أثر طريقة التوفيق بين تقديرات المقيّمين للمهمات الكتابية وعدد فئات دليل التصحيح في الدرجة الإجرائية

عبدالحافظ قاسم الشايب*

جامعة الباحة، السعودية

استلم بتاريخ: ٢٠١٢١١٢٢ **عدل بتاريخ: ٢٠١٣١٢١٣** **قُبل بتاريخ: ٢٠١٣١٣١٢**

هدفت الدراسة إلى التعرّف على مؤشرات ثبات الدرجة الإجرائية المحسوبة بثلاث طرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية هي: المتوسط الحسابي للتقديرين الأصيلين، المتوسط الحسابي لتقدير المقيّم الخبير والتقديرين الأصيلين، المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين لدى استخدام دليلين للتصحيح (خماسي التدريج، سباعي التدريج) لتقييم المهمة الكتابية نفسها، بالإضافة إلى الكشف عن أثر كل من عاملي طريقة التوفيق بين تقديرات المقيّمين لحساب الدرجة الإجرائية، وعدد فئات دليل التصحيح في الدرجة الإجرائية. شملت بيانات الدراسة إجابات ٢٣٢ معلماً ومعلّمة لمهمة كتابية معدّة مسبقا، حيث قام ستة مقيّمين بتقييم الإجابات باستخدام دليلي التصحيح المعدّين مسبقاً. وكشفت النتائج عن أن قيم مؤشرات الثبات تختلف باختلاف الطريقة المُستخدمة لحساب الدرجة الإجرائية بصرف النظر عن دليل التصحيح المُستخدم. وكشفت نتائج تحليل التباين ذو القياسات المتكرّرة عن وجود أثر رئيس لكل من عاملي طريقة التوفيق ودليل التصحيح في الدرجة الإجرائية المحسوبة بإحدى طرق التوفيق المُستخدمة. ١٤٦ كلمة.

الكلمات المفتاحية: طريقة التوفيق، المهمات الكتابية، دليل التصحيح، الدرجة الإجرائية.

The Effect of Score Resolution Method among Evaluators' Ratings of Writing Tasks and Number of Scoring Rubric Categories on the Operational Score

Abdelhafez Q. Al-Shayeb*
Al-Baha University, Kingdom of Saudi Arabia

The study aimed at investigating the reliability indices of the operational score calculated by three different methods of score resolution among evaluators' ratings of a writing task (the averaged score of the original ratings, the averaged score of the expert rating and the original ratings, the averaged score of the expert rating and the closest rating of the original ratings). Two scoring rubrics (five categories, seven categories) were used. The effect of score resolution method was examined, and scoring rubric in the operational score calculated in one of the above mentioned methods. Data were obtained from the answers of 232 male and female teachers to a previously developed writing task. The writings were blindly assessed by six raters using the two pre-developed scoring rubrics. The results revealed differences among reliability indices due to the resolution method used to calculate the operational score regardless of the scoring rubric being used. Repeated measures ANOVA with between-subjects factor revealed significant main effect of both factors i.e., score resolution method, and scoring rubric in the calculated operational score using one of the aforementioned resolution methods.

Keywords: resolution method, writing tasks, scoring rubric, operational score.

*alshayeb99@yahoo.com

شهدت بدايات الألفية الثالثة ثورة في مفهوم التقويم واستراتيجياته وأدواته، وانصب الاهتمام على عمليات التفكير العليا كبلورة الأحكام واتخاذ القرارات، وحل المشكلات باعتبارها مهارات عقلية عليا تمكّن المتعلّم من التعامل مع معطيات عصر المعلوماتية وتفجّر المعرفة والتقنية متسارعة التطور. ولجأ المشتغلون بالقياس والتقويم إلى ما يُسمى بالتقويم الواقعي Authentic Assessment، وهو التقويم الذي يعكس إنجازات المتعلمين في مواقف حقيقية، ويجعلهم ينغمسون في مهمات ذات معنى بالنسبة لهم حتى تبدو كنشاطات تعلّم وليست اختبارات بالمعنى التقليدي، ويُمارسون مهارات التفكير العليا، ويُوائمون بين طيف واسع من المعارف لبلورة الأحكام واتخاذ القرارات، أو لحل المشكلات الحياتية التي يواجهونها، ويُطوّرون القدرة على التفكير التأملي الذي يساعدهم في معالجة المعلومات ونقدها وتحليلها (علام، ٢٠٠٩). وفي الآونة الأخيرة، تزايد الاهتمام باستراتيجية تقييم الأداء Performance Assessment التي تُعدّ واحدة من بين الأشكال المختلفة لاستراتيجيات التقويم الجديدة أو ما يُسمى بالتقويم البديل. وتُعدّ المهمات الكتابية أو أسئلة المقال واحداً من بين الأشكال المختلفة الأكثر شيوعاً التي ينطوي عليها تقييم الأداء (Linn, 1993; Johnson, Penny, & Gordon, 2000).

ومع تزايد الاهتمام باستراتيجية تقييم الأداء، أدرك المربون وخبراء القياس والتقويم في الولايات المتّحدة الأمريكية جوانب قصور الأسئلة الموضوعية السائدة منذ عقود طويلة في مؤسساتها التعليمية بمختلف مستوياتها، ونادوا بإحداث تغييرات جذرية في استراتيجيات التقويم بحيث تستند إلى الأداء (علام، ٢٠٠٩). واستجابة لهذه المناداة، أدرك القائمون على البرامج الاختبارية في عدد كبير من هذه الولايات ضرورة اشتمال البرامج الاختبارية على مهمات اختبارية من النوع الكتابي بجانب الأسئلة الموضوعية للتّمكن من تقييم مهارات التفكير العليا عند الطلبة في المراحل التعليمية المختلفة (Olson, Bond, & Andrews, 1999). ويرى البعض أن اشتمال البرامج الاختبارية على مهمات كتابية أصبح حاجةً ملحّة لأن الأسئلة الموضوعية لا توفّر مؤشرات دقيقة عن طبيعة الأداء مقارنة بالمهمات الكتابية (Johnson et al., 2000). ويؤكّد بيرلمان Perlman (2003) على ذلك في إشارته إلى مزايا اختبارات المقال بالقول بأن نتاجات التعلّم المعقّدة كمهارات التفكير الناقد، والاتصال، وحل المشكلات لا تُسلم نفسها بشكل جيّد للصور الاختبارية المألوفة التي تتطلب من المفحوص اختيار أحد بدائل الإجابة التي يوفّرها الفاحص كأسئلة الاختيار من متعدّد أو أسئلة الصواب والخطأ، مما يستدعي اللجوء إلى استخدام اختبارات المقال.

ومع أن لهذا الرأي ما يبرّره، إلا أن مسألة ذاتية تصحيح المهمات الكتابية كانت وما زالت تؤرق المشتغلين بالقياس والتقويم، لأنها تُعدّ مصدراً أساسياً من مصادر أخطاء القياس التي تؤثر بشكل مباشر في خاصيّة الثبات، الأمر الذي ينعكس سلباً على دقة القرارات المستندة لعملية التقييم وصدقها، ويزداد الأمر خطورة عندما تمسّ القرارات المترتّبة على نتائج الاختبارات مستقبل الأفراد بشكل مباشر، وتغيّر من مسار حياتهم، فعلى سبيل المثال، يعتمد قرار النجاح في نهاية المرحلة الثانوية في عدد من الولايات المتّحدة الأمريكية جزئياً على أدائه على اختبارات المقال (National Education Goals Panel, 1996). ويعتمد تحديد مستوى الطالب في أحد مستويات اللغة الانجليزية على نوعية أدائه الكتابي (Weigle, 1999; Smith, 1993; Haswell, 1998; Hayes, Hatch, & Silk, 2000; Willard-Traub, Decker, Reed, & Johnston, 2000). وتولي بعض اللجان والمجالس في الولايات المتّحدة الأمريكية كالمجلس القومي لمعايير مهنة التدريس National Board for Professional Teaching Standards، والمجلس القومي للفاحصين في ميدان الطب National Board of Medical Examiners أهمية خاصة للمهمات الكتابية في برامجها الاختبارية، حيث يتقرّر في ضوئها قرار منح المفحوص لرخصة مزاولة المهنة (Margolis & Ross, 1995; National Board for Professional Teaching Standards, 1993).

وفي إطار السعي للتغلّب على مشكلة ذاتية تصحيح المهمات الكتابية، وتحسين ثبات تقديرات المقيمين Inter-rater Reliability للمهمات الكتابية، وزيادة دقة القرارات المترتّبة على نتائج التقييم، فقد جرت العادة أن يقوم مقيّمين اثنين على الأقل بتقييم إجابات الأفراد للمهمات الكتابية (Cherry & Meyer, 1993; Breland, 1983). وقد تبيّن في مسح أجري في الولايات المتحدة الأمريكية أن (٧٨.٤% من دوائر التربية الأمريكية تستخدم هذه الآلية في برامجها الاختبارية (Johnson, Penny, & Johnson, 1998).

ومع أن زيادة الثبات بين المقيّمين هو بلا شك أمر أساسي، إلا أن ارتفاع مؤشر الثبات لا ينفي وجود اختلاف بين التقديرات التي يمنحها المقيّمون للمهمات الكتابية، مما يُشكّل مأزقاً لمتخذي القرار بشأن القرار النهائي المتعلّق بالدرجة الإجرائية Operational Score كما يُشار لها في أدبيات القياس والتقويم (Johnson et al., 2000). ويُشير الأدب النظري في هذا السياق إلى أنه مهما بلغت دقة دليل التصحيح Scoring Rubric، ومستوى مهارة المقيّمين، وجودة البرامج التدريبية، وصرامة إجراءات

المراقبة، تبقى مسألة وجود فروق بين تقديرات المقيّمين للمهمات الكتابية أمراً محتوماً لأنها تخضع في النهاية لأحكام شخصية (Johnson, Penny, Fisher, & Kuhs, 2003). ويعود اختلاف التقديرات إلى عدة عوامل مثل أثر الحمل من مفحوص لآخر examinee-to-examinee carryover كأن يتأثر تقدير إجابة أحد المفحوصين سلباً بعد قراءة المقيّم لعدد من الإجابات الجيدة لعدد من المفحوصين قبله مباشرة، أو يتأثر تقديره إيجاباً بعد قراءة المقيّم لعدد من الإجابات السيئة لعدد من المفحوصين قبله مباشرة (Hopkins, 1998). أو نتيجة لأثر الحمل من فقرة لأخرى item-to-item carryover كأن يتأثر تقدير إجابة المفحوص على سؤال مقالي معيّن سلباً أو إيجاباً نتيجة تقدير إجابته على فقرة سابقة (Bracht, 1967, as cited in Hopkins, 1998)، أو نتيجة لما يُسمى بأثر الهالة التلازمي Concurrent Hallo Effect أو الانطباع الملازم لعملية التصحيح والذي يرتبط بأمور شكلية كالخط والترتيب والإملاء...الخ (عودة، ٢٠٠٤)، أو نتيجة لطول المهمة الكتابية (Lumley, 2002)، أو تشدّد المقيّمين الناتج عن اندافعهم وحماسهم (Weigle, 1998)، أو نتيجة لعامل التعب أو الملل الناتج عن طول الفترة الزمنية لجلسة التقييم (Penny, 2003).

من هنا تبرز مسألة البحث عن طريقة مناسبة للتوفيق بين التقديرات المختلفة، والتوصّل إلى درجة إجرائية بحيث تعكس أداء الفرد بدرجة مقبولة من العدالة والموضوعية. ويزداد الأمر إلحاحاً عندما تكون القرارات المترتّبة على درجة الفرد ذات طبيعة حسّاسة، كتلك المتّصلة بقرار تخرّج الطالب، أو قبوله في برنامج ما للدراسات العليا، أو منح الفرد رخصة مزاولة المهنة (Johnson et al., 2003).

وفي هذا السياق، يلجأ القائمون على معظم البرامج الاختبارية إلى الاعتماد على قيمة المتوسط الحسابي لتقدير المقيّمين الاثنين إذا كان التقديران متجاورين كأن يكون تقدير المقيّم الأول ٣، وتقدير المقيّم الثاني ٤، مثلاً (Johnson, et al., 1998). أما في الحالات التي لا يكون فيها التقديران متجاوران، فقد جرت العادة أن يتم اللجوء إلى حكم قضائي Adjudication من خلال تكليف مقيّم خبير بتقييم المهمة بشكل مستقل، ثم حساب المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين؛ أو حساب المتوسط الحسابي للتقديرات الثلاثة؛ أو الاعتماد على تقدير المقيّم الخبير وإهمال التقديرين الأصيلين (Brennan, 1996; Livingston, 1998). وجدير بالذكر أن المشتغلين في هذا المجال، يختلفون حول تعريف المقيّم الخبير؛ ففي حين يرى البعض بأن المقيّم الخبير هو الأكثر خبرة مقارنة بالمقيّمين الأصيلين (Livingston, 1998)، لا يكتفي البعض الآخر بعامل الخبرة، بل يضيفوا إلى ذلك عوامل أخرى كالخضوع لبرنامج تدريبي مكثّف في موضوع التقييم، والألفة بقدرات المفحوصين، والتمتع باحترام الزملاء، وامتلاك درجة جيدة من مهارات الاتصال (Hieronymous, Hoover, Cantor, & Oberley, 1987).

وفي سياق المقارنة بين الطرق السابقة، يعتقد بريلاند (Breland, 1983) أن الطريقة التي تقوم على حساب المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب له من التقديرين الأصيلين تُفضي إلى معامل ثبات أكبر مقارنة بالطرق الأخرى إذا كان تباين تقديرات المقيّم الخبير قليلاً. ويرى جونسون ورفاقه (Johnson, Penny & Gordon, 2001) أن الطريقة التي تعتمد على تقدير المقيّم الخبير وإهمال التقديرين الأصيلين هي الطريقة الفضلى مقارنة بالطرق الأخرى انطلاقاً من كون الخبير خبيراً. ويعتقد فريق ثالث أن الطريقة التي تقوم على حساب المتوسط الحسابي للتقديرات الثلاثة (تقدير المقيّم الخبير والتقديرين الأصيلين) هي الطريقة الفضلى بين الطرق المختلفة استناداً إلى أن معامل الثبات يزداد بزيادة عدد المقيّمين من ناحية نظرية كما تُشير نظرية القياس إذا لم يكن هناك تقديرات متطرّفة بشكل واضح (Mehrens & Lehmann, 1991; Shevelson & Webb, 1991).

ومع أن نتائج بعض الدراسات توصّلت إلى عدم وجود فروق بين بعض طرق التوفيق بين تقديرات المقيّمين، أشارت نتائج معظم الدراسات إلى أفضلية الطريقة التي تعتمد على حساب المتوسط الحسابي للتقديرات الثلاثة مقارنة بالطرق الأخرى. ففي دراسة جونسون ورفاقه Johnson et al. (2000) التي تناولت العلاقة بين أربع طرق مختلفة للتوفيق بين تقديرات المقيّمين (المتوسط الحسابي للتقديرين الأصيلين، تقدير المقيّم الخبير، المتوسط الحسابي للتقديرات الثلاثة، المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب له من بين التقديرين الأصيلين) والثبات بين المقيّمين لدى تقييم إجابات ١٢٠ مفحوصاً في الصف الحادي عشر في مدرسة جورجيا الثانوية على اختبار المدرسة الكتابي المؤهل للتخرّج، واستخدمت دليل تصحيح تحليلي (Analytical Scoring Rubric)، ووظّفت ثلاثة مؤشرات لمعامل الثبات بين المقيّمين هي نسبة الاتفاق، ومعامل ارتباط الرتب (سبيرمان)، ومعامل الاستقلالية المعبّر عنه بمعامل فاي (Phi Coefficient) استناداً إلى استخدام نظرية التعميم (G-Theory)، تبيّن أن الطريقة التي تقوم على حساب المتوسط الحسابي للتقديرات الثلاثة تُفضي إلى معاملات ثبات أكبر مقارنة بالطرق الأخرى. وفي دراسة أخرى استهدفت اختبار معاملات الثبات والصدق للدرجات الإجرائية المحسوبة بثلاث طرق مختلفة (المتوسط الحسابي

للتقديرين الأصيلين، المتوسط الحسابي للتقديرات الثلاثة، المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب له من التقديرين الأصيلين)، واستخدم فيها اختبار نهاية المرحلة الثانوية في إحدى الولايات الأمريكية (Johnson et al., 2003)، توصّل الباحثون إلى النتيجة السابقة نفسها.

في المقابل، لم يتبيّن وجود أثر للطريقة التي تقوم على حساب متوسط التقديرين الممنوحين من قِبل مقيّمين اثنين في زيادة دقة الدرجة الإجرائية مقارنة بطريقة المناقشة بين المقيّمين، علماً بأن طريقة المناقشة قد توفّر درجة أكبر من الدقة في بعض الحالات التي تنطوي على قرارات ذات مستوى عالٍ من الأهمية كقرار قبول أو رفض طلب الهجرة مثلاً (Johnson, Penny, Gordon, Shumate & Fisher, 2005)، ولدى توظيفهما لنموذج راش (Rasch Model) أحادي المعلمة، أحد نماذج نظرية الاستجابة للفقرة (Item Response Theory, IRT)، أشارت نتائج دراسة مايفورد وولف (Myford & Wolfe, 2002) التي وظّفت برمجية (FACETS) إلى أن الدرجات التي كشف التحليل عن تطرّفها قبل إجراء عملية توفيقية بينها لم تُظهر تطرفاً بعد تعديلها باستخدام أي طريقة من الطرق المختلفة.

يتبيّن من خلال مراجعة الأدب السابق أن مسألة المقارنة بين الطرق المختلفة للتوفيق بين تقديرات المقيّمين للأداء الكتابي لم تحسم الجدل لصالح طريقة معيّنة من هذه الطرق، مع أنها تميل إلى ترجيح الطريقة التي تقوم على حساب متوسط التقديرات الثلاثة (الدرجتين الأصيلتين ودرجة المقيّم الخبير). وبما أن طرق التوفيق بين تقديرات المقيّمين تُستخدم في سياقات وظروف مختلفة، يُعتقد بأن مسالة المفاضلة بين الطرق المختلفة هي مسألة نسبية تتوقف على عوامل أخرى تؤثّر في العلاقة بين الطريقة المستخدمة والدرجة الإجرائية، كعدد فئات دليل التصحيح المُستخدم، أو نوعه (كلّي، تحليلي)، أو نوع المهمة الكتابية (مقالية مقيّدة، مقالية حرّة، بورتفوليو...إلخ)، أو عدد المهمات، أو مستوى صعوبتها، أو نوع القرار المترتّب على الدرجة الإجرائية (مطلق، نسبي)، أو الإطار المُستخدم لتحليل البيانات (نظرية القياس التقليدية، نظرية الاستجابة للفقرة IRT). لهذا، تأتي الدراسة الحالية كخطوة ضمن الجهود البحثية الرامية إلى تعميق الفهم للعوامل المؤثرة في العلاقة بين الطرق المختلفة للتوفيق بين تقديرات المقيّمين للمهمات الكتابية من جهة، والدرجة الإجرائية من جهة ثانية. وتبرز أهمية الدراسة من تناولها لعامل عدد فئات دليل التصحيح كعامل يمكن أن يؤثر في الدرجة الإجرائية بجانب عامل طريقة التوفيق بين تقديرات المقيّمين للمهمات الكتابية، والذي لم يسبق أن تناوله الباحثون من قبل. ويُؤمل أن تُسهم نتائج هذه الدراسة في تقدّم المعرفة النظرية في هذا الميدان، وأن تفتح الباب أمام الباحثين لإجراء المزيد من الدراسات التي تتناول الجوانب المختلفة للمسألة. من ناحية ثانية، ربما تُسهم نتائج الدراسة في تقديم بعض المؤشرات والأدلة الإمبريقية للمشتغلين في ميدان التقييم حول الطرق المختلفة للتوفيق بين تقديرات المقيمين وأثرها في الدرجة الإجرائية، وبخاصة أننا نشهد اليوم اهتماماً غير مسبوق بموضوع تقييم الأداء في الوطن العربي بشكل عام، وفي الأردن على وجه التحديد بعدما اتجهت وزارة التربية والتعليم في الآونة الأخيرة إلى استبدال استراتيجيات التقويم التقليدية باستراتيجيات التقويم البديل التي تركّز على المهارات العقلية العليا، والتي يستلزم تقييمها استخدام أحد أشكال تقييم الأداء كالمهمات الكتابية.

مشكلة الدراسة وأسئلتها:

يلجأ المشتغلون في مجال تقييم الأداء بشكل عام، وتقييم المهمات الكتابية بشكل خاص إلى تكليف مقيّمين اثنين على الأقل بتقييم الأداء، إضافة إلى استخدام إجراءات ضابطة كإعداد دليل للتصحيح يوضّح كل مستوى من مستويات الأداء، وعقد ورش عمل لتدريب المقيّمين على إجراءات التقييم بهدف التغلّب على مشكلة ذاتية التصحيح المقترنة بمفهوم تقييم الأداء، وتحسين ثبات التقديرات، وزيادة دقة القرارات المترتّبة عليها. وبشكل عام، تمرّ عملية تقييم الأداء في مرحلتين: يقوم في المرحلة الأولى مقيّمين اثنين بتقييم المهمة نفسها بشكل مستقل، ومن ثم تُقارن تقديرات المقيّمين، فإذا تطابق التقديران تنتهي عملية التقييم، ويكون التقدير المتّفق عليه هو التقدير النهائي لمستوى الأداء (الدرجة الإجرائية). أما إذا اختلف المقيّمان في تقديرهما للأداء، فإما أن يُحسب المتوسط الحسابي للتقديرين ليمثّل الدرجة الإجرائية، أو يتم الانتقال إلى المرحلة الثانية أو المرحلة القضائية، حيث يتم تكليف مقيّم ثالث (مقيّم خبير) بتقييم المهمة بشكل مستقل عن المقيّمين الأصيلين، ثم تُحسب الدرجة الإجرائية بطريقة معيّنة من الطرق المختلفة للتوفيق بين التقديرات، كأن يُحسب المتوسط الحسابي للتقديرات الثلاثة، أو المتوسط الحسابي لتقدير الخبير والتقدير الأقرب لتقديره من التقديرين الأصيلين، أو يحل تقدير الخبير مكان التقديرين الأصيلين. من هنا تبرز مشكلة المفاضلة بين الطرق السابقة للتوفيق بين التقديرات في حال اختلف المقيّمين الأصيلين لأن أثر هذه الطرق في الدرجة الإجرائية قد يختلف من طريقة لأخرى. لكن في المقابل، ليست الطريقة هي العامل الوحيد الذي يؤثر في الدرجة الإجرائية، وإنما هناك عوامل أخرى يمكن أن تؤثر فيها كنوع دليل التصحيح وعدد فئاته، ونوع المهمة ودرجة صعوبتها، ونوع القرار. وهذا يعني أن أثر طريقة التوفيق

بين التقديرات المختلفة في الدرجة الإجرائية قد يتوقف على عوامل أخرى؛ فقد تكون الطريقة التي تقوم على حساب المتوسط الحسابي للتقديرات الثلاثة هي الأفضل مقارنة بالطرق الأخرى في حال كان دليل التصحيح تحليلياً وليس كلياً، وكان عدد فئات الدليل هو خمسة مثلاً، وكانت المهمة الكتابية من النوع المقيّد، ونوع القرار نسبياً وليس مطلقاً على سبيل المثال. من هنا تبلورت مشكلة الدراسة التي تتلخص في التعرّف على مؤشرات ثبات الدرجة الإجرائية المحسوبة بثلاث طرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية (المتوسط الحسابي للتقديرين الأصيلين، المتوسط الحسابي لتقدير المقيّم الخبير والتقديرين الأصيلين، المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين) لدى استخدام دليلين للتصحيح (خماسي التدريج، سباعي التدريج)، بالإضافة إلى الكشف عن أثر كل من عاملي طريقة التوفيق بين تقديرات المقيّمين لحساب الدرجة الإجرائية، وعدد فئات دليل التصحيح في الدرجة الإجرائية، وبالتحديد سعت الدراسة للإجابة عن الأسئلة الآتية

ما مؤشر ثبات الدرجة الإجرائية المحسوبة بثلاث طرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية نفسها باستخدام كل من دليلي التصحيح خماسي التدريج وسباعي التدريج؟

هل تختلف الدرجة الاجرائية المحسوبة بطرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية نفسها تبعاً لطريقة التوفيق المُستخدمة بدرجة دالة إحصائياً عند مستوى (٠,٠٥)؟

هل تختلف الدرجة الاجرائية المحسوبة بطرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية نفسها تبعاً لدليل التصحيح المُستخدم بدرجة دالة إحصائياً عند مستوى (٠,٠٥)؟

هل تختلف الدرجة الاجرائية المحسوبة بطرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية نفسها نتيجة للتفاعل بين عاملي طريقة التوفيق ودليل التصحيح بدرجة دالة إحصائياً عند مستوى (٠,٠٥)؟

التعريفات الإجرائية:

ورد في الدراسة الحالية عدد من المصطلحات التي تحتاج إلى تعريفات بصورة إجرائية على النحو الآتي:

طريقة التوفيق بين تقديرات المقيمين: ويُقصد بها الطريقة الإحصائية المُستخدمة للتوصّل إلى تقدير توفيقي واحد يعكس تقديري المقيّمين الأصيلين إذا كان التقديران مختلفين وبصرف النظر عن مقدار الاختلاف، وهي ثلاث طرق في سياق الدراسة الحالية: الطريقة الأولى وتقوم على حساب المتوسط الحسابي لتقديري المقيّمين الأصيلين الأول والثاني، والطريقة الثانية وتقوم على حساب المتوسط الحسابي لتقديرات المقيّمين الأصيلين والمقيّم الخبير، والطريقة الثالثة وتقوم على حساب المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين الأول والثاني.

الاختلاف بين المقيّمين: أي فرق مطلق بين التقديرات التي يمنحها المقيّمون المختلفون للمهمة الكتابية المستخدمة في هذه الدراسة بصرف النظر عن قيمته.

المهمة الكتابية: هي سؤال مقالي من نوع الإجابة المقيّدة، يدور حول فلسفة المعلّم الشخصية في العملية التعليمية.

الدرجة الإجرائية: هي التقدير التوفيقي المعبّر عنه كمياً من خلال استخدام إحدى طرق التوفيق الثلاث المستخدمة في هذه الدراسة.

المقيّم الخبير: هو أحد المشرفين التربويين الذين تولوا عملية تقييم إجابات المعلمين عن المهمة الكتابية المستخدمة في الدراسة ممن ينطبق عليهم شرط العمل في التدريس لمدة لا تقل عن عشر سنوات، بالإضافة إلى خبرة لا تقل عن خمس سنوات في مجال الإشراف التربوي.

دليل التصحيح: هو الدليل الذي يستخدمه المقيّم لتقييم الإجابة عن المهمة الكتابية المستخدمة في الدراسة باستخدام الطريقة الكليّة Holistic ذو المهمة المحدّدة Task Specific، ويمثّل الإجابات المحتملة التي تتدرّج على سلّم تدريج يبدأ بالدرجة ١ التي تمثّل النهاية الدنيا، وينتهي بالدرجة ٥ أو الدرجة ٧ التي تمثّل النهاية القصوى.

حدود الدراسة:

يتحدّد تعميم نتائج الدراسة بطرق التوفيق المستخدمة في الدراسة وهي ثلاث طرق (المتوسط الحسابي للتقديرين الأصيلين، المتوسط الحسابي لتقديري المقيّمين الأصيلين وتقدير المقيّم الخبير، المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين). ويتحدّد تعميم نتائج الدراسة أيضاً بالمهمة الكتابية المستخدمة في هذه الدراسة والتي اقتصرت على سؤال مقالي واحد من نوع الإجابة المقيّدة يدور حول الفلسفة الشخصية للمعلّم في العملية التدريسية. من جانب آخر، يتحدّد تعميم النتائج بدليلي التصحيح المستخدمين في الدراسة (خماسي التدريج، سباعي التدريج) من النوع الكلي ذو المهمة المحدّدة.

بيانات الدراسة:

اشتملت بيانات الدراسة على إجابات ٢٣٢ معلّماً ومعلّمة؛ منهم ١١٢ معلّماً و ١٢٠ معلّمة، يعلّموا في

عشر مدارس أساسية من المدارس التابعة لمديرية قصبة المفرق للعام الدراسي ٢٠١٠/٢٠١١ تمّ اختيارها بطريقة عشوائية طبقية محدّدة من بين المدارس التابعة لتلك المديرية، وبواقع خمس مدارس للذكور وخمس مدارس للإناث. هذا بالإضافة إلى ستة مشرفين تربويين تم اختيارهم بطريقة قصدية من مجتمع المشرفين العاملين في المديرية نفسها ممن لا تقل خبراتهم التدريسية عن عشر سنوات، ولا تقل خبراتهم في مهنة الإشراف التربوي عن خمس سنوات ليتولوا عملية تقييم إجابات المعلمين على المهمة الكتابية.

أداة الدراسة:

قام الباحث بكتابة سؤال مقالي واحد من نوع الإجابة المقيّدة يدور حول "الفلسفة الشخصية للمعلّم في العملية التدريسية والقيم الأساسية"، ويهدف إلى بيان فلسفة المعلّم الشخصية من خلال توضيح رؤيته المهنية، ورسالته، والأهداف الأساسية التي يسعى إلى تحقيقها، والقيم الأساسية التي يتمثّلها ويلتزم بها، ومفهومه لدوره ودور المتعلّم في ضوء الفكر التربوي المعاصر. وقام الباحث أيضاً بإعداد دليلي تصحيح من النوع الكلّي ذو المهمة المحدّدة: واحد خماسي التدريج والثاني سباعي التدريج. وقد مرت عملية إعداد دليلي التصحيح بالمراحل الآتية (Brookhart, 1999):

تحديد أوصاف مستوى الأداء الأنموذج والذي يعكس الأداء المتميّز، وأعطي الدرجة ٥.

تحديد أوصاف مستوى الأداء الأدنى والذي يعكس فهماً محدوداً للمهمة، وأعطي الدرجة ١.

تحديد أوصاف مستوى الأداء المتوسط من خلال المقارنة بين الأداء الأنموذج والأداء الأدنى، وأعطي الدرجة ٣.

تحديد أوصاف مستوى الأداء الذي يقع بين الأداء الأنموذج والأداء المتوسط من خلال المقارنة بين أوصاف المستويين الأنموذج والمتوسط، وأعطي الدرجة ٤.

تحديد أوصاف مستوى الأداء الذي يقع بين الأداء المتوسط والأداء الأدنى من خلال المقارنة بين أوصاف المستويين المتوسط والأدنى، وأعطي الدرجة ٢.

إعادة تكميم أوصاف مستويي الأداء الأنموذج والأدنى والمتوسط لتصبح على النحو ٧، ١، ٤ بدلاً من ٥، ١، ٣ بهدف بناء دليل التصحيح سباعي التدريج من خلال استحداث أربعة أوصاف لمستويات أخرى: اثنان منها يقعان بين مستوى الأداء الأنموذج ٧ والأداء المتوسط ٤ وأعطيا الدرجات ٦، ٥، واثنان آخران يقعان بين مستوى الأداء المتوسط والأداء الأدنى ١ وأعطيا الدرجات ٣، ٢.

عرض دليلي التصحيح على ثلاثة من أعضاء هيئة التدريس المختصين في القياس والتقويم في الجامعات الأردنية للتأكد من وضوح وسلامة اللغة المستخدمة، وتدرّجهما وفقاً للتقديرات الكميّة المناظرة لكل مستوى من مستويات الأداء أو أوصاف الإجابة. وبناءً على مقترحات المختصين، قام الباحث بتعديل بعض أوصاف الإجابة التي اتفق اثنين من المختصين حول ضرورة تعديلها.

إجراءات الدراسة:

قام الباحث في بداية الفصل الدراسي الأول من العام الدراسي ٢٠١٠/٢٠١١ باختيار المدارس التي شملتها الدراسة، ثم التقى الباحث بالمشرفين (المقيّمين) لتعريفهم بالدراسة وأهدافها، وشرح المهمة الكتابية، وآلية التقييم باستخدام دليلي التصحيح. وبعد ذلك خضع المشرفون لبرنامج تدريبي استغرق عشر ساعات موزّعة على خمسة أيام بواقع ساعتين يومياً، حيث تولى الباحث نفسه عملية التدريب على آلية التقويم من خلال تقييم خمس إجابات (نسخ مصوّرة) تمّ اختيارها عشوائياً من بين الإجابات، وباستخدام أسلوب المناقشة بين المقيّمين أنفسهم بالإضافة إلى الباحث. وبعد التأكد من فهم المقيّمين لآلية التقييم، تمّ اختيار مقيّمين اثنين من بين المقيّمين الستة ليمثلوا المقيمين الخبراء من خلال الاسترشاد بآراء المسئولين في المديرية، وتمّ تعيينهم عشوائياً في مجموعتين مستقلتين بالإضافة إلى تعيين المقيّمين الأربعة الآخرين عشوائياً في المجموعتين بحيث تولى اثنين منهم بالإضافة إلى المقيّم الخبير مهمة التقييم باستخدام دليل التصحيح خماسي التدريج، وتولى الاثنين الآخرين بالإضافة إلى المقيّم الخبير مهمة التقييم باستخدام دليل التصحيح سباعي التدريج. بعد ذلك تمّ توزيع أوراق الإجابة عشوائياً وبالتساوي على المجموعتين بحيث اشتملت كل مجموعة على ١١٦ ورقة إجابة حيث قام كل مقيّم بتقييم الإجابات بشكل مستقل عن زملاءه. وبعد الانتهاء من عملية التقييم، قام الباحث بفرز الإجابات التي اختلف المقيّمان الاثنان حولها في كل مجموعة، وعرض نسخة ثالثة من الإجابة على المقيّم الخبير في المجموعة المعنية بحيث اقتصرت مهمة المقيّم الخبير في كل مجموعة من المجموعتين على تقييم الإجابات التي اختلف المقيّمان في المجموعة الواحدة حولها. من ثمّ تم إدخال البيانات في ذاكرة الحاسوب تمهيداً لتحليلها.

النتائج ومناقشتها

للإجابة عن السؤال الأول من أسئلة الدراسة المتعلق بدلالة الثبات بين المقيّمين للمهمة الكتابية لدى استخدام دليل التصحيح خماسي التدريج ودليل التصحيح سباعي التدريج، تّم استخدام ثلاثة مؤشرات للثبات بين المقيّمين Inter-rater Reliability هي: نسبة الاتفاق (Percentage Agreement)، ومعامل ارتباط الرتب (سبيرمان)، ومعامل الثبات المستند إلى نظرية التعميم G-Theory (2) لدى استخدام كل من دليلي التصحيح الخماسي والسباعي. ويوضّح جدول١ التكرارات والنسب المئوية المقابلة لنقاط الفرق بين تقديرات المقيّمين للمهمة الكتابية، وذلك ضمن كل نوع من دليلي التصحيح الخماسي والسباعي.

جدول ١

التوزيع التكراري والنسب المئوية للفرق بين تقديرات المقيمي َن الأصيلين للمهمة الكتابية باستخدام دليلي التصحيح خماسي وسباعي التدرج

دليل التصحيح	الفرق بالنقاط	التكرار	%
الخماسي	٠	٤٥	٣٨,٧٩
	١	٥٣	٤٥,٦٩
	٢	١٧	١٤,٦٥
	٣	١	٠,٨٦
	٤	٠	٠,٠٠
المجموع		١١٦	١٠٠,٠٠
السباعي	٠	٣٠	٢٥,٨٦
	١	٤٨	٤١,٣٨
	٢	٣٠	٢٥,٨٦
	٣	٨	٦,٩٠
	٤	٠	٠,٠٠
	٥	٠	٠,٠٠
	٦	٠	٠,٠٠
المجموع		١١٦	١٠٠,٠٠

يُلاحظ من جدول١ أن عدد الإجابات التي اتفق المقيّمان في تقديرها لدى استخدام كل من دليلي التصحيح الخماسي والسباعي قد بلغت ٤٥ و ٣٠ إجابة على الترتيب، وتشكّل هذه التكرارات ما نسبته ٣٨,٧٩% و ٢٥,٨٦% من مجموع الإجابات التي خضعت للتقييم باستخدام كل من دليلي التصحيح الخماسي والسباعي وهو ١١٦ إجابة في كل مجموعة, ويُلاحظ أيضاً أن عدد الإجابات التي اختلف المقيّمان في تقديرهما لدى استخدام كل من دليلي التصحيح الخماسي والسباعي وبصرف النظر عن عدد نقاط الفرق بين التقديرات التي منحها المقيّمان قد بلغت ٧١، ٨٦ على الترتيب من مجموع الإجابات، وتشكّل هذه التكرارات ما نسبته ٦١,٢١% و ٧٤,١٤%. وهذا يعني أن نسبة الاتفاق بين التقديرات كانت أكبر لدى استخدام دليل التصحيح خماسي التدريج. ولعل هذا يعود إلى ضيق مدى دليل التصحيح خماسي التدريج مقارنة بدليل التصحيح سباعي التدريج؛ بمعنى أن هامش الحرية لدى استخدام دليل التصحيح خماسي التدريج أقل مما هو عليه لدى استخدام دليل التصحيح سباعي التدريج، الأمر الذي يسمح للمقيّم باستخدام مدى أكبر من التقديرات لدى استخدام دليل التصحيح سباعي التدريج مقارنة بدليل التصحيح خماسي التدريج.

جدول ٢

نسبة الاتفاق بين تقديرات المقيّمين للمهمة الكتابية باستخدام دليلي التصحيح خماسي وسباعي التدريج محسوبة وفقاً لطريقة التوفيق بين التقديرات.

دليل التصحيح	الطريقة: الأولى		الطريقة: الثالثة	
	متوسط التقديرين الأصيلين		متوسط تقدير الخبير والتقدير الأقرب لتقديره	
	ك	%	ك	%
الخماسي	٤٥	٣٨,٧٩	٤٠	٥٦,٣٤
السباعي	٣٠	٢٥,٨٦	٢٦	٣٠,٢٣

ولدى المقارنة بين نسب الاختلاف المقترنة بعدد نقاط الاختلاف، يتبيّن أن أكبر نسبة اختلاف بين المقيّمين كانت في التقديرات المتجاورة سواء في دليل التصحيح الخماسي أو السباعي؛ حيث بلغت نسبة التقديرات المتجاورة لدى استخدام دليل التصحيح خماسي التدريج ٤٥,٦٩% . في حين بلغت النسبة٤١,٣٨% لدى استخدام دليل التصحيح سباعي التدريج. ويُلاحظ أيضاً أن نسبة الاختلاف المقترنة بثلاث نقاط أو أكثر لدى استخدام أي من الدليلين هامشية، حيث بلغت هذه النسبة لدى استخدام دليل التصحيح الخماسي٠,٨٦%. في حين بلغت النسبة٦,٩٠% لدى استخدام دليل التصحيح السباعي، وهذا يؤشر على انخفاض نسبة التقديرات المتطرفة بشكل عام. ويتأكد هذا لدى ملاحظة عدم وجود تقديرات متطرفة بشكل واضح النهاية العظمى أو النهاية الدنيا لدى استخدام أي من الدليلين الخماسي أو السباعي. وربما يشير هذا إلى دقة وصف مستويات الأداء في دليلي التصحيح، ودرجة اهتمام وجديّة المقيّمين، وأهمية تدريب المقيّمين قبل البدء بعملية التقييم، وآلية اختيارهم كما يُشير هيرونيموس ورفاقه (Hieronymous, Hoover, Cantor, & Oberley, 1987). كما يوضّح جدول٢ نسب الاتفاق بين تقديرات المقيّمين وفقاً لطريقة التوفيق بين التقديرات.

يتّضح من جدول ٢ أن نسبة الاتفاق بين تقديرات المقيّمين لدى استخدام الطريقة الثالثة للتوفيق بين التقديرات، والتي تقوم على حساب متوسط تقدير

المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين أكبر منها مقارنة بالطريقة الأولى، والتي تقوم على حساب المتوسط بين التقديرين الأصيلين لدى استخدام أي من دليلي التصحيح الخماسي أو السباعي.

وبما أن مؤشر نسبة الاتفاق يعتمد في حسابه على وجود مقيّمين اثنين، بالإضافة إلى أنه يميل للتضخّم كلما مال المقيّمون إلى الابتعاد عن أطراف تدريج دليل التصحيح، أو الميل إلى النقطة الوسطى في التدريج (Cronbach, Linn, Brennan, & Haertel, 1995)، فقد تمّ اللجوء إلى مؤشر ثان للثبات هو معامل ارتباط الرتب (سبيرمان ρ) بين تقديرات المقيّمين للمهمة الكتابية لدى استخدام كل من دليلي التصحيح الخماسي والسباعي، والذي يوفّر مؤشراً عن الثبات بين المقيّمين أكثر دقة من مؤشر نسبة الاتفاق لأنه يأخذ بالاعتبار التغيّر في رتب الأفراد بصرف النظر عن التقديرات التي يمنحها المقيّمون سواء كانت بعيدة أو قريبة عن الأطراف، أو تميل إلى النقطة الوسطى في التدريج. وقد بلغت قيمتي معامل ارتباط الرتب بين تقديرات المقيّمين للمهمة الكتابية لمجموعة البيانات كاملة (٠,٥٢٨، ٠,٤٣٥) لدى استخدام كل من دليلي التصحيح الخماسي والسباعي على الترتيب بالاعتماد على الطريقة الأولى (متوسط التقديرين الأصيلين) للتوفيق بين التقديرات، في حين بلغت قيمتي معامل ارتباط الرتب بين تقديرات المقيّمين لدى استخدام كل من الدليلين الخماسي والسباعي لمجموعة الإجابات التي اختلف المقيّمان على تقديرهما بصرف النظر عن مقدار الفرق بين تقديريهما مما أدى إلى اللجوء إلى المقيّم الخبير والاعتماد على الطريقة الثالثة للتوفيق بين التقديرات المختلفة والتي تقوم على حساب متوسط تقدير المقيّم الخبير والتقدير الأقرب لتقديره (٠,٨١٤ و ٠,٦٧٣) على الترتيب أيضاً كما هو موضّح في جدول٣.

وبما أن مؤشر معامل ارتباط الرتب يتحدّد بوجود مقيّمين اثنين فقط، كما أنه قد يضخّم الثبات بين المقيّمين عندما يميلوا للتساهل أو التشدّد (Shevelson & Webb, 1991)، فقد تم توظيف نظرية التعميم Generalizability Theory (G-Theory) كأسلوب ثالث يمكّن من تقدير معامل الثبات G- Coefficient ρ2) (عندما يتعلق الأمر بقيام أكثر من مقيّمين اثنين لتقييم المهمة الكتابية كما هو الحال في الدراسة الحالية عندما تمّ اللجوء إلى المقيّم الخبير لتقييم الإجابة إذا تبيّن وجود اختلاف بين تقديري المقيّمين الأصيلين، هذا بجانب إمكانية تجزئة تباين الخطأ الكلي إلى مكوّناته واستخدام هذه المكوّنات في تقدير معامل الثبات (ρ2) بدلاً من استخدام تباين الخطأ الكلي في تقدير الثبات كما هو الحال في نظرية القياس التقليدية Classical Test Theory (CTT) (Brennan, 1996). وقد استخدم تصميم المظهر الواحد المتقاطع One-facet crossed design (p x r) حيث تُشير (p) إلى موضوع القياس Object of Measurement وهم الأفراد (الإجابات)، وتُشير (r) إلى المظهر Facet وهو المقيّمين. ويوضّح جدول٣ معاملات الثبات المحسوبة باستخدام معامل ارتباط الرتب، وتلك المحسوبة في إطار نظرية التعميم.

جدول ٣

معاملات ارتباط الرتب (سبيرمان ρ) ومعاملات الثبات المحسوبة باستخدام نظرية التعميم (G-Coefficient ρ2) لدى استخدام دليلي التصحيح خماسي وسباعي التدريج

دليل التصحيح	الطريقة	الأولى متوسط التقديرين الأصيلين	الثانية متوسط التقديرات الثلاث	الثالثة متوسط تقدير الخبير والتقدير الأقرب لتقديره
الخماسي		ن = ١١٦	ن = ٧١	ن = ٧١
	ρ	٠,٥٢٨*	-----	٠,٨١٤*
	ρ2	٠,٧٠٦	٠,٧٨٩	٠,٨٨٩
السباعي	ρ	٠,٤٣٥*	-----	٠,٦٧٣
	ρ2	٠,٦٣٩	٠,٧٣٨	٠,٨٣٦

* مستوى الدلالة ≤ ٠,٠١

يُلاحظ من جدول٣ أن قيم مؤشرات الثبات سواء تلك المحسوبة باستخدام معامل ارتباط الرتب (سبيرمان ρ)، أو تلك المحسوبة باستخدام نظرية التعميم (ρ2) تختلف باختلاف الطريقة المُستخدمة للتوفيق بين تقديرات المقيّمين، وأن الطريقة الثالثة للتوفيق بين تقديرات المقيّمين والتي تقوم على حساب متوسط تقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين تُفضي إلى مؤشرات ثبات أكبر مقارنة بكل من الطريقة الأولى (متوسط التقديرين الأصيلين)، والطريقة الثانية (متوسط التقديرات الثلاث). وبينما تتفق هذه النتيجة مع نتائج دراسة بريلاند Breland (1983)، فإنها تخالف وجهة نظر مهرنز وليهمان (Mehrens & Lehman, 1991) ورأي شيفلسون و ويب (Shevelson & Webb, 1991) الذين يعتقدون بأن مؤشر ثبات الدرجة الإجرائية ينبغي أن يزداد بزيادة عدد المقيّمين من الناحية النظرية، وتختلف أيضاً عن نتائج دراسة جونسون، وبيني، وجوردون (Johnson et al., 2000)، ونتائج دراسة جونسون، وبيني، وفيشر، وكوهس (Johnson et al., 2003)، والتي أشارت إلى أفضلية الطريقة الثالثة التي تقوم على حساب المتوسط الحسابي للتقديرات الثلاثة (تقدير المقيّم الخبير والتقديرين الأصيلين).

ومع أن مؤشر ثبات الدرجة الإجرائية المحسوب بالطريقة الثالثة، والتي تقوم على حساب المتوسط الحسابي للتقديرات الثلاثة (تقدير المقيّم الخبير والتقديرين الأصيلين) كان أكبر منه مقارنة بالطريقة الأولى التي تقوم على حساب المتوسط الحسابي لتقديري المقيّمين الأصيلين، إلا أنه يقل عن مؤشر الثبات المحسوب بالطريقة الثالثة، والتي تقوم على حساب المتوسط الحسابي لتقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين. وهذا يعني أن زيادة عدد المقيّمين للمهمة الكتابية لا يفضي دائماً إلى زيادة في مؤشر الثبات، إذ يعتمد ذلك على مواصفات المقيّمين أنفسهم، هذا بالإضافة إلى أن ذلك قد يختلف باختلاف عدد من العوامل كنوع المهمة الكتابية، ودرجة صعوبتها، ونوع دليل التصحيح المُستخدم (كلي، تحليلي)، وعدد فئات دليل التصحيح. ويعتقد الباحث أن مسألة المقارنة بين مؤشرات ثبات الدرجة الإجرائية المحسوبة بطرق مختلفة لا تنفصل عن السياق الذي تتم فيه عملية التقييم، ولا تتوقف على عدد المقيّمين فقط كما يُشير الأدب النظري، الأمر الذي يبرّر التوصية بإجراء المزيد من الدراسات حول هذا الموضوع مع الأخذ بعين الاعتبار السياق الذي تتم فيه عملية التقييم.

من ناحية ثانية، أشارت النتائج إلى أن مؤشرات الثبات المحسوبة استناداً إلى نظرية التعميم أكبر من تلك المحسوبة اعتماداً على معامل ارتباط الرتب وبصرف النظر عن عدد فئات دليل المُستخدم لتقييم المهمة الكتابية، وعن طريقة حساب الدرجة الإجرائية. وتتفق هذه النتيجة مع نتائج معظم الدراسات التي قارنت بين طرق تقدير الثبات مثل دراسة جونسون، وبيني، وجوردون (Johnson et al., 2000)، ونتائج دراسة جونسون، وبيني، وفيشر، وكوهس (Johnson et al., 2003). ويعود السبب في ذلك إلى الأساس النظري الذي يعتمد عليه في تقدير الثبات، فبينما يقوم تقدير الثبات كما يعكسه معامل ارتباط الرتب على تقدير قيمة واحدة لتباين الخطأ، يستند تقدير الثبات المحسوب باستخدام نظرية التعميم على تجزئة تباين الخطأ إلى مكوّناته.

ومن جهة ثالثة، أشارت النتائج إلى أن قيم مؤشرات الثبات المحسوبة بأي من طرق تقدير معامل الثبات أكبر لدى استخدام دليل التصحيح خماسي التدريج مقارنة بدليل التصحيح سباعي التدريج. ولعل هذا أمر طبيعي أن تميل نسبة الاتفاق بين التقديرات إلى الزيادة كلما مال عدد فئات دليل التصحيح إلى النقصان.

وللإجابة عن أسئلة الدراسة الثاني والثالث والرابع المتعلقة باختلاف الدرجة الإجرائية المحسوبة بطرق مختلفة للتوفيق بين تقديرات المقيّمين للمهمة الكتابية نفسها تبعاً لطريقة التوفيق المُستخدمة، ودليل التصحيح المُستخدم، والتفاعل بين عاملي طريقة التوفيق ودليل التصحيح المُستخدمين، فقد تمّ حساب ثلاث تقديرات توفيقية (درجات إجرائية) لكل فرد من أفراد الدراسة ممن اختلف المقيّمان الأول والثاني في تقدير إجاباتهم، وبصرف النظر عن مقدار الاختلاف، وذلك اعتماداً على طرق التوفيق الثلاث المُستخدمة في الدراسة. وقد بلغ عدد الإجابات التي اختلف المقيّمان الأصيلان في تقديراتها لدى استخدام دليل التصحيح الخماسي ٧١ إجابة، بينما بلغ عدد الإجابات التي اختلف المقيّمان الأصيلان في تقديراتها ٨٦ إجابة. بعد ذلك تمّ حساب المتوسطات الحسابية والانحرافات المعيارية للدرجات الإجرائية المحسوبة بطرق التوفيق الثلاث وذلك باستخدام كل من دليلي التصحيح الخماسي والسباعي، وهي موضّحه في جدول٤.

يُلاحظ من جدول٤ وجود فروق ظاهرية بين المتوسطات الحسابية للدرجات الإجرائية المحسوبة باستخدام طرق التوفيق الثلاث بين التقديرات سواء لدى استخدام دليل التصحيح الخماسي أو السباعي، ويُلاحظ أيضاً أن متوسط الدرجات الإجرائية المحسوب بالطريقة الثالثة (متوسط تقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين) والذي بلغ ٣,٠٤٢، و ٣,٣٩٥ لدى استخدام كل من دليلي التصحيح الخماسي والسباعي على الترتيب هو الأكبر مقارنة بمتوسطات الدرجات الإجرائية المحسوبة بالطريقتين الأولى (متوسط تقديري المقيّمين الأصيلين) والثانية (متوسط التقديرات الثلاثة)، وأن متوسط الدرجات الإجرائية المحسوب بأي من الطريقتين الأولى (متوسط تقديري المقيّمين الأصيلين) أو الثانية (متوسط التقديرات الثلاثة) هو نفسه لدى استخدام دليل التصحيح الخماسي، بينما يزيد متوسط الدرجات

جدول ٤

المتوسطاتالحسابية والانحرافات المعيارية للدرجات الإجرائية المحسوبة بطرق التوفيق الثلاث موزّعة بحسب دليل التصحيح

طريقة التوفيق	ن	م	ع	ن	م	ع	ن	م	ع
الأولى	٧١	٢,٩٨٦	٠,٨٧٤	٨٦	٣,١٨٦	١,٠٥٧	١٥٧	٣,٠٩٥	٠,٩٨١
الثانية	٧١	٢,٩٨٦	٠,٨١٧	٨٦	٣,٢٨٣	٠,٩٢٧	١٥٧	٣.١٤٩	٠,٨٨٩
الثالثة	٧١	٣,٠٤٢	٠,٩٦٦	٨٦	٣,٣٩٥	١,٠٣٥	١٥٧	٣,٢٣٦	١,٠١٧
المجموع	٧١	٣,٠٠٥	٠,٨٢٤	٨٦	٣,٢٨٨	٠,٨٦١	١٥٧	٣,١٦٠	٠,٨٩١

الإجرائية المحسوب بالطريقة الثانية (متوسط التقديرات الثلاثة) والبالغ ٣,٢٨٣عن المتوسط المحسوب بالطريقة الأولى (متوسط تقديري المقيّمين الأصيلين) والبالغ ٣,١٨٦لدى استخدام دليل التصحيح السباعي. من جهة ثانية، يُلاحظ انخفاض تباين التقديرات المحسوب بالطريقة الثانية (متوسط التقديرات الثلاثة) مقارنة بالطريقتين الأولى (متوسط التقديرات الثلاثة) والثالثة (متوسط تقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين) لدى استخدام أي من دليلي التصحيح الخماسي أو السباعي. وكشف التحليل عن اختلاف متوسط الدرجات الإجرائية بصرف النظر عن الطريقة المُستخدمة لحسابها وذلك لدى استخدام دليل التصحيح الخماسي مقارنة بدليل التصحيح السباعي.

وللكشف عن دلالة الفروق الظاهرية بين متوسطات الدرجات الإجرائية المحسوبة بطرق التوفيق الثلاث باستخدام كل من دليلي التصحيح الخماسي والسباعي، فقد تمّ استخدام أسلوب تحليل التباين ذو القياسات المتكرّرة (الطرق الثلاث لحساب الدرجة الإجرائية) باعتبار عامل دليل التصحيح بمستوييه (خماسي، سباعي) هو متغيّر بين الأفراد repeated measures ANOVA with between – subjects factor. ويوضّح جدول٥ نتائج التحليل.

جدول ٥

نتائج تحليل التباين ذو القياسات المتكرّرة للكشف عن أثر كل من عاملي طريقة التوفيق بين التقديرات ودليل التصحيح في الدرجة الإجرائية

مصدر التباين	مجموع المربعات	د.ح.	متوسط المربعات	قيمة ف
داخل طرق التوفيق				
طريقة التوفيق	١,٤٠٦	٢	٠,٧٠٣	٣,٤٩١
طريقة التوفيق x دليل التصحيح	٤,٦٦	٢	٠,٦٦	١,١٥٧
الخطأ	٦٢,٤٠٧	٣١٠	٠,٢٠١	
بين دليلي التصحيح				
دليل التصحيح	٩,٣٧٢	١	٩,٣٧٢	*٤,٠١
الخطأ	٣٦٢,٣٠١	١٥٥	٢,٣٣٧	

* مستوى الدلالة أقل من ٠,٠٥

يتبيّن من جدول٥ وجود أثر رئيس لعامل طريقة التوفيق بين تقديرات المقيّمين الأصيلين، حيث بلغت قيمة "ف" (٣,٤٩١) وهي ذات دلالة إحصائية بمستوى يقل عن ٠,٠٥ مما يُشير إلى أن الدرجة الإجرائية المحسوبة بالطرق المختلفة للتوفيق بين تقديرات المقيّمين تختلف باختلاف طريقة التوفيق المُستخدمة. وتتفق هذه النتيجة مع نتائج الدراسات التي تناولت أثر طريقة التوفيق بين تقديرات المقيّمين في الدرجة الإجرائية مثل دراسة جونسون، وبيني، وجوردون (Johnson et al., 2000)، ونتائج دراسة جونسون، وبيني، وفيشر، وكوهس Johnson et al. (2003). وكشف التحليل أيضاً عن وجود أثر رئيس لمتغيّر عدد فئات دليل التصحيح (خماسي، سباعي) في الدرجة الإجرائية، حيث بلغت قيمة "ف" (٤,٠١٠) وهي أيضاً ذات دلالة إحصائية بمستوى يقل عن ٠,٠٥، وهذا يتسق مع الافتراض الذي انطلقت منه الدراسة الحالية والذي يُشير إلى أن عدد فئات دليل التصحيح هو أيضاً عامل آخر من بين العوامل التي ينبغي أن تؤخذ بعين الاعتبار لدى حساب الدرجة الإجرائية للتوفيق بين تقديرات المقيّمين.

وللكشف عن مصدر الفروق الإجمالية التي أسفرت عنها نتائج تحليل التباين، والتي أشارت إلى وجود فروق ذات دلالة إحصائية بين متوسطات الدرجات الإجرائية تبعاً لعامل طريقة التوفيق بين تقديرات المقيّمين، تمّ استخدام أسلوب توكي Tukey بهدف المقارنة بين متوسطات الدرجات الإجرائية بين تقديرات المقيّمين باستخدام طرق التوفيق الثلاث. ويوضّح جدول٦ النتائج التي أسفر عنها التحليل.

جدول ٦

نتائج المقارنات البعدية بين متوسطات الدرجات الإجرائية تبعاً لطريقة التوفيق المستخدمة بين التقديرات

طريقة التوفيق	الأولى (م = ٣,٠٩٥)	الثانية (م = ٣,١٤٩)	الثالثة (م =٣,٢٣٦)
الأولى	١,٠٠٠	٠,٠٥٤	*٠,١٤١
الثانية		١,٠٠٠	٠,٠٨٧

* مستوى الدلالة ≤ ٠,٠٥

يتّضح من جدول٦ أن مصدر الفروق الإجمالية التي كشف عنها التحليل هو اختلاف متوسط الدرجات الإجرائية المحسوب بالطريقة الأولى (متوسط تقديري المقيّمين الأصيلين) عن متوسط الدرجات الإجرائية المحسوب بالطريقة الثالثة (متوسط تقدير المقيّم الخبير والتقدير الأقرب لتقديره من بين التقديرين الأصيلين).

في ضوء النتائج التي أسفرت عنها الدراسة الحالية، يوصي الباحث المشتغلين بالقياس والتقويم بصفة عامة، والقائمين على البرامج الاختبارية التي توظّف المهمات الكتابية في برامجها الاختبارية بصفة خاصة أخذ مسألة عدد فئات دليل التصحيح المصمّم لتقييم المهمات الكتابية بعين الاعتبار بصرف النظر عن مستوى دقته، وعن مستوى مهارة المقيّمين الذين يستخدمونه. ويوصي الباحث أيضاً بتحري الدقة لدى اختيار طريقة التوفيق بين تقديرات المقيّمين للمهمات الكتابية، والنظر إلى العوامل الأخرى التي يمكن أن تؤثر في ثبات الدرجات الإجرائية كعدد فئات دليل التصحيح، وبخاصة إذا كانت عملية التقييم تترتّب عليها قرارات حاسمة في حياة الفرد. ولعله من المناسب في هذا الشأن أن يتم تجربة دليل التصحيح المُستخدم قبل اعتماده بشكل رسمي من خلال حساب ثبات الدرجات

الإجرائية المحسوبة بطرق مختلفة للتوفيق بين تقديرات المقيّمين. من ناحية أخرى، يوصي الباحث بإجراء المزيد من الدراسات حول ثبات الدرجات الإجرائية المحسوبة بطرق التوفيق المختلفة، والعوامل التي يمكن أن تؤثر فيها كنوع دليل التصحيح (تحليلي، كلي)، ونوع المهمّة، ومستوى صعوبتها، ومستوى القرار (نسبي، مطلق)، والتي لم يتناولها الباحثون في البيئة العربية، وبخاصة أن هناك اهتماماً متزايداً نحو استخدام طرق التقويم البديل في الدول العربية، واهتماماً ملحوظاً بتقييم الأداء. كما يوصي الباحث بتوظيف النظرية الحديثة في القياس النفسي والتربوي، وإجراء المزيد من الدراسات المتعلقة بمسألة ثبات الدرجات الإجرائية في سياق هذه النظرية.

المراجع

المراجع العربية:

علام، صلاح الدين (٢٠٠٩). **القياس والتقويم التربوي في العملية التدريسية**. عمّان: دار المسيرة.

عودة، أحمد (٢٠٠٤). **القياس والتقويم في العملية التدريسية**. إربد: دار الأمل للنشر والتوزيع.

المراجع الأجنبية:

Breland, H. (1983). *The direct assessment of writing skill: A measurement review*. (Tech. Rep. No. 83-6). Princeton, NJ: College Entrance Examination Board.

Breland, H., & Jones, R. (1984). Perception of writing skills. *Written Communication, 1*, 101-109.

Brennan, R. (1996). Generalizability of performance assessment. In G. Philips (Ed.), *Technical issues in large-scale performance assessment* (pp. 19-58). Washington, DC: National Center for Education Statistics.

Brookhart, S. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Cherry, R., & Meyer, P. (1993). Reliability issues in holistic assessment. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton.

Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1995). *Generalizability analysis for educational assessment*. Los Angeles: Center for the Study of Evaluation, Standards, and Students, and Student Testing, University of California at Los Angeles.

Haswell, R. (1998). Rubrics, prototype, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing, 5,* 231-268.

Hayes, J., Hatch, J., & Silk, C. (2000). Does holistic assessment predict writing performance? Estimating the consistency of student performance on holistically scored writing assignments. *Written Communication, 17,* 3-26.

Hieronymous, A., Hoover, H., Cantor, N., & Oberley, K. (1987). *Handbook for focused holistic scoring*. Chicago: Riverside.

Hopkins, K. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Needham Heights, MA: Allyn & Bacon.

Johnson, R., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education, 16*(4), 299-322.

Johnson, R., Penny, J., & Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13,* 121-138.

Johnson, R., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Writing Communication, 18,* 229-249.

Johnson, R., Penny, J., Gordon, B., Shumate, S., & Fisher, S. (2003). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores. *Language Assessment Quarterly, 2*(2), 117-146.

Johnson, R. L., Penny, J., & Johnson, C. (1998, June). *Score resolution in the rating of performance assessments: Practices and issues*. Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.

Johnson, R., Penny, J., & Johnson, C. (2000, April). *A conceptual framework for score resolution in the rating of performance assessment: The union of validity and reliability*. Paper presented at the annual meeting of the AERA, Orleans, LA.

Linn, R. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Livingston, S. (1998, April). *Results of the pilot test of the School Leaders' Licensure Assessment*. Paper presented at the annual meeting of the AERA, San Diego, CA.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246-276.

Margolis, M., & Ross, L. (1995, April). *Halo and related effects in ratings by standardized patients in clinical evaluation*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Mehrens, W., & Lehmann, I. (1991). *Measurement and evaluation in education and psychology*. Fort Worth, TX: Harcourt Brace.

Myford, C., & Wolfe, E. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3,* 300-324.

National Board for Professional Teaching Standards. (1993). *Candidate guide* [Brochure]. San Antonio, TX: Author. National Education Goals Panel. (1996). *Profile of 1994-1995 state assessment systems and reported results*. Washington, DC: Author.

Olson, J., Bond, L., & Andrews, C. (1999). *Annual survey of state student*

assessment programs: A summary report. Washington DC: Council of chief

State School Officers.

Penny, J. (2003). My life as a reader. *Assessing Writing, 8*, 192-215.

Perlman, C. (2003). *Performance assessment: Designing appropriate performance tasks scoring rubrics*. (ERIC Document Reproduction Service No. ED 480070).

Shevelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. Williamson & B.

Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Cresskill, NJ: Hampton.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-287.

Weigle, S. (1999). Investigating rater / prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6,* 145-178.

Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (2000). The development of large-scale portfolio placement assessment at the University of Michigan: 1992-1998. *Assessing Writing, 6,* 41-84.