّة وقدرتهم أثناء عمليَّة التقييم للأداء (علام، ٢٠١٦؛ وزارة التعليم، ٢٠١٣).

وفي هذا السياق، أُجريت العديد من الدراسات التي اهتمت بالتقييم القائم على الأداء، والتي ركّزَت بشكل كبير على مدى ممارسة المُعلِّمين له ومشكلاته، ومنها دراسة البشير وبرهم (٢٠١٢)، والخالدي (٢٠١٤)، والمطرفي (٢٠١٥)، وبرهم (٢٠١٨)، والفهيدي (٢٠١٨)، والزبيدي (٢٠١٩)، والخروصي والذهلي (٢٠٢٠)، وفي معظم هذه الدراسات تم استخدام الاستبانة من أجل تحقيق أهدافها، في حين اهتمت بعض الدراسات بالمقارنة في الخصائص السيكومتريَّة بين الاختبارات الموضوعية واختبارات التكملة التي يقوم فيها الطلَّاب بصياغة استجابة، منها دراسة أبو جرد (٢٠١٢).

وتتطلَّب عمليَّة تصحيح وتقدير المهامّ الأدائيَّة تصميم قائمة بالمحكات التي تُشكِّل أساسًا لعمليَّة التصحيح، ورَصْد الدرجات مع تحديد واضح للنقاط الأساسية لخصائص الأداء الجيد وكيفية تقييمه، ويستخدم المُقيّمون المحكات نفسها لتصحيح المهام؛ وذلك للوصول إلى اتّساق الأحكام بين المقيِّمين، وقد تختلف درجات الطلَّاب من مهمة إلى

أخرى ومن مُقيِّم إلى آخر، وعندما يختلف تقدير الأداء فإن ذلك يؤدي إلى حدوث خطأ عشوائي في القياس مصدره اختلاف أدوات التقييم والمقيِّمين وأيضًا اختلاف عيِّنة المهام، أي أنّ هناك نسبة من التباين الكلي ترجع إلى الأخطاء العشوائيَّة في القياس وعدم اتساق التقديرات في المهام (علام، ٢٠٠٢؛ علام، ٢٠٠٩؛ فرج، ٢٠١٧؛ Gao, Shavelson & Baxter, 1994).

وتوجد طريقتان رئيسَتان لتقدير أداء الطلَّاب في التقييم القائم على الأداء، هما: الطريقة الكليَّة، والطريقة التحليليَّة؛ وتُركِّز الطريقة الكليَّة على مهارات محدَّدة جدًّا؛ لذا فهي تتضمَّن معايير بعبارات وصفية محدَّدة وموجزة وقابلة للملاحظة، وتقدير درجة للمهمة الأدائيَّة بوجه عام؛ في حين أن الطريقة التحليليَّة يتم فيها تحليل المهمة الأدائيَّة بعبارات وصفيَّة للمحكات المكوّنة لهذه المهمة التي تتدرَّج بمستويات متباينة، ويتم تقدير درجات لكل محك ثم حساب الدرجة الكليَّة لجميع المحكات (علام، ٢٠٠٩، Popham, 2003;).

وعند الرغبة في تقدير درجات عيِّنة من أداء الطلَّاب من مجموعة من المهام الأدائيَّة، ثم تُقيَّم بالاستعانة بمقيِّمين من بين مجموعة من المقيِّمين المدرَّبين باستخدام سلالم تقدير لتقييم الأداء من بين مجموعة من أدوات التقييم البديل، فإنه يمكن الاعتماد على نظريَّة إمكانيَّة التعميم في تحليل البيانات الخاصة بتقييم الأداء في مناسبة واحدة، ثم تضمين مكونات التباين التي تؤثر في تقدير درجات الطلَّاب، وتقدير قِيَم التباين لهذه المكونات والتفاعلات فيما بينها وتأثيراتها على صدق وثبات درجات الأداء للطلَّاب (Webb, Schlackman & Sugrue, 2000; Shavelson, Gao & Baxter, 1993).

ويتحدد الافتراض الأساسي لنظريَّة إمكانيَّة التعميم هو أن موضوع القياس (Object of Measurement) لسمة معينة، أو سلوك، أو معارف، أو مهارة يُفترَض أنها مستقرَّة لا تتغيَّر، كما أن التباين بين الأفراد فيها هو تبايُن مرغوب فيه؛ لذا فإن أيّ اختلافات ملحوظة بين الدرجات التي تم الحصول عليها أثناء عمليَّة القياس في المُقيِّمين، أو فقرات مختلفة؛ ترجع فقط إلى مصادر تبايُن الخطأ المنتظِم، وموضوع القياس في تقييم مهام الأداء يُركِّز على التباين بين الطلَّاب في جودة وتعقيد الأداء الذي قام به الطلَّاب في موضوع مُعيَّن، ويجب التمييز بينه وبين بقية مصادر التباين التي قد تؤدي إلى حدوث

خطأ في القياس، والتي تشمل الأوجه الفردية والتفاعلات بين هذه الأوجه في التصميم الذي أعدَّه الباحث (Shavelson & Webb, Yin & Shavelson, 2004; 1991).

وتعد الدرجات وفق هذه النظريَّة هي درجات مثالية مقبولة لهذا الغرض، وهي النتيجة التي سيستند إليها القرار، فهي واحدة من مجموعة من الدرجات تخدم نفس الغرض، وبالتالي لا يهم استجابة الأفراد على فقرات محدَّدة في موقف مُعيَّن في لحظة محدَّدة وتسمى الدرجة الشاملة (Universe Score)، وهي تلك الدرجة المتوقَّعة من بين مجموعة الدرجات المستنتَجة من مجموعة من الملاحظات في مختلف المواقف التي تنتمي إلى النطاق الشامل للملاحظات المقبولة (Universe of Admissible Observations)، وتُعرف بأنها: متوسط الدرجات التي يحصل عليها الفرد من مختلف الاختبارات والمواقف والمقيِّمين والمهام وغيرها. بمعنى آخر: هي كل الشروط الممكنة للأوجه (حبيب وعزيز، ٢٠١٨؛ الطريري، ٢٠١٤؛ العرايضة والشريفين، ٢٠٢٢؛ Brennan, 2001).

والدرجات التي يحصل عليها الطلَّاب والمُعبّرة عن قدراتهم لا تتغيَّر بتغيُّر إعادة تطبيق أداة القياس، أو اختلاف في ظروف التطبيق، أو تطبيق فقرات مختلفة، أو تحت أيّ ظروف متغيرة، وتُسمَّى في نظريَّة إمكانيَّة التعميم بالدرجة الملاحظة (Observed score)، وهي مجموع الدرجة الشاملة ودرجات الخطأ في كلّ مصادر التباين المتضمَّنة في التصميم؛ حيث يتم استخلاص العوامل التي تسهم في انحراف الدرجات الملاحَظة عن الدرجة الحقيقية لاتخاذ قرار حول كيفية التقليل من مقدارها في الدراسات المستقبلية (Bottema-Beutel, Lloyd, Carter & Asmus, 2014; Webb & Shavelson, 2005; Wobiele & Queensoap, 2020).

وتتعامل نظريَّة إمكانيَّة التعميم مع مصادر تبايُن الخطأ المحتملة، فهي تؤكد وجود مصادر تبايُن متنوعة من الخطأ المنتظم أثناء القياس الواحد تُسمَّى أوجُهًا أو أبعادًا (Facets) لا بد أن تُؤخذ بعين الاعتبار، ففي دراسة إمكانيَّة التعميم يتم وصف الملاحظة سواء كانت الفقرات أو المقيِّمين أو المواقف التي تُعتبر مصدرًا مهمًا من مصادر التباين الضرورية في القياس، كما يُطلَق على مستويات هذه الأوجه الشروط أو الظروف (Conditions)، مثل الفقرات المختلفة أو المقيِّمين المختلفين أو مواقف مختلفة؛ حيث

يتم تقدير قيم تباينات الخطأ المنتظم المتعدِّدة الأوجُه بشكل منفصل والتفاعلات بين الأوجه في التصميم (رونالد وسيسيل، ٢٠١٣؛ علام، ٢٠٠٢ ; Shavelson & Webb, 1991).

ويمكن أن تكون الأوجه عشوائيَّة (random facet)، أو ثابتة (fixed facet)، أو محدودة العشوائيَّة (random finite) بناءً على طريقة أَخْذ العينات الخاصة بالأوجه. ويقال عن الوجه إنه وجه عشوائي إذا تم اختيار شروط الوجه في التحليلات بصورة عشوائيَّة من مجتمع الشروط المحتملة، في حين يُعتبر الوجه ثابتًا إذا تمت تغطية جميع حالات شروط المجتمع الشامل، ويُسمَّى الوجه محدود العشوائيَّة أو مختلطًا في حال تطبيق المعاينة على مجتمعات محدَّدة من مجتمع الشروط. وبوجهٍ عامّ كلما زاد عدد الأوجه في التصميم قلَّت قِيَم تبائُنات الخطأ المنتظم، وازدادت قيم معاملات إمكانيَّة التعميم (الأحمدي، ٢٠١٨؛ Brennan, 2011; Scheuneman, 1982; Suen, Lu, & Neisworth & Bagnato. 1993).

وقد استخدم عدد من الباحثين نظريَّة إمكانيَّة التعميم من أجل تقييم المهام الأدائيَّة في مقرَّر العلوم، ومن هذه الدراسات (Shavelson, Gao & Baxter, 1993; Gao, Shavelson & Baxter, 1994; Ruiz–Primo, Baxter & Shavelson, 1993; Webb, Schlackman & Sugrue, 2000).

ومما تجدر الإشارة إليه، أن هناك العديد من الدراسات في نظريَّة إمكانيَّة التعميم التي ركّزَت على تحديد عدد المهام الأدائيَّة وعدد المقيّمين الأنسب، بالإضافة إلى بعض العوامل المؤثرة في تقييم الأداء ومن هذه الدراسات، دراسة شافيلسون وروبز بريمو ووايلي (Shavelson, Ruiz–Primo & Wiley, 1999) حيث هدفت إلى معرفة أثر التباين في المهام الأدائيَّة، باستخدام تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المواقف x المهام x المقيّمين)؛ حيث تم اختيار (٢٧) طالبًا واستخدام ثلاث طرائق للتقييم، هي: الملاحظة المباشرة لأداء الطلَّاب العملي التجريبي، وتدوين ما قاموا به من تجربة في دفتر العلوم، وإجراء تجربتين باستخدام المختبر الافتراضي باستخدام الحاسوب. وتُعزَى أكبر نسبة للتباينات إلى تبائُن البواقي بقيمة (٥٩%) من إجمالي التباين الكلي، تليها تفاعُل (الطلَّاب x المهام) بنسبة (٣٢%) من التباين الكلي؛ مما يدل على أن الطلَّاب يتغيَّر أداؤهم من مهمة إلى أخرى، وقد بيّنَت نتائج دراسة القرارات أن عدد

المهام الأدائيَّة ينبغي أن يتراوح ما بين (٦) و(٢٠) مهمة للحصول على معاملات إمكانيَّة التعميم موثوقةً.

وأيضًا من الدّراسات التي اهتمَّت بتقييم الأداء في مقرَّر العلوم دراسة ويب وشليكمان وشوجر (webb, Schlakman & Sugrue, 2000) التي تهدف إلى معرفة أهميَّة المناسبات كمصدر من مصادر التباين، باستخدام تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x المناسبات x المقيِّمين)، بإعداد مهمتين، هما: اختبار عملي تطبيقي، واختبار ورقة وقلم يتضمَّن صورًا للعناصر نفسها المستخدَمة في الاختبار العملي باستثناء أن الطلَّاب تم إعطاؤهم صورًا بدلًا من الأدوات، وطُبّقت على (٥٧) طالبًا. وأظهرت النتائج ارتفاع نسبة التباين العائد إلى موضوع القياس في حالة الاختبار العملي بنسبة (٦٧%) مقارنةً بالتطبيق الورقي بقيمة تبايُن (٥٧%)، يليها نتيجةً للتفاعل بين (الطلَّاب x المهام) بنسبة (٤١%) في المناسبة الأولى، وبنسبة (٣٣%) في المناسبة الثانية؛ أما مهمة الورقة والقلم فإن أكبر نسبة للتباين عائدة في كلتَا المناسبتين إلى موضوع القياس. وبيّنَت النتائج أنّه لزيادة قيم معامل إمكانيَّة التعميم النسبية والمطلقة للاختبار العملي فإن ذلك يتطلَّب مهمتين أو أكثر، في حين أن قيم معامل إمكانيَّة تعميم النسبية والمطلقة مرتفعة بوجه عام في الاختبار المعتمِد على الورقة والقلم.

في حين ركّز الباحثان هان وهوانغ (Han & Huang, 2017) على تقييم (٣٦) مقالة للطلَّاب، بالطريقة الكليَّة ثم التحليليَّة من قِبل عشرة مُقيّمين بفارق زمني بين التقييمَيْن، واستُخدم تصميم ثنائي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x طريقة التقييم x المقيِّمين). وقد بيّنَت النتائج: أن أكبر مصدر للتباين يرجع إلى موضوع القياس بنسبة بلغت (٤٠%) من إجمالي التباين الكلي، يليه تبايُن البواقي وأخطاء القياس، كما أظهرت النتائج وجود تفاعُل بين (المقيِّمين x طريقة التقدير). وبيّنَت النتائج أيضًا أن قيمة التباين الراجع لموضوع القياس والفرق بين الطلَّاب في جودة المقالات جاءت مرتفعةً وفقًا للتقييم باستخدام الطريقة الكليَّة والطريقة التحليليَّة، وقد أظهرت نتائج دراسات القرارات أنه يمكن الحصول على معامل إمكانيَّة تعميم يمكن الوثوق به بقيمة (٠,٨) في كلٍّ من الطريقتَيْن في عمليَّة التقييم، وهو الاعتماد على ستة مُقيّمين.

أما دراسة تاشيديلن وشاهين وبايتيمير (Taşdelen, Şahin & Baytemir, 2016) فقد اهتمَّت بدراسة فعالية تقييم الأقران كإحدى تقنيات التقييم القائم على الأداء في مهام الأداء. وتكوّنَت عيّنة الدراسة من (٤١) طالبًا، تم اختيار ثلاثة من الطلَّاب بصورة عشوائيَّة كمُقيّمين، قاموا بتقييم أداء (٣٨) من زملائهم بعد عمليَّة تدريبهم، ثم حساب متوسط التقييم للأقران، وقام المعلم أيضًا بتقييم المهام للطلَّاب، وتم تحليل البيانات وَفْق تصميم ثنائي الوجه المتقاطع كليًّا المتضمّن (الطلَّاب x المهام x المقيّمين). وأظهرت النتائج قيمًا مرتفعة لمعاملات إمكانيَّة التعميم النسبية والمطلقة (٠,٨٦ و٠,٨٢) على التوالي، ووفقًا لنتيجة دراسات القرارات يمكن الحصول على معامل إمكانيَّة تعميم مقبول من خلال مُقيّمين بقيمة معاملات (٠,٨٢ و٠,٧٨) على التوالي، وأن تكون المهام المُوكَلة للطلَّاب من (٦) إلى (٨) مهامّ أدائيَّة للحصول على قيم معاملات إمكانيَّة التعميم للقرارات المطلقة والنسبية في المدى المقبول.

وقد اهتمت دراسة بوير، بوجين ، ساندرز وفان Bouwer, B´eguin, (Sanders & Van, 2015) بمعرفة أثر التنوع للمهام الكتابية في التمييز بين الأداءات الكتابية لمجموعة من الطلَّاب، وعدد المهام والمقيّمين الأنسب للوصول إلى معامل إمكانيَّة تعميم موثوق، واختيار عيّنة عشوائيَّة من (٦٧) طالبًا أكملوا (١٢) مهمة كتابيَّة تتضمَّن كتابة (رسالة، مقال، قصص من نوع مغامرة، وقصص شخصية) في ثلاث فترات، وتقييمها باستخدام سُلَّم تقدير وصفي من قِبل ثلاثة مُقيّمين تم اختيارهم بصورة عشوائيَّة، واستخدام تصميم ((p x (r:t:g)). واتضح من النتائج أن أكبر قيم للتباين راجعة للبواقي (p(r:t:g), e)، يليها التباين الذي يعود إلى (الطلَّاب x تنوُّع المهام: المهام) بنسبة (١٩%)، وللوصول إلى معامل إمكانيَّة تعميم ينبغي للطلَّاب كتابة ما لا يقلّ عن أربعة نصوص لكل ستة أنواع من المهام الكتابية وتقييمها من قِبل ثلاثة مُقيّمين على الأقل، كما بيّنَت النتائج وجود تفاعل بنسبة بلغت (٢٤%) بين (المقيّمين x تنوع المهام) واختلاف أحكام المقيّمين وفقًا لتنوع المهام، وأن الحكم يعتمد على نوع المهمة الكتابية التي يقوم بها الطلَّاب، وأنه للحصول على تقييمات يمكن الاعتماد عليها وموثوقة يجب على الطلَّاب على الأقل كتابة ثلاثة نصوص في كل نوع من الأنواع الأربعة، مُقيَّمة من قِبل اثنَيْن على الأقل من المقيّمين.

واهتمّت دراسة غراهام وهيبرت وساندبانك وهاريس (Graham, Hebert, Sandbank and Harris.2016) باختبار عدد عينات الكتابة اللازمة للوصول إلى تقديرات موثوقة في القدرات على الكتابة، والاستعانة بـ (١٢) معلمًا طُلب منهم تحديد الطلَّاب الذين يعتقدون أنهم يعانون من ضعف في الكتابة ثم طُبق عليهم اختبار معياري لتحديد مستوى مهارة الكتابة، واختيار الضعيفِين في مهارة الكتابة الحاصلين على ٢٥% أو أقل في الاختبار؛ حيث تم اختيار (٨١) طالبًا و(٥٦) طالبة، وبعد جَمْع البيانات قام مُقيّمون بتقييم كتابات الطلَّاب من دون معرفة الهدف من الدراسة، حيث كان الاتفاق بين المقيّمين مرتفعًا (٠,٩)، وتم تطبيق نظريَّة إمكانيَّة التعميم على درجات الطلَّاب في ثلاثة مقاييس للكتابة تتضمَّن (العدد الإجمالي للكلمات، تنوُّع المفردات، جودة الكتابة)، بالاعتماد على تصميم أحادي الوجه المتقاطع المتضمِّن (الطلَّاب x مهام الكتابة)، واتضح من النتائج أنه للحصول على تقديرات ثبات مقبولة فإنه ينبغي أن تكون عينات الكتابة بوجهٍ عام من ٦ عينات فأكثر للحصول على معامل إمكانيَّة تعميم بقيمة (٠,٨).

في حين هدفت دراسة القحطاني (Alkatani, 2012) إلى المقارنة بين أدوات تقييم الأداء للتلاوة القرآنية باستخدام طريقتين؛ طريقة تقييم الأداء الشفوي الكليَّة، والتحليليَّة؛ من خلال ثلاثة مُقيّمين. واشتملت العينة على (٩٣) طالبًا، باستخدام تصميم ثنائي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المقيّمين x طريقة التقييم)، وبيّنَت النتائج أن أكبر تباين يرجع إلى موضوع القياس بنسبة (٨١%)، وللحصول على تقييمات يمكن الاعتماد عليها في الأداء الشفوي وتلاوة القرآن فإنه يمكن الاعتماد على مُقيّم واحد في عمليَّة تقييم تلاوة القرآن؛ حيث كانت قيم معاملات التعميم أكبر من (٠,٨)، في حين تبيَّن عند استخدام تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المقيّمين x طريقة التقييم x أبعاد الأداة) قيمة مرتفعة أيضًا للتباين العائد إلى موضوع القياس بنسبة مرتفعة، وأيضًا تضمّنَت دراسات القرارات أنه يمكن الاعتماد على مُقيّم واحد في عمليَّة تقييم طالب واحد في التلاوة القرآنيَّة.

ومن الدراسات التي بيّنَت أهميَّة استخدام سلالم التقدير الوصفي وتأثُّرها بدقة تقدير المقيّمين وأحكامهم على تقديرات أداء الطلَّاب، دراسة كانكوي وأوزدر (Cankoy & Ozder, 2017)؛ بهدف بناء نموذج سُلَّم تقدير وصفي لتقييم مهارات حل المشكلات في مقرّر الرياضيات، ثم التأكُّد من ثبات درجات الطلَّاب باستخدام نظريَّة إمكانيَّة التعميم لدى

(٢٥) طالبًا في ثلاث مهام أدائيَّة، واستُخدم تصميم (الطلَّاب x الفقرات x المقيِّمين). وقد أوضحت النتائج: ارتفاع نسبة التبايُن الذي يعود إلى المقيِّمين بقيمة (٢٤%) من إجمالي التباين الكلي في حالة عدم الاعتماد على أداة التقييم في تقدير درجات الطلَّاب، في حين انخفضت قيم التبايُن بين المقيِّمين إلى (٠,١%) في حالة الاعتماد على أداة التقدير، أما بالنسبة للتفاعل بين (المقيِّمين x الفقرات) فإن نسبة التباين الخاصة به في حالة عدم استخدام نموذج سُلَّم تقدير وصفي بلغت (١٠%) من إجمالي التباين الكلي، وكانت قِيَمها صفرًا في حالة استخدام نموذج سُلّم التقدير الوصفي، كما أظهرت النتائج أن الزيادة أو التقليل من الفقرات لهما تأثير على معاملات إمكانيَّة التعميم أكثر من تأثير الزيادة أو التقليل من عدد المقيِّمين.

كما أجرى كلٌّ من كساب وفداء ورضوان وحسن أبو حجلة وأوكونور (Kassab, Fida, Radwan, Hassan, Abu-Hijleh & OConnor, 2016) دراسة بهدف تطبيق خرائط المفاهيم وإيجاد معاملات إمكانيَّة التعميم لأداة قياس لمهام من خرائط، حيث طُلب من (١١٦) طالبًا جامعيًّا إنشاء ثلاث خرائط مفاهيم متكاملة، وتم تقييم خرائط المفاهيم بشكل مستقل من قِبل أربعة مُقيِّمين، واستخدام تصميم ثنائي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x الأبعاد أداة قياس خرائط المفاهيم x المقيِّمين)؛ حيث جاءت أكبر قيمة لنسب التباين الراجع لموضوع القياس "تمثيل درجات الطلَّاب في خرائط المفاهيم" بنسبة (٤٧%) من إجمالي التباين الكلي، وبقيمة معامل إمكانيَّة تعميم مرتفع. في حين بيّنَت النتائج الخاصة بدراسات القرارات أنه للحصول على معامل إمكانيَّة تعميم يبلغ (٠,٨) فإنه ينبغي أن يُجري الطلَّاب خريطتي مفاهيم يتم تقييمها بثلاثة مقيِّمين، أو خريطة مفاهيم واحدة يُقيّمها خمسة مُقيِّمين، وأظهرت النتائج أن معامل إمكانيَّة التعميم ترتفع قيمته بزيادة عدد المقيِّمين بدلًا من زيادة عدد أبعاد تقييم أداة قياس خريطة المفاهيم.

كما ركزت دراسة طباع (٢٠٢٠) على تقدير ثبات اختبار تقييم كفاءة الطلَّاب في الرياضيات، بالاستعانة باختبار يتضمَّن تسع مهمَّات معقَّدة موزَّعة على ثلاث صيغ، وذلك بهدف تقييم كفاءة الأعداد والحساب، على عيّنة بلغت (٣٣١). وقُيّمت أداءات الطلَّاب بوساطة ثلاثة مُقيِّمين مُدرَّبين على استخدام سُلَّم التقدير التحليلي، واستخدام تصميم ثنائي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x المقيِّمين). وأظهرت

نتائج الدراسة وجود مصادر تبايُن خطأ أثر تفاعُل (الطالب x المهمة) والتأثير الرئيس للمهمة. ومن أجل ضمان مستويات ثبات مقبولة يجب زيادة عدد المهمات وليس عدد المقيِّمين.

ومن الدراسات التي طبّقت نظريَّة إمكانيَّة التعميم في مقرّر الرياضيات دراسة العرايضة والشريفين (٢٠٢٢)، وذلك بهدف الكشف عن إسهام مصادر التباين المتعددة في ثبات اختبار الرياضيات، باستخدام تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x المقيِّمين x الفترات الزمنية)، وتكوّنَت عيِّنة الدراسة من (٢٤٣)، طُبق عليهم اختبار الرياضيات المكوَّن من (١٢) مهمة مرتين بفاصل زمني مدته أسبوعان، وقُيّم من قِبل ثلاثة مُقيِّمين. وأظهرت النتائج: أن أكثر مصادر التباين تأثيرًا في المعاملات هي تفاعل (الطلَّاب x المهام)، وأن أداء الطلَّاب تحسَّن في الموقف الثاني، كما أظهرت دراسات القرارات أنه ينبغي زيادة عدد المهام وعدد الفترات الزمنية في حالة الرغبة في رفع معاملات التعميم مقارنةً بزيادة عدد المقيِّمين.

مشكلة البحث:

إن عمليَّة تحديد عدد المهام الأدائيَّة المناسبة التي ينبغي تطبيقها لإصدار حكم حول كفاءة الطلَّاب في مهام الأداء في مقرَّر العلوم –من الأمور المعقَّدة، ففي حالة الاختبار التحصيلي فإن عيِّنة الفقرات التي يتم إدراجها في الاختبار تكون ممثِّلة للمحتوى، في حين أن النطاق السلوكي غير محدَّد في حالة التقييم القائم على الأداء (Dunbar, Kortez & Hoover, 1991).

وذكر كلٌّ من: علام (٢٠٠٩) ووزارة التعليم (٢٠١٣) أنه في حالة الرغبة في تحقيق أهداف عمليَّة تعليم العلوم لا بد من النظر إلى التقنيات المتّبَعة في عمليَّة تقدير أداء الطلَّاب التي تتطلَّب معاييرَ وطرقَ تقييمٍ منقحةً واعيةً؛ لضمان عدم تدخُّل عوامل خطأ غير مرغوبة أثناء تقييم الأداء والتقليل من تأثيرها؛ للحصول على دقة في القياس بأقل قَدْر من الأخطاء في عمليَّة تقدير المهام الأدائيَّة في مقرَّر العلوم.

فالمعوّق الأكبر في عمليَّة تنفيذه في الميدان التربوي يكمن في الجانب السيكومتري، خصوصًا ثبات الدرجات. وهذا ما أشار إليه الدوسري (٢٠٠١) الذي بيَّن أن طبيعة المهام في تقييم الأداء يصعب الحصول فيها على معامل ثبات درجات مناسب؛

نتيجةً لتعدُّد مصادر الأخطاء العشوائيَّة، ووجود عوامل متعددة تؤثر سلبًا على عمليَّة التقييم.

كما أشار برينان (Brennan, 2001) إلى أن نظريَّة إمكانيَّة التعميم مناسبة بصورة خاصة للتقييمات التي تعتمد على مهام الأداء. وهذا ما أكده طباع (٢٠٢٠) حيث بيَّن أنه يمكن الوصول إلى دقة في عدد المهام الأدائيَّة وعدد المقيِّمين الأنسب للحصول على معامل ثبات يمكن الاعتماد عليه عند استخدام أداة سُلَّم التقدير الوصفي في تقدير درجات الطلَّاب.

ومن هذا المنطلق يتضح أن إجراءات نظريَّة إمكانيَّة التعميم التي تُستعمل لجمع المعلومات وتحليلها إحصائيًّا، والكشف عن مصادر التباين المختلفة والتفاعلات بينها وتأثيرها في دقة النتائج –هي الأنسب في تقدير درجات المهام الأدائيَّة للطلَّاب، وتحديد عدد المهام الأدائيَّة وطريقة التقدير والمقيِّمين الأمثل في تقدير ثبات درجات سلالم التقدير الوصفية، والحصول على نتائج يمكن الاعتماد عليها؛ وبالتالي فإن مشكلة البحث يمكن صياغتها في التساؤلات الآتية:

١– ما نسبة إسهام تباين الخطأ الذي يُفسِّره كلٌّ من أوجُه المهام الأدائيَّة وطريقة التقدير والمقيِّمين والتفاعل بينها في التباين الكلي في ثبات درجات المهام الأدائيَّة في مقرَّر العلوم؟

٢– ما الشروط الواجب توافرها لتحقيق أفضل مستويات عدد المهام الأدائيَّة وطريقة التقدير وعدد المقيِّمين على معاملات إمكانيَّة تعميم درجات أداء الطلَّاب في مقرَّر العلوم؟

أهداف البحث:

يسعى البحث الحالي إلى:

- الكشف عن الأهميَّة النسبية لجميع مصادر التباين في التصميم المستخدَم (الطلَّاب x المهام x طريقة التقدير x المُقيِّمين)؛ وذلك لتقدير ثبات درجات المهام الأدائيَّة في مقرَّر العلوم.
- تحديد عدد مستويات الأوجه المتمثِّل في المهام الأدائيَّة وطريقة التقدير والمقيِّمين الذي يعد تباين خطأ منتظم وأثره على ثبات درجات الطلَّاب؛ وذلك بهدف خفض مصادر تباين الخطأ، ورَفْع قيمة معامل إمكانيَّة التعميم.

أهميَّة البحث:

تتمثل أهميَّة البحث فيما يلي:

- يُعَدّ البحث من البحوث التي اهتمَّت بتقدير نتاجات وتقييم الأداء للطلَّاب في مقرَّر العلوم، من خلال استخدام أدوات التقييم البديل.
- تسليط الضوء على مميزات استخدام نظريَّة إمكانيَّة التعميم للتحقُّق من الخصائص السيكومتريَّة للتقييم القائم على الأداء.
- يُقدِّم البحث الحالي عدد عينات أدوات التقييم البديل ومهام الأداء والمقيِّمين الكافي حتى تكون القرارات موثوقة ويمكن الاعتماد عليها.

التعريفات الإجرائية:

نظريَّة إمكانيَّة التعميم: هي مختلف الإجراءات والأساليب الإحصائية المتّبَعة لتقدير مصادر تبايُن الخطأ في التصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x طريقة التقدير x المُقيِّمين)، ومعاملات إمكانيَّة التعميم النسبيَّة والمطلقة (طباع، ٢٠٢٠).

الثبات: يُعبَّر عنه بمعامل إمكانيَّة التعميم النسبي الذي يشير إلى اتساق درجات الطلَّاب عَبْرَ المهام وطرق التقدير والمقيِّمين لتحديد المكانة النسبية للطلَّاب بين زملائهم، ومعامل إمكانيَّة التعميم المطلق الذي يشير إلى اتساق درجات الطلَّاب عَبْرَ المهام وطرق التقدير والمقيِّمين؛ بهدف مقارنة أدائهم بمحكّ أداء خارجي (طباع، ٢٠٢٠).

طرق التقدير: وهي الطريقة التي تُستخدَم في تحديد درجة الطلَّاب في المهام الأدائيَّة، بالاعتماد على أدوات التقييم البديل، وهي في البحث الحالي سُلَّم التقدير الكلي والتحليلي، والمتضمّنة مجموعة مؤشرات الأداء معرفة إجرائيًا لتقدير مستوى كفاءة الطلَّاب في أدائهم لمهمَّة معينة (علام، ٢٠٠٩).

مهام الأداء: أيّ نشاط تعليمي (شفوي أو تحريري أو عملي) يُطلَب من المتعلِّم القيام به لإثبات معرفته وفهمه وكفاءته، وتطبيق القواعد والقوانين والمبادئ التي تمت دراستها، ثم تقييم ما أتقنه الطلَّاب (وزارة التعليم، ٢٠١٣). وتُعرَّف إجرائيًّا بأنها: درجة الطالبات على المهام الأدائيَّة الفردية في مقرَّر العلوم.

المُقيِّمون: عيِّنة المعلمات المدرَّبات على عمليَّة تقييم نتاجات الطالبات في المهام الأدائيَّة الثلاث.

حدود البحث:

تمثلت حدود البحث فيما يأتي:

- عيِّنة البحث وقوامها (١١٥) من طالبات الصف الثاني المتوسط بمدينة مكَّة المكرَّمة، وعيّنة من ثلاث مهام أدائيَّة صُمّمت للإجابة عن تساؤلات البحث، وعينة عشوائيَّة من مُقيّمتَيْن من معلمات مادة العلوم للصف الثاني المتوسط بعد تدريبهن على عمليَّة تقدير درجات الطالبات وتقييمها بالاعتماد على سلالم التقدير.
- سُلَّم تقدير كلي وسُلَّم تقدير وصفي تحليلي لتقدير درجات المهام الأدائيَّة الثلاث في مقرَّر العلوم للصف الثاني المتوسط الفصل الدراسي الثالث لعام ١٤٤٣هـ.

منهج البحث:

استُخدم المنهج الوصفي؛ وذلك لدراسة الخصائص السيكومتريَّة لأداة سُلَّم التقدير الوصفي التحليلي، ووَصْف مصادر التباين الخطأ المنتظم، وتحديد أكثر مصادر التباين تأثيرًا في التصميم، وتقدير معامل إمكانيَّة التعميم النسبية والمطلقة، والمنهج شبه التجريبي في الجزء الخاص بدراسات القرارات في نظريَّة إمكانيَّة التعميم؛ من أجل الكشف عن تأثير تغيير شروط أوجه التصميم الأكثر فاعليّة وكفاءة في رَفْع قيم معاملات إمكانيَّة التعميم النسبية والمطلقة إلى قيم مقبولة يمكن الوثوق بها.

عينة البحث:

يتكون مجتمع البحث من جميع طالبات الصف الثاني متوسط المنتظمات والبالغ عددهن (١٥١٠٧) طالبة، وتم تطبيق أدوات البحث على عيّنة عشوائيَّة قوامها (١١٥) من طالبات الصف الثاني المتوسط بمدينة مكَّة المكرَّمة لعام ١٤٤٣هـ، وعينة عشوائيَّة من مقيمتَيْن من المعلمات المسند إليهن تدريس العلوم للصف الثاني المتوسط، وعينة من ثلاث مهام أدائيَّة صُمّمت للإجابة عن تساؤلات البحث، وتم اختيار مجالها بطريقة عشوائيَّة بسيطة من مجتمع يضمّ جميع المجالات التي تتضمَّن مهمة واحدة لجميع مقررات العلوم في المرحلة المتوسطة في الفصل الدراسي الثالث لعام ١٤٤٣هـ.

أداة البحث:

أداة البحث تمثَّلَت في سُلَّم تقدير كلي وسُلَّم تقدير وصفي تحليلي لتقييم ثلاث مهام أدائيَّة في الفصل العاشر من مقرر العلوم للصف الثاني المتوسط؛ مما يتطلَّب من الطالبات (الكتابة العلميَّة، البحث عن المعلومات، إعداد ملصق) من الفصل العاشر (موارد البيئة وحمايتها)، وذلك بعد الاطّلاع على الأدب التربوي المتعلِّق بسلالم التقدير الوصفي بشقَّيْه الكلي والتحليلي؛ ومن هذه الأدبيات دراسة الخليلي (١٩٩٨)، وخليل (٢٠١١)، وعلام (٢٠٠٩)، وزيتون (٢٠٠٧)، وبيرك (Burke, 2011)، وهاموند (Hammond, 2014). وتحديد الأهداف لمهام الأداء والتي ترتبط بصورة وثيقة بالمقرّر الدراسي، والمتطلَّبات التي ينبغي أداؤها من قِبل الطالبات، وتوقُّع واضح لنوعية هذه المهام وطرق إنجازها، ثم تحديد محكات تقدير الأداء بأوصاف النوعية توضِّح خصائص المنتج، وكمثال لمهمة أدائية (اعداد ملصق: قم بإعداد ملصق مدعوم بصور بالمعلومات التي تجدها حول ترشيد الاستهلاك، وإعادة الاستخدام، وإعادة التدوير مستعيناً بالمراجع عبر شبكة الإنترنت، وذكر امثلة حول حماية الموارد البيئية والحفاظ عليها يستطيع زملاؤك في المدرسة القيام بها). ويكون تقدير درجات الطالبات في أدائهن لكل مهمة أدائيَّة بشكل كلي في سُلَّم التقدير الكلي دون تجزئة الأداء إلى مستويات، وبالاعتماد على محكات عامة للأداء، أما تقدير الدرجات لأداة سُلَّم التقدير الوصفي التحليلي فمن خلال محكَّات تتضمَّن مستويات الأداء بصورة أوصاف هي: (مبتدئ، متوسط، متقدم، ماهر)، والتعبير عنها بمتصل الأرقام التي تتدرج من الرقم ٤ (الأعلى)، ٣، ٢، ١ (الأدنى)؛ حيث تم التحقق من توفر مؤشرات الصدق من خلال عرض المهام الأدائيَّة وسلالم التقدير على مجموعة من المحكّمين البالغ عددهم (٩) محكمين؛ بهدف إبداء آرائهم في مدى صحة قياس النتاجات التعليميَّة، وتم الأخذ بملاحظاتهم في الاعتبار وإجراء التعديلات المناسبة.

وللتحقُّق من الموضوعية تم التأكُّد من إكمال عمليَّة التدريس قبل البدء بتقديم المهام الأدائيَّة للطالبات، والتأكُّد من وضوح التعليمات، وطريقة الاستجابة، وطريقة تقدير الدرجات، والزمن المخصَّص للتنفيذ، وتوفير نسخة من سُلَّم التقدير الكلي ومصفوفة التقدير التحليليَّة المعتمدة في تقدير الدرجات للنتاجات لجميع الطالبات، والحرص على أن تكون طريقة تقدير الدرجات لنتاجات الطالبات واحدة من قِبل المقيّمين، وذلك من خلال تنفيذ البرنامج التدريبي لعينة المقيّمين.

وللتحقُّق من الصدق التقاربي لأداة البحث تمَّ حساب مدى تقارُب الطرق المختلفة لتقدير درجات التقييم القائم على الأداء، بحيث يتم اعتبار طرق تقدير الدرجات وجهًا منتظمًا عشوائيًّا، ويتكون من شرطَيْن، هما: (سُلَّم التقدير الكلي وسُلَّم التقدير الوصفي التحليلي)؛ لتقييم درجات عيِّنة حساب الخصائص السيكومترية مكونة من (٢٧) طالبة من طالبات الصفّ الثاني المتوسط بمدينة مكَّة المكرَّمة ، وتحليل البيانات بالاعتماد على تصميم ثنائي الوجه المتقاطع كليًّا (الطلاب× طريقة التقدير× المهام)، ويشار إليها باختصار (P x m x t)؛ ويوضح الجدول (١) نتائج تقدير مكونات التباين في تصميم ثنائي الوجه المتقاطع كليًّا (P x m x t)؛ للكشف عن مؤشر الصدق التقاربي لأدوات التقييم المستخدَمة.

جدول (١):
نتائج تقدير مكونات التباين في تصميم ثنائي الوجه المتقاطع كليًّا (الطلَّاب× طريقة التقدير× المهام) للكشف عن مؤشر الصدق التقاربي لأدوات التقييم المستخدَمة.

مصادر التباين		مجموع المربعات	درجات الحرية	متوسط المربعات	المكون العشوائي	نسبة تأثير التباين
الطلّاب	p	٤٨٨٨٦,١٦	٢٦	١٨٧,٩٢٩	٢٨,٩٢٧	%٧١,٩
طريقة التقدير	m	٢,٢٢٨	١	٢,٢٢٨	٠,٠١٩	٠
المهام	t	٤١٨,٧١٦	٢	٢٠٩,٣٥٨	٣,٦١٨	%٩
الطلّاب x طريقة التقدير	pm	٢٥,٩٣٨	٢٦	٠,٩٩٧	٠,٠٠٥	٠
الطلّاب x المهام	pt	٧٤٦,٢٨٣	٥٢	١٤,٣٥١	٦,٦٨٤	%١٦,٦
طريقة التقدير x المهام	mt	١,٢٣٤	٢	٠,٦١٧	٠,٠١	٠
البواقي	pmt,e	٥١,٠٩٨	٥٢	٠,٩٨٢	٠,٩٨٢	%٢,٤

ويتضح من جدول (١) أن قيم مكونات التباين للتأثير الرئيس لطريقة التقدير والتفاعلات بين المهام وطريقة التقدير مساوية للصفر؛ مما يدل على أن تقدير درجات الطلَّاب في المهام الأدائيَّة لم يتأثر بطريقة التقدير سواء كانت الطريقة الكليَّة أو التحليليَّة، التي تعطي مؤشرًا للصدق التقاربي.

وتبيَّن من نتائج التحليل الخصائص السيكومتريَّة لثنائي الوجه المتقاطع كليًّا (p x m x t)، أن قيم معاملات إمكانيَّة التعميم النسبية والمطلقة للتحقق من ثبات درجات تقدير الطلَّاب في مهام الأداء باستخدام طريقتَي التقدير مقبولة ويمكن الوثوق بها، حيث يمكن الاستفادة من النتائج من أجل تعميم استخدام هذه الأدوات على عيِّنة البحث؛ ويوضح جدول (٢) قيم معاملات إمكانيَّة التعميم النسبية والمطلقة للتصميم (p x m x t) للتحقق من ثبات تقدير درجات الطلَّاب في المهام الأدائيَّة.

جدول (٢):
قيم معاملات إمكانيَّة التعميم النسبية والمطلقة للتصميم (p x m x t) للتحقق من ثبات تقدير درجات الطلَّاب في المهام الأدائيَّة

نوع معامل إمكانيَّة التعميم	قيمة المعامل	قيمة معامل الخطأ المعياري	الهدف
معامل إمكانيَّة التعميم النسبي	٠,٩٢	١,٤٥	التعميم على طريقة التقدير والمهام الأدائيَّة
معامل إمكانيَّة التعميم المطلق	٠,٨٩	١,٩٠	

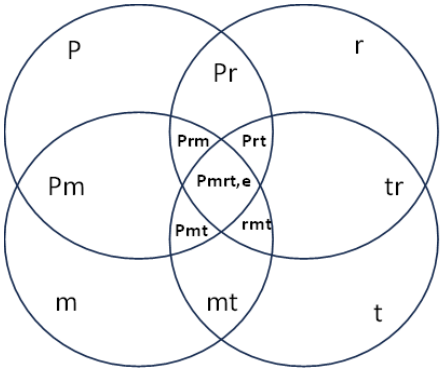
ويتضح من نتائج جدول (٢) نجد أن قيم معاملات إمكانيَّة التعميم النسبية والمطلقة أعلى من القيمة المقبولة التي يمكن الوثوق بها (٠,٨) وَفْق ما ذكرَهُ (Shavelson et all,1999)؛ مما يدل على أن النتائج ذات موثوقية وقابلة للتعميم على كلٍّ من طرق التقدير والمهام؛ حيث فسّرت النتائج ما نسبته (٧١,٩%) من مصادر التباين في الدرجات الشاملة لأداة البحث.

الأساليب الإحصائية:

تم الاعتماد على برنامج (EduG)؛ وذلك لتحليل بيانات وحساب التباينات لتصميم البحث والإجابة عن الأسئلة.

تصميم البحث:

تم الاعتماد على تصميم ثلاثي الوجه المتقاطع كليًّا المتضمِّن (الطلَّاب x المهام x طريقة التقدير x المُقيِّمين)، وتحليل بيانات البحث باستخدام نظريَّة إمكانيَّة، ويشتمل التصميم على الطلَّاب (P) Persons، والمهام الأدائيَّة (t) Tasks، والمقيِّمين (r) Raters، وطريقة التقدير (m)Methods؛ حيث تم اعتبار عيِّنة الطلَّاب والمهام وطريقة التقدير والمقيِّمين أوجهًا عشوائيَّة، ويشار إليها باختصار بـ (P x t x m x r).



رسم توضيحي للتصميم ثلاثي الوجه المتقاطع

كليًّا (p x t x m x r)

نتائج البحث ومناقشتها وتفسيرها:

نتائج السؤال الأول ومناقشتها وتفسيرها:

ما نسبة إسهام تبايُن الخطأ الذي يُفسِّره كلٌّ من أوجه المهام الأدائيَّة وطريقة التقدير والمقيِّمين والتفاعل بينها في التباين الكلي في ثبات درجات المهام الأدائيَّة في مقرَّر العلوم؟

للإجابة عن السؤال الأول، تم تقدير مصادر التباين الخطأ والتفاعل بين هذه المصادر ونِسب تأثيرها على الدرجة الشاملة في أداء الطلَّاب، ومعاملات إمكانيَّة التعميم النسبية والمطلقة؛ ويوضح جدول (٣) نتائج تقدير مكونات التباين تصميم ثلاثي الوجه المتقاطع كلِّيًا (P x t x m x r) بالاعتماد على درجات الطلَّاب في المهام الأدائيَّة المستمدة من طريقة التقدير الوصفي الكلي والتحليلي.

جدول (٣)
نتائج تقدير مكونات التباين تصميم ثلاثي الوجه المتقاطع كلِّيًا (الطلَّاب x المهام x طريقة التقدير x المُقيِّمين)

مصادر التباين		درجات الحرية	مجموع المربعات	متوسط المربعات	المكون العشوائي	نسبة تأثير التباين
الطلَّاب	P	١١٤	٣٣٣٦٦,٤١	٢٩٢,٦٩	٢١,٣٨	%٦٠,٨
طريقة التقدير	m	١	٠,٠٣	٠,٠٣	٠,٠	٠,٠
المقيِّمُون	r	١	٣٣,١٩	٣٣,١٩	٠,٠٥	%٠,١
المهام	t	٢	٢٠٧٧,٦٨	١٠٣٨,٨	٢,١٨	%٦,٢
الطلَّاب x طريقة التقدير	pm	١١٤	٥٥٠,٨١	٤,٨٣	٠,١٤	%٠,٤
الطلَّاب x المقيِّمُون	pr	١١٤	٨٣٠,٦٥	٧,٢٤	٠,٢٥	%٠,٧
الطلَّاب x المهام	pt	٢٢٨	٧٤٥٠,٣٢	٣٢,٦٨	٦,٩٤	%١٩,٨
طريقة التقدير x المقيِّمُون	mr	`	١,٤	١,٤	٠,٠٣	٠,٠
طريقة التقدير x المهام	mt	١	١٥,٩٨	٧,٩٩	٠,٠١	٠,٠
مصادر التباين		درجات الحرية	مجموع المربعات	متوسط المربعات	المكون العشوائي	نسبة تأثير التباين	مصادر التباين
المقيِّمُون x المهام	rt	١	١٨,٨٥	٩,٤٣	٠,٠٢	٠,٠
الطلَّاب x طريقة التقدير x المقيِّمُون	Pmr	١١٤	٤٢٥,٧٦	٣,٧٣	٠,٣٨	%١,١
الطلَّاب x طريقة التقدير x المهام	Pmt	٢٢٨	٦٥١,٦٩	٢,٨٦	٠,١٣	%٠,٤
الطلَّاب x المقيِّمُون x المهام	Prt	٢٢٨	١٠٦٠,٨١	٤,٦٥	٠,٠٣	%٢,٩
طريقة التقدير x المقيِّمُون x المهام	Mrt	٢	٢٢,١١	١١,٠٦	٠,٠٧	%٠,٢
البواقي	pmrt,e	٢٢٨	٥٩٣,٢٢	٢,٦	٢,٦	%٧,٤

ويتضح من الجدول (٣) أن أكبر نسبة للتباين تأثيرًا على معاملات إمكانيَّة التعميم في الدرجات الشاملة ترجع إلى موضوع القياس (الطلَّاب)، وأن التباين بين الطلَّاب في جودة وتعقيد الأداء في المنتجات المُقدمة جاء بنسبة (٦٠,٨%) من إجمالي التباين الكلي؛ حيث أشار كلٌّ من برينان وشافيلسون وويب (Brennan, 2001; Shavelson & Webb, 1991) إلى أنه من المتوقَّع أن تكون النسبة المئوية لموضوع القياس في التباين الكلي أكبر من مصادر التباين للأوجه في التصميم والتفاعلات بين هذه الأوجه وتباين البواقي. يليها نسبة تبايُن تفاعل (الطلَّاب x المهام) بنسبة (١٩,٨%) من إجمالي التباين الكلي، في حين جاء تأثير التباين الذي يعود إلى البواقي والذي يعكس التفاعل بين موضوع القياس (الطلَّاب) والمهام الأدائيَّة وطريقة

التقدير والمقيّمين، بالإضافة إلى الأخطاء العشوائيَّة غير المدرجة في التصميم، بنسبة بلغت (٧,٤%) من إجمالي التباين الكلي.

كما تم تحليل دراسة التعميم G–Study لتصميم البحث؛ بهدف حساب نِسَب تباين الخطأ النسبي والمطلق لكلّ مصدر من مصادر التباين؛ ويوضح جدول (٤) نتائج دراسة التعميم للتصميم ثلاثي الوجه المتقاطع كليًّا (P x t x m x r).

جدول (٤)
نتائج دراسة التعميم للتصميم ثلاثي الوجه المتقاطع كليًّا (P x t x m x r)

مصدر التباين	تباين الدرجة الشاملة	تباين الخطأ النسبي	النسبة المئوية	تباين الخطأ المطلق	النسبة المئوية
الطلّاب	٢١,٣٨	.....	.....	.....	.....
طريقة التقدير	.....	.....	.....	٠,٠	٠,٠
المقيّمُون	.....	.....	.....	٠,٠٢	%٠,٦
المهام	.....	.....	.....	٠,٧٣	%١٩,٣
طريقة التقدير xالطلّاب	.....	٠,٠٧	٢,٣	٠,٠٧	%١,٩
الطلّاب x المقيّمُون	.....	٠,١٣	٤,٢	٠,١٣	%٣,٣
الطلّاب x المهام	.....	٢,٣١	٧٦,٨	٢,٣١	%٦١,٤
المقيّمُونxطريقة التقدير	.....	.....	.....	٠,٠	٠,٠
المهامxطريقة التقدير	.....	.....	.....	٠,٠	٠,٠
المقيّمُون x المهام	.....	.....	.....	٠,٠	٠,٠
x طريقة التقدير xالطلّاب المقيّمُون	.....	٠,٠٩	٣,١	٠,٠٩	%٢,٥
x طريقة التقدير xالطلّاب المهام	.....	٠,٠٢	٠,٧	٠,٠٢	%٠,٦
المهامx المقيّمُون xالطلّاب	.....	٠,١٧	٥,٧	٠,١٧	%٤,٥
x المقيّمُون xطريقة التقدير المهام	.....	.....	.....	٠,٠١	%٠,٢
البواقي	.....	٠,٢٢	٧,٢	٠,٢٢	%٥,٨
مجموع التباينات	٢١,٣٨				
الانحراف المعياري	٤,٦٢				
معامل التعميم النسبي	٠,٨٨				
معامل التعميم المطلق	٠,٨٥				

ويوضح الجدول (٤) أن أكبر مكون لتباين الخطأ المنتظم في القياس النسبي والمطلق يرجع إلى تفاعُل الطلَّاب مع المهام بنسبة بلغت (٧٦,٨%– ٦١,٤%) من إجمالي نِسب التباين الكلي؛ مما يدلّ على أن أداء الطلَّاب يختلف حسَب نوع المهام المطلوب أداؤها نتيجةً لاختلاف المهام الأدائيَّة؛ ويمكن تفسير ذلك باختلاف أداء الطلَّاب للمهام الذي قد يعود إلى اختلاف خصائص مهام الأداء فيما بينها، والاختلاف في درجة

صعوبة المهام لجميع الطلَّاب. وتتفق هذه النتائج مع دراسة شافيلسون ورويز بريمو ووايلي (Shavelson, Ruiz-Primo & Wiley, 1999)، ودراسة ويب وشليكمان وشوجر (webb, Schlakman & Sugrue, 2000)، ودراسة بووير، بوجين ، ساندرز وفان (Bouwer, B´eguin, Sanders & Van, 2015)، ودراسة طباع (٢٠٢٠)، ودراسة العرايضة والشريفين (٢٠٢٢)؛ حيث تُعتبر من المصادر الرئيسة لتباين الخطأ المنتظم؛ إذ لُوحِظ في هذه الدراسات ظهور هذا التأثير للتفاعل بين الطلَّاب والمهام. ثم تأتي نسبة تباين الخطأ المطلق الراجع إلى المهام الأدائيَّة بنسبة (١٩,٣%)؛ مما يدلّ على مستوى صعوبة المهام بالنسبة للطلَّاب، ويُلاحَظ أن قيم نسب باقي مكونات التباين منخفضة. كما أن قيم معاملات إمكانيَّة التعميم النسبي والمطلق بلغت (٠,٨٨و ٠,٨٥) على التوالي، وهي قيم أعلى من الحدّ الأدنى للقيمة التي يمكن الوثوق بها (٠,٨) وَفْق ما ذكره (Shavelson et al., 1999).

نتائج السؤال الثاني ومناقشتها وتفسيرها:

ما الشروط الواجب توافرها لتحقيق أفضل مستويات عدد المهام الأدائيَّة وطريقة التقدير وعدد المقيِّمين على معاملات إمكانيَّة تعميم درجات أداء الطلَّاب في مقرَّر العلوم؟

ويكون ذلك من خلال زيادة عدد شروط طرق التقدير والمهام الأدائيَّة والمقيِّمين؛ بهدف خَفْض مصادر تباين الخطأ المنتظم، ورَفْع قيم معاملات إمكانيَّة تعميم، ومعرفة الأوجه الأكثر تأثيرًا في قيم هذه المعاملات؛ ويوضح الجدول (٥) دراسات القرارات للتصميم ثلاثي الوجه المتقاطع كليًّا (P x t x m x r) مُدرجةً بها قيم تباينات الخطأ النسبية والمطلقة، ومعاملات إمكانيَّة التعميم النسبية والمطلقة لكلّ دراسة قرار بناءً على دراسة إمكانيَّة التعميم الأولية.

جدول (٥):
دراسات القرارات للتصميم ثلاثي الوجه المتقاطع كليًّا (P x t x m x r)

طريقة التقدير	عدد المقيِّمين	عدد المهام	تباين الخطأ النسبي المتوقع	تباين الخطأ المطلق المتوقع	معامل إمكانيَّة التعميم النسبي	معامل إمكانيَّة التعميم المطلق
m	r	t	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	Φ
١	١	١	١١,٤٧	١٣,٧٧	٠,٦٥	٠,٦١
٣	١	١	٩,٣	١١,٥٦	٠,٧٨	٠,٧٥
١	٣	١	٨,٦٣	١٠,٨٥	٠,٧١	٠,٦٦
١	١	٣	٤,٣٣	٥,١٣	٠,٨٣	٠,٨١
١	٢	٢	٥,٩٠	٦,٠٣	٠,٨١	٠,٧٨
١	٢	٣	٣,٤٢	٤,١٨	٠,٨٦	٠,٨٤
٢	١	٢	٥,١٨	٦,٣٣	٠,٨١	٠,٧٧
٢	١	٣	٣,٦٢	٤,٤١	٠,٨٦	٠,٨٣
٢	٢	٢	٤,٣٧	٥,٥	٠,٨٣	٠,٨

يبيِّن جدول (٥) أثر زيادة مستويات أوجه التصميم على كلٍّ من تباين الخطأ المطلق والنسبي وقيم معاملات إمكانيَّة التعميم النسبية والمطلقة، من خلال زيادة عدد شروط الأوجه العشوائيَّة للمهام الأدائيَّة وطريقة التقدير وعدد المقيِّمين؛ وتشير النتائج إلى أنه لا يمكن الاعتماد على مهمة أدائيَّة واحدة تُقيَّم باستخدام طريقة تقدير واحدة يقوم بتقييمها مُقيِّم واحد؛ وذلك لانخفاض قيم معامل إمكانيَّة التعميم النسبي بقيمة (٠,٦٥) ومعامل إمكانيَّة التعميم المطلق بقيمة (٠,٦١)، وتتحقَّق قيم معاملات إمكانيَّة تعميم مقبولة في حالة الاعتماد على ثلاث مهام أدائيَّة أو أكثر تُقدّر بأيّ طريقة تقدير يُقيِّمها مُقيِّم واحد، أو الاعتماد على طريقتَيْ تقدير أو أكثر وعدد من المقيِّمين بالإضافة إلى مهمتين أدائيَّتَيْن أو أكثر.

وتشير النتائج إلى أنه في حالة زيادة عدد شروط أوجه المهام الأدائيَّة يعطي ذلك مؤشرًا يسهم بشكل أكبر في رَفْع قيم معاملات إمكانيَّة التعميم النسبية والمطلقة مقارنةً بزيادة طرق التقدير وعدد المقيِّمين؛ لذا يُفضَّل الاستعانة بعدد أكثر من المهام الأدائيَّة في مقرَّر العلوم.

كما تشير النتائج إلى أن قيم معاملات إمكانيَّة التعميم النسبية والمطلقة بوجه عام مرتفعة وأعلى من (٠,٨) في حالة الاعتماد على ثلاث مهام أدائيَّة أو أكثر، وأظهرت النتائج أيضًا أن تأثير عدد المقيِّمين في دراسات القرارات طفيف مقارنةً بتأثير عدد المهام الأدائيَّة، وهذا ما أشار إليه عديد من الدراسات؛ منها دراسة ويب وشليكمان وشوجر

(weeb, Schlakman & Sugrue.2000)، ودراسة شافيلسون ورويز بريمو ووايلي (Shavelson, Ruiz–Primo and Wiley,1999)، ودراسة كانكوي وأوزدر (Cankoy & Ozder, 2017)، ودراسة بووير، بوجين ، ساندرز وفان Bouwer, (B´eguin, Sanders & Van, 2015)، ودراسة طباع (٢٠٢٠)، ودراسة العرايضة والشريفين (٢٠٢٢)، ودراسة دراشى تاشيديلن وهوك وبايتيمي (Taşdelen, Şahin & Baytemir,2016). وللحصول على معاملات إمكانيَّة تعميم موثوقة بقيم أعلى من (٠,٨) يمكن الاعتماد على مُقيِّم واحد. وتتفق هذه النتيجة مع ما أشار إليه القحطاني (Alkatani, 2012) أنه يمكن الحصول على قيمة مرتفعة لمعاملات إمكانيَّة التعميم النسبيَّة والمطلقة بالاعتماد على مُقيِّم واحد في عمليَّة تقييم بالاستعانة بإحدى طرائق التقييم البديل التقييم الأداء الشفوي الكليَّة والتحليليَّة. وتتفق النتائج مع دراسة كلٍّ من كساب وفداء وروضة وحسان وأبو حجلة وأوكونور (Kassab, Fida, Radwan, A., Hassan, Abu–Hijleh & OConnor, 2016) في أنّ قيم معاملات إمكانيَّة التعميم ترتفع بزيادة عدد المقيِّمين بدلًا من زيادة أبعاد طريقة التقدير المستخدَمة.

توصيات البحث:

في ضوء ما تقدَّم من عرضٍ لنتائج البحث؛ يُوصَى بالتالي:

- الاهتمام بزيادة عدد المهام الأدائيَّة قَدْرَ الإمكان، وذلك في تقدير درجات الطلَّاب في مهام الأداء للحصول على معاملات إمكانيَّة تعميم تتمتَّع بموثوقية.
- أفضلية زيادة عدد المقيِّمين بدلًا من زيادة طريقة التقدير المستخدَمة، وذلك لتأثيرها في رَفْع قيم معاملات إمكانيَّة التعميم النسبية والمطلقة مقارنةً بزيادة طرق التقدير.

البحوث المقترحة:

يمكن اقتراح بعض البحوث في ضوء نتائج البحث الحالي، كما يلي:

- دراسة تطبيق نظريَّة إمكانيَّة التعميم في تحديد عدد التجارب المعمليَّة وعدد المقيِّمين الأمثل الذي يُقدّم معاملات إمكانيَّة تعميم يمكن الوثوق بها لتقدير أداء طلَّاب المرحلة الثانوية في نظام المقررات في التخصُّصات المختلفة.
- المقارنة بين معاملات إمكانيَّة التعميم لتقدير درجات أداءات الطلَّاب في المختبرات الافتراضية بالتطبيق العملي في قيم معاملات إمكانيَّة التعميم، باستعمال أداة التقييم المناسبة، التي تعتمد على الملاحظة المباشرة، مثل أداء التجارِب العمليَّة أو الأداءات الشفوية، وذلك من خلال إدراج تصميم يتضمَّن (الطلَّاب x التجارب x المقيِّمين).

المراجع:

أبو جراد، حمدي بونس (٢٠١٢). الخصائص السيكومترية لاختبارات مكتبة الاختيار من متعدد والتكميل دراسة مقارنة باستخدام نظرية الاستجابة للمفردة. *مجلة العلوم التربوية والنفسية*، ١٣ (٣)، ٣٧٦–٤٠٤.

الأحمدي، شرف حامد (٢٠١٨). صدق وثبات الطلاب والطالبات للتدريس باستخدام نظرية إمكانية التعميم. *مجلة جامعة طيبة للعلوم التربوية*، ٣ (١٣)، ٢٦٩–٢٨٦.

برهم، أريج عصام (٢٠١٨). أغراض استخدام التقويم الواقعي لدى معلمي الرياضيات في صفوف مرحلة التعليم الأساسي العليا في الأردن. ٤٥ (٤)، ٣١٥–٣٣٣.

البشير، أكرم عادل؛ برهم، أريج عصام (٢٠١٢). استخدام استراتيجيات التقويم البديل وأدواته في تقويم تعلم الرياضيات واللغة العربية في الأردن. *مجلة العلوم التربوية والنفسية*١ (١٣)، ٢٤١–٢٧٠.

حبيب، صفاء طارق؛ عزيز، صادق عبد النور (٢٠١٨). بناء المقاييس النفسية وفقاً لنظرية الاستجابة للفقرة باستعمال الانموذج الكشفي المعمم. عمان: الدار المنهجية للنشر والتوزيع.

الخالدي، عادي كريم عادي (٢٠١٤). درجة ممارسة معلمي العلوم الطبيعية بالمرحلة المتوسطة لمهارات التقويم البديل. *مجلة كلية التربية*، ٣٨ (٣)،٤١٥–٤٦٣.

الخليلي، خليل يوسف (١٩٩٨). التقييم الحقيقي في التربية. *مجلة التربية*، ١٢٦ (٢٧)، ١١٨–١٣٢.

الخروصي، حسين على؛ الذهلي، ربيع المر (٢٠٢٠). واقع التقويم البديل لدى معلمات الحلقة الأولى من التعليم الأساسي في سلطة عمان: الاستخدام والكفاءة الذاتية والاتجاه. *مجلة البحوث التربوية والنفسية*، ١٧ (٦٥)، ٢٨٦–٣٢٣.

خليل، محمد أبو الفتوح حامد (٢٠١١). التقويم التربوي بين الواقع والمأمول. المدنية المنورة: الشقري للنشر والتوزيع.

الدوسري، إبراهيم مبارك (٢٠٠١). إطار مرجعي للتقويم التربوي. مكتب التربية العربي لدول الخليج، دار الشروق للنشر والتوزيع: عمان.

رونالد، ليفينجستون؛ سيسيل، رينولدز. (٢٠١٣). إتقان القياس النفسي. *ترجمة: أ.د. صلاح الدين محمود علام*. عمان: دار الفكر للنشر والتوزيع.

الزبيدي، محمد حسن يحيى (٢٠١٩). واقع التقويم البديل في التعليم العام في منطقة مكة المكرمة من وجهة نظر معلمي التعليم العام. *المجلة التربوية*، ٦٦ (٦٦)، ٢٤٥–٢٨٤.

زيتون، حسن حسين (٢٠٠٧). أصول التقويم والقياس التربوي: المفهومات والتطبيقات. القاهرة: الدار الصوليته للتربية.

طباع، فاروق (٢٠٢٠). استخدام نظرية إمكانية التعميم في تقدير ثبات اختبار تقييم كفاءة الرياضيات لدى طلاب السنة الرابعة ابتدائي. *المجلة الأردنية في العلوم التربوية*، ١٦ (١)، ١–١٨.

الطريري، عبد الرحمن سليمان (٢٠١٤). القياس النفسي والتربوي نظريته– اسسه –تطبيقاته. الرياض: مكتبة الرشد.

العرايضة، محمد أكرم؛ الشريفين، نضال كمال (٢٠٢٢). الكشف عن إسهام مصادر التباين المتعددة في ثبات اختبار الرياضيات باستخدام نظرية إمكانية التعميم، *مجلة العلوم التربوية*، ٣٤ (٢)، ١٩٣–٢١٨.

علام، صلاح الدين محمود (٢٠٠٢). القياس والتقويم التربوي والنفسي أساسياته وتطبيقاته وتوجهاته المعاصرة. القاهرة: دار الفكر العربي.

علام، صلاح الدين محمود (٢٠٠٩). التقويم التربوي البديل: أسسه النظرية والمنهجية وتطبيقاته الميدانية. القاهرة: دار الفكر العربي.

علام، صلاح الدين محمود (٢٠١٦). القياس والتقويم التربوي في العملية التدريسية. القاهرة: دار الفكر العربي.

فرج، صفوت (٢٠١٧). القياس النفسي. ط٦. مكتبة الانجلو المصرية: القاهرة.

الفهيدي، هذال عبيد عياد (٢٠١٨). تقويم واقع مشكلات تعليم العلوم بالمرحلة الابتدائية في محافظة شرورة من وجهة نظر معلمي العلوم. *مجلة جامعة أم القرى للعلوم التربوية والنفسية*، ٩ (١)، ٢٤١–٢٧٤.

المطرفي، ثامر هزاع عواد (٢٠١٥). مدى أهمية واستخدام أساليب التقويم البديل لدى معلمي العلوم الشرعية بالمرحلة الثانوية وأهم الصعوبات التي يوجهونها. *مجلة كلية التربية*، ٣١ (٣)، ٥٤٤–٦٠١.

وزارة التعليم (٢٠١٣). التقويم الأدائي في دروس العلوم الصف الثاني متوسط. العبيكان: الرياض.

Alkahtani, s.f (2012).Oreal Performance Scoring Using Generalizability theory and Many-Facet Rasch measurement: A Comparison Study. PhD, The Pennsylvania State University.

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

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, 1–21. Psychology, 16, 137–163.

Burke, K. (2009). How to Assess Authentic Learning. 5th ed. ISBN 978-1-4129-6278-0 (celothi ISBN 978-1-4129-6279-7.

Burke, K. (2011). From Standards Rubrics in Six Steps: Tools for Assessing Student Learning. Third Edition SAGA Publishing. ISBN 978-4129-8701-1.

Bouwer, R., B´eguin, A., Sanders, T., & Van, D., H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32 (1), 83–100.

Bottema-Beutel, K., Lloyd, B., Carter, E., & Asmus, J. (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. M J Intellect Dev Disabil ;119(6):589-605.

Cankoy, O. & Ozder, H. (2017). Generalizability Theory Research on Developing a Scoring Rubric to Assess Primary School Students Problem Posing Skills. *EURASIA Journal of Mathematics Science and Technology Education*. ISSN: 1305-8223 (online).13(6):2423-2439.

Dunbar, S. B., Kortez, D.M. & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-304.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, 7(4), 323-342.

Graham, S.; Hebert, Ml; Sandbank, M; and Harris, K. (2016), "Assessing the Writing Achievement of Young Struggling Writers: Application of Generalizability Theory" Special Education and Communication Disorders Faculty Publications. 118

Han, T. & Huang, J. (2017) Examining the Impact of Scoring Methods on the Institutional EFL Writing Assessment: A Turkish Perspective. PASAA. 53 (112-147).

Hammond, L. D. (2014). Next Generation Assessment: Moving Beyond the Bubble Test to Support 21st Century Learning. John Wiley & Sons, Incorporated. EBOOK PUB DATE.

Kassab, S. E., Fida, M., Radwan, A., Hassan, A. B., Abu-Hijleh, M., & OConnor, B. P. (2016). Generalizability theory analyses of concept mapping assessment scores in a problem-based medical curriculum. *Medical Education*, 50(7), 730–737.

Popham, W. J. (2003). Test Better, Teach Better: The Instructional Role of Assessment, Association for Supervision Curriculum Development. ProQuest eBook Central.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, 30, 41–53. doi:10.1111/jedm.1993.30.

Scheuneman, J. D. (1982). A new look at bias in aptitude tests. *New Directions for Testing & Measurement*, 12, 3–35.

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

Shavelson, R., Ruiz-Primo, M. & Wiley, E. (1999). Note on Sources of Sampling Variability in Science Performance Assessments. *Wiley Journal of Educational Measurement*. 36 pp. 61-71

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research*, 2, pp. 538–543. New York: Macmillan.

Suen, H. K., Lu, C., Neisworth, J. T., & Bagnato, S. J. (1993). Measurement of team decision-making through generalizability theory. *Journal of Psychoeducational Assessment*, 11, 120–132.

Taşdelen T., G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment applications. *Journal of Human Sciences*, 13(3), 5574-5586.

webb. N. M., Schlakman. J. & Sugrue.B. (2000). The Dependability and Interchangeability of Assessment Methods in Science. *Applied Measurement in Education*. 13(3), 277-301. Published online: 07 Jun 2010

Webb, N. M., Shavelson, R. J. (2005). Generalizability Theory: Overview. *Encyclopedia of Statistics in Behavioral Science*, V2. pp. 717-719. ISBN-13: 978-0-470-86080-9 ISBN-10: 0-470-86080-4.

Wobiele, O. & Queensoap, M. (2020). Application Of Generalizability Theory in Reliability Assessment of Chemistry Essay Achievement Test. *EPRA International Journal of Research & Development (IJRD)*, 5(4).

Yin, Y. & Shavelson R. J. (2004). Application of generalizability theory to concept map assessment research. *Applied Measurement Education*, (3),273–9.