تدريج الاختبارات مختلطة الفقرات باستخدام نظرية الاستجابة للمفردة مقارنة بين النماذج الأحادية ومتعددة الأبعاد.[1]

أ.د/ محمد حبشي حسين[2]

كلية التربية – جامعة الإسكندرية

ملخص

يوجد شبه إجماع بين المتخصصين في مجال القياس والتقويم أن لكل من أسئلة الاختيار من متعدد والاسئلة المقالية نقاط قوة ونقاط ضعف، وأن الدمج بين هذين النوعين من الفقرات داخل اختبار واحد يزيد من دقة الاختبار في قياس السمة المراد قياسها من منطلق ان كلا النوعين يكمل أحداهما الأخر. إلا إن الجمع بين هذين النوعين من الأسئلة داخل اختبار واحد اثار بعض القضايا السيكومترية من بينها هل تلك الأسئلة تقيس نفس القدرة أو قدرات متماثلة أم إنها كل نوع يقيس قدرة أو قدرات مختلفة عن تلك يقيسها النوع الاخر، وترتب على هذا السؤال سؤال اخر هل تصلح نماذج نظرية الاستجابة للمفردة أحادية البعد لتدريج هذا النوع من الاختبارات، أم أن هذا النوع من الاختبارات يتطلب استخدام النماذج متعددة الأبعاد. لهذا فقد هدف البحث الحالي إلى مقارنة بين كفاءة النماذج أحادية البعد ومتعددة الأبعاد في تدريج الاختبارات مختلطة الفقرات. وللإجابة عن هذا السؤال فقد استخدم الباحث اختبار لقياس القدرة الرياضة لتلاميذ الصف السادس الابتدائي وقد تكون الاختبار من ١٥ سؤال مقسمة إلى ١٠ أسئلة من نوع الاختيار من متعدد وخمسة أسئلة مقالية، وقد تكونت عينة البحث من ٧٣٨ تلميذ من تلاميذ الصف السادس الابتدائي ٤٠٠ تلميذة و ٣٣٨ تلميذ. وقد استخدم الباحث النموذج ثنائي المعلم لتدريج أسئلة الاختيار من متعدد، واستخدم نموذج التقدير الجزئي العام لتدريج الأسئلة مفتوحة النهايات باستخدام برنامج PARSCALE، واستخدم النموذج ثنائي العامل في حالة النماذج متعددة الابعاد واستخدم حزمة mirt في برنامج R. وقد أظهرت النتائج تفوق النماذج متعددة الابعاد في تدريج الاختبار مقارنة بالنماذج الأحادية وذلك اعتماداً على قيم الدالة المعلوماتية للفقرات وجودة مطابقة النماذج للبيانات.

---

[1] تم استلام البحث في ٢٠٢٢/٨/٢٥/ تقرر صلاحيته للنشر في ٢٠٢٢/٩/٢٩

[2] استاذ علم النفس التربوي ت: 01226032477 Email:mhussein@alexu.edu.eg

مقدمة

تلعب الاختبارات أدوار جوهرية في حياة الأشخاص؛ نظراً لأنها تستخدم لأغراض مختلفة، مثل انتقاء وتسكين الأفراد، وتحديد المجالات المعرفية التي تحتاج إلى تحسين وتخطيط وتطوير البرامج التربوية. تعد علميات تصميم، وتحليل درجات الاختبار وتفسير نتائج الاختبارات جوانب مهمة لقياس مستويات سمات الممتحنين (Kinsey, 2003). عزز اهتمام عامة الناس بالاختبارات المناقشات المتعقلة بثبات وصدق الاختبارات التي تتأثر بالعديد من العوامل مثل طول الاختبار، وصيغة أو شكل الأسئلة وطريقة التصحيح.

تعد أسئلة الاختيارات من متعدد النمط الأكثر انتشاراً في الاختبارات، وبالرغم من حقيقة ان أسئلة الاختيار من متعدد تعرضت للعديد من الانتقادات التي من بينها إمكانية تخمين الإجابة الصحيحة بواسطة الممتحن، فإن الكثير من الاختبارات تتضمن فقط أسئلة اختيار من متعدد بسبب سهولة تصحيحها وقدرتها على تغطية جزء كبير من المادة المتعلمة في وقت قصير، وبالرغم من تلك المميزات إلا أن أسئلة الاختيار من متعدد لا تستطيع قياس التفكير عالي الرتبة.

لهذا فإن اقتصار الاختبارات على أسئلة الاختيار من متعدد تجعل تركيز عمليتي التعليم والتعلم لا تركز على مهارات التحليل والتركيب والتقويم لدي المتعلم، ومن ثم يفتقد المتعلم القدرة على البناء الفعال للمعرفة. ولتقليل جوانب القصور الرئيسية في أسئلة الاختيار من متعدد، فيمكن دمج الأسئلة ذات النهايات المفتوحة أو الأسئلة التي تتطلب بناء الاستجابة بجانب أسئلة الاختيار من متعدد أو أسئلة اختيار الاستجابة في الاختبار. على الجانب الاخر فإن الأسئلة ذات النهايات المفتوحة يصعب تصحيحها بصورة موضوعية وثابتة بالرغم انها تقيس فهم المتعلم للمحتوى على مستوى اعمق (Kim, Walker & MCHale, 2008).

ونتيجة لذلك فقد تحولت العديد من الاختبارات من الاعتماد المطلق على أسئلة الاختيار من متعدد في منتصف القرن العشرين إلى الاستخدام الراهن للاختبارات مختلطة الفقرات التي تشمل أسئلة الاختيار من متعدد واسئلة بناء الاستجابة (Ercikan et al., 1998; Kim et al., 2010, Kuechler and Simkim, 2010). على سبيل المثال، وظفت أسئلة الاختيار من متعدد واسئلة بناء الاستجابة في اختبارات عديدة منها التقييم الوطني للتقدم التربوي National Assessment of Educational Progress (NAEP) وبرنامج التسكين المتقدم Advanced Placement Program (AP) لطلاب الجامعة، اختبار الاستدلال للاستعداد الجامعي SAT

Reasoning Test واختبارات مهارات ما قبل المهنة Pre-Professional Skills Tests (PPST). توصل Lane (2005) إلى أن ٦٣% من الاختبارات التي تستخدم علميات التقييم على مستوى الولاية تبنت الشكل المختلط لأسئلة الاختيار من متعدد وبناء الاستجابة وأن هذا العدد في تزايد.

لا يقتصر مفهوم الفقرات المختلطة على أسئلة الاختيار من متعدد واسئلة بناء الاستجابة او النهايات المفتوحة بل أن ذلك يشمل أي نوع من الاختبارات تتضمن فقرات تتطلب طرق مختلفة للتصحيح، ولهذا فاختبارات الشخصية والاتجاهات التي تتضمن عبارات موجبة وأخرى سالبة تدخل في نطاق هذا المفهوم. فقد أشار كل من Chin and Jin (2020) إلى أن العديد من أدوات تقييم المشهورة لقياس الشخصية تدمج الفقرات ذات الصياغة الموجبة والسالبة؛ فعلى سبيل المثال يتضمن مقايس روزنبرغ لتقدير الذات Rosenberg Self-Esteem خمس فقرات صيغت بصورة موجبة فين حين أن العبارات الخمسة المتبقية صيغة بصورة سالبة، وأن الأدبيات تشير إلى أن التصميم مختلط الفقرات يشجع المستجيب على قراءة ومعالجة الفقرة بعناية، ومن ثم فإن الميل إلى الإذعان Acquiescence المقصود به النزعة إلى الموافقة على جميع الفقرات بغض النظر عن محتواها ربما يقل (Wong et al., 2003).

تتميز الاختبارات مختلطة الفقرات التي تشتمل على أسئلة الاختيار من متعدد وبناء الاستجابة بأنها أدوات قياس عالية الكفاءة في علميتي التعليم والتعلم لقدرتها على التغلب على جوانب القصور الناتجة عن الاقتصار على نمط واحد فقط؛ فعندما يدمج النوعين في اختبار واحد ينتج عن ذلك درجات كلية أكثر استقراراً وتقدير أكثر دقة للسمة الكامنة (Sykes & Yen, 2000). إلا إنه، وكما ذكر Hollingworth, Beard and Proctor (2007) أن بعض التربويين وصانعي القرارات السياسية يرون أن أسئلة بناء الاستجابة والاختيار من متعدد لا تقيس نفس المفهوم عندما توضع في نفس الاختبار.

يعتقد كثير من الباحثين أن دمج أسئلة الاختيار من متعدد وبناء الاستجابة يزيد من دقة القياس بصورة عامة، لأن كلا النوعين من الأسئلة يكمل أحداهما الأخر، تتطلب أسئلة انشاء الاستجابة وقت أطول للاختبار بينما تقيس مهارات الاستدلال والمعرفة العميقة التي يصعب قياسها بأسئلة الاختيار من متعدد. على الجانب الأخر، تعد أسئلة الاختيار من متعدد أكثر فعالية إلا أن البعض يقدم الحجج على أنها تقيس حقائق المعرفة. بالإضافة إلى أن أسئلة الاختيار من متعدد ربما تكون عرضة التأثير السلبي بالتخمين. كما إن أسئلة بناء الاستجابة يمكنها تقديم معلومات

عن الطلاب مرتفعي أو منخفضي القدرة بصورة متطرفة الذين يصعب تقيمهم بواسطة أسئلة الاختيار من متعدد (Ercikan et al., 1998).

يمثل تحليل الاختبارات مختلطة الفقرات تحدي، فلكل نوع من الأسئلة طريقة تصحيح مختلفة، الأسئلة متعددة البدائل يتم تصحيحها بصورة ثنائية في حين ان الأسئلة مفتوحة النهاية يتم تصحيحها بواسطة مقياس متعدد البدائل (Alagoz, 2000, Kim & Lee, 2004). ومن ثم نحتاج إلى طريقة تسمح بتحليل الأسئلة ذات الطرق المختلفة في التصحيح داخل مجموعة واحدة من فقرات الاختبار. تصعب تطبيق نظرية القياس التقليدية على الاختبارات مختلطة الفقرات بسب عدم وجود نموذج مصمم خصيصاً لمعالجة تركيبة الفقرات ذات الطرقة المختلفة في علمية التصحيح المدمجة داخل نفس الاختبار. ولهذا فإن نظرية الاستجابة للمفردة تكون مفيدة بصورة خاصة في التحليل المتزامن للاستجابات الثنائية ومتعددة البدائل، طالما أن شرط أحادية البعد متعدد. التدريج المتزامن ربما تكون أما من خليط من النماذج المختلفة أو نموذج واجد. يتطلب التدريج المتزامن تنفيذ مرة واحدة لبرنامج نظرية الاستجابة للمفردة (Ercikan et al., 1998; Lee & Ansley, 2007)

ففي دراسة Supriyati, Falani and Maulana (2021) قارن بين التدريج المنفصل لكل نوع من أنواع الأسئلة في اختبار الفيزياء بالتدريج المدمج، وقد تكون الاختبار من ٣٠ سؤال من نوع الاختيار من متعدد وخمس أسئلة مفتوحة النهاية. وقد تكونت العينة من ٣٠٠ طالب من ست مدارس ثانوية. وقد تم اختيار العينة بطريقة قصدية، وتم تحيل استجابات الاختبار متعدد البدائل باستخدام نماذج نظرية الاستجابة للمفردة، وقد استخدم نموذج الاختيار المتعدد لأسئلة الاختيار من متعدد، ونموذج الاستجابة المتدرجة للأسئلة مفتوحة النهاية. وقد تم التدريج بطريقتين: بصورة متزامنة وبصورة منفصلة، وبعد ذلك تم مقارنة النتاج للكشف عن قيمة الدالة المعلوماتية الأعلى. أظهرت نتائج التحليل أن الدالة المعلوماتية للفقرات التي تم تدريجها بصورة متأنية للاختبار مختلط الفقرات أعلى من الدالة المعلوماتية الناتجة من التدريج بصورة منفصلة.

بالرغم من أن الدمج المختلط لأسئلة الاختيار من متعدد والنهايات المفتوحة "بناء الاستجابة" يحمل الكثير من المميزات السيكومترية، إلا إنها تؤدي ايضاً إلى أسئلة مهمة عديدة. الأول، ربما واحد من الأسئلة الجوهرية المتعلق بالاختبار مختلط الفقرات ما إذا كان نوعي الأسئلة تقيسا نفس السمات أو سمات متشابهة بدرجة كبيرة. هذا السؤال يقودنا إلى سؤال أكثر أهمية يتعلق بالاختبارات مختلطة الفقرات: هل من المناسب استخدام نظرية الاستجابة للمفردة

إحادية البعد Unidimensional Item Response theory للتحليل المتزامن للبيانات الناتجة من نوعي الأسئلة؟

تطبيق النماذج أحادية البعد على فئات الفقرات أو العبارات متعددة الأبعاد اشغل بال المتخصصين في مجال القياس لعقود عديدة. فقد استدل كلا من Humphreys (1985) and Reckase (1997) على أن اختبارات التحصيل المعرفي في الغالب الأعم متعددة الأبعاد؛ وذهب Humphreys (1986) أبعد من ذلك عندما استدل على أن الأبعاد الثانوية يجب أن تتضمن في أي اختبار لتحسين الثبات. كما أن العوامل التي لا ترتبط بالاختبار مثل الاستراتيجيات لتسريع الإجابة على الاسئلة واستراتيجيات التخمين وغيرها من الاستراتيجيات حكمة الاختبار test-wiseness ربما تخلق بصورة غير متعمدة اختبار متعدد الابعاد.

وفي سياق الاختبارات مختلطة الفقرات، تصمم وتبنى أسئلة النهايات المفتوحة والاختيار من متعدد لأغراض قياس مختلفة بواسطة مجموعات مختلفة من معدي الاختبارات، لهذا فإن قد يكون من الطبيعي أن تقيس قدرات كامنة مختلفة. فقد توصل Thissen et al. (1994) عند تحليل اختبارات الكيمياء وعلوم الكمبيوتر في برنامج التسكين المتقدم إلى وجود درجة ما من تعددية الأبعاد. كما توصل Ercikan and Schwarz (1995) إلى أن النماذج ثنائية العامل تتطابق بصورة متسقة مع الاختبارات مختلطة الفقرات من النماذج أحادية البعد. تدل تلك الشواهد على أنه ربما يكون من غير المناسب استخدام النماذج أحادية البعد مع الاختبارات مختلطة الفقرات. وقد توصل Erickan et al. (1998) إلى أن تدريج الاختبارات مختلطة الفقرات باستخدام النماذج أحادية البعد يسبب فقد في دالة المعلومات للأسئلة ذات النهاية المفتوحة "إنشاء الاستجابة".

يمكن لنماذج نظرية الاستجابة للمفردة متعددة الأبعاد أن تكون مفيدة جداً في محاولة تحسين فهمنا لما تقيسه المفردة أو السؤال، ومستوى كفاءة الطالب في كل سمة، ودرجة دقة التركيبات المختلفة من القدرة محل القياس. يمكن أن تستخدم دالة المعلومات متعددة الأبعاد وإحصاءات المفردة متعددة الأبعاد (أي الصعوبة والتمييز) في المساعدة على تحسين فهمنا لبنية البيانات وتحسين تصنيفات الاختبار (Reckase, 2009). وقد أشار محمد حبشي (٢٠١٨) إلى أن النماذج متعدد الأبعاد تختلف عن النماذج الأحادية في أنها تفترض وجود أكثر من قدرة تسهم في تحديد احتمالية إجابة الفرد عن المفردة إجابة صحيحة، وهذا عكس النماذج أحادية البعد التي

تفترض أن هناك قدرة سائدة أو وحيدة هي تحدد احتمالية الإجابة عن السؤال أو الفقرة بطريقة صحيحة.

كما تختلف النماذج متعددة الأبعاد عن النماذج الأحادية أنها تعطي عدد من معاملات التمييز تساوي عدد القدرات المتضمنة في النموذج، في حين ان النماذج الأحادية تعطي قيمة واحدة لمعامل التمييز، يفترض أنها ثابتة كما هو الحال في نموذج راش، أو يفترض انها متغيرة ويتم تقديرها بواسطة البرنامج المستخدم في التدريج كما هو الحال في النموذج ثنائي البارامتر. ومن منطق المقابلة بين معاملات التمييز في نظرية الاستجابة للمفردة والميل في المعادلة الانحدارية فإن قيم معامل التمييز تشير إلى الاسهام النسبي لكل سمة من السمات في تحديد احتمالية الإجابة الصحيحة على السؤال. ومن هذا المنظور، تنقسم النماذج متعددة الأبعاد إلى نماذج تعويضية ونماذج غير تعويضية أو شبه تعويضية، وذلك وفقاً لتصور الباحث حول أدوار القدرات اللازمة للإجابة عن كل سؤال، هل النقص في قدرة ما يمكن تعويضه بالزيادة في مستوى قدرة أخرى، أم انها غير تعويضية بمعنى أنها لا يعوض الارتفاع في أحد القدرات النقص في قدرة أخرى، فحل المسائل الحسابية اللفظية يتطلب قدرة رياضية وقدرة لغوية، هل نقص قدرة الطالب الرياضية يمكن أن تعوضه ارتفاعه في القدرة اللفظية، أم أن ذلك غير محتمل (محمد حبشي، ٢٠١٩).

كما تنقسم النماذج متعددة الأبعاد بين نماذج استكشافية ونماذج تحققية، ففي النماذج الاستكشافية لا يعرف الباحث بالضبط عدد القدرات اللازمة للإجابة عن فقرات المقياس، لهذا فهو يتبع المدخل الاستكشافي الذي يعتمد على البيانات في تحديد عدد الأبعاد أو العوامل الكامنة، في حين يعتمد المدخل التحققي على معرفة الباحث المسبقة بعدد العوامل المتضمنة في الاختبار وذلك اعتماداً على نظرية أو إطار نظري، أو خبرة سابقة. كما يمكن للباحث أن يقارن بين اعداد مختلفة من العوامل بالاعتماد على جودة مطابقة كل نموذج للبيانات المستخدمة في البحث (محمد حبشي، ٢٠١٨).

ونظراً أن غالبية المقاييس النفسية والتربوية تتكون من عدة أبعاد وليس بعد واحد، وأن تدريج الفقرات داخل كل بعد بصورة مستقلة عن البعد الأخر يؤثر سلبياً على دقة التقدير، كما أن هناك الكثير من المقاييس التي يهتم فيها الباحث بالدرجة الكلية بالإضافة إلى الدرجة على كل بعد من ابعاد المقياس، اضف إلى ذلك أن هناك فقرات تتشبع على عاملين أو أكثر، مما يتطلب معه أن يتم تدريج فقرات المقياس باستخدام نماذج نظرية الاستجابة للمفردة متعددة الأبعاد.

وقد استخدمت النماذج متعددة الأبعاد في أكثر من دراسة، فقد استخدم محمد حبشي (٢٠١٨) نماذج نظرية الاستجابة للمفردة متعددة الأبعاد في تدريج مقياس اليقظة العقلية على عينة من طلاب كلية التربية بلغ حجمها ٥٤٤ طالب وطالبة، وقد أظهرت النتائج أن الدالة المعلوماتية في حالة النماذج متعددة الأبعاد كانت أكثر تمايزاً من نظيرتها في النماذج أحادية البعد.

كما استخدم كلا من Friyatmi and Haryanto (2020) نماذج نظرية الاستجابة للمفردة متعددة الأبعاد في تدريج مقياس مهارات التفكير عالي الرتبة لطلاب الثانوية العامة بماليزيا، وقد تكونت العينة من ٧٥٠ طالب وطالبة، بمعدل ٣٠٨ طالب و٤٤٢ طالبة، وقد أظهرت النتائج أن نماذج نظرية الاستجابة للمفردة تعطي نتائج دقيقة لقدرات الأفراد. كما استخدمت نماذج نظرية الاستجابة للمفردة متعددة الأبعاد في دراسة Quansah et al (2022) لتدريج قائمة مواجهة المواقف الضاغطة لدى عينة من معلمي التربية الرياضية قوامها ٤٨٤ وقد أظهرت النتائج أن المقياس يتكون من اربع أبعاد وهو ما يتفق مع النسخة الأصلية للمقياس، وقد أظهرت النتائج أن الدالة المعلوماتية للاختبار وهي مؤشر على دقة المقياس تصل إلى قمتها في المنتصف أي لدى الأفراد متوسطي القدرة، في حين أنها تكون منخفضة عند الأطراف، أي للأفراد مرتفعي ومنخفضي القدرة على مواجهة الضغوط، ويجدر الإشارة إن هذه النتيجة تكررت في دراسات كثيرة ويرجع هذا إلى أن مستوى صعوبة الأسئلة تكون غالباً حول المتوسط.

وعليه فإن هدف الدراسة الحالية هو مقارنة جودة مطابقة النماذج أحادية البعد ومتعددة الأبعاد في تدريج الاختبارات مختلطة الفقرات، والكشف عن الفروق في قيم الدالة المعلوماتية للاختبار، وقيم معالم الفقرات التي تشمل القدرة التميزية والصعوبة.

المنهج والإجراءات

العينة

تكونت عينة البحث من ٧٣٨ تلميذ من تلاميذ الصف السادس الابتدائي بمحافظة الإسكندرية بإدارة الجمرك التعليمية، وقد انقسمت عينة البحث إلى ٤٠٠ تلميذة و ٣٣٨ تلميذ، بمتوسط عمر زمني ١٢,٢ وانحراف معياري ٠،٨٧.

المقياس

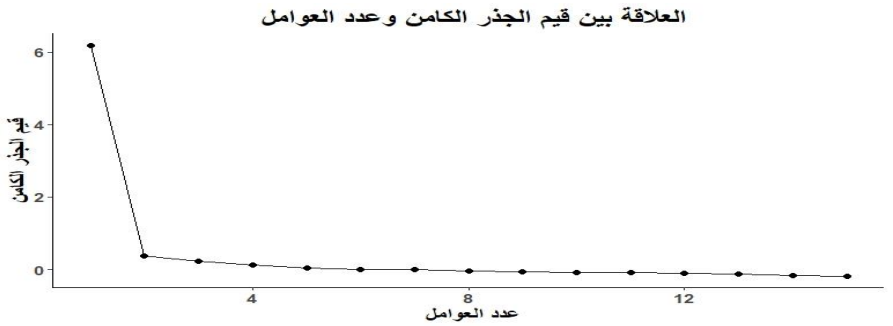
استخدمت الباحث اختبار التحصيل الدراسي المكون من ١٥ سؤال، بواقع ١٠ أسئلة اختيار من متعدد، وخمس أسئلة مفتوحة النهاية، وقد تم تصحيح أسئلة الاختيار من متعدد بواقع

درجة للسؤال الصحيح وصفر للإجابة الخاطئة، في حين تم تصحيح أسئلة النهايات المفتوحة باستخدام rubrics المكون من ٦ مستويات متدرج من صفر إلى خمسة.

وقد استخدم الباحث برنامج Parscale لتدريج المقياس حيث يسمح البرنامج بتدريج الفقرات المختلطة، وذلك بتقسيمها إلى تكتلات Blocks يشمل كل تكتل نوع من الأسئلة، إلا إن البرنامج يستخدم فقط في حالة الاختبارات أحادية البعد، لهذا فعند تدريج الاختبار باستخدام نماذج نظرية الاستجابة للمفردة متعددة الابعاد استخدم الباحث برنامج R والحزمة الإحصائية Multidimensional Item Response Theory (mirt).

نتائج الدراسة

لتحقق من عدد ابعاد المقياس استخدم الباحث التحليل العاملي الاستكشافي، وقد أظهرت النتائج كما يوضحها شكل الانتشار Scree plot الذي يوضح العلاقة بين عدد العوامل وقيم الجذر الكامن، وقد جاءت النتائج كما يوضحها شكل (١)



شكل (١) العلاقة بين عدد العوامل وقيم الجذر الكامن لكل عامل

يتضح من شكل (١) وجود عامل واحد رئيسي يفسر ٤١,٢% من التباين بين الفقرات في حين ان العامل التالي له يفسر فقط ٢,٥% من التباين، مما يدل على وجود عامل وحيد سائد في البيانات، وبالرغم من أن هناك عامل واحد وأن العوامل الأخرى تفسر قدر من التباين أقل بكثير من العامل الأول إلا أن هدف الدراسة الحالية التحقق من النتائج التي تترتب على تجاهل تلك العوامل والاكتفاء بعامل واحد، لهذا فإن الدراسة الحالية سوف تختبر عدة نماذج محتملة تم حصرها في أربع عوامل هي:

١. عامل وحيد عام

٢. عامل عام يشمل أسئلة الاختيار من متعدد واسئلة الصواب والخطأ، بالإضافة إلى عامل يخصص للأسئلة المفتوحة.

٣. عاملين منفصلين: احداهما لأسئلة الاختيار من متعدد والثاني للأسئلة مفتوحة النهاية.

٤. عامل عام يشمل أسئلة الاختيار من متعدد وعامل نوعي لأسئلة الصواب والخطأ وعامل نوعي اخر للأسئلة مفتوحة النهاية.

وقد استخدم الباحث النموذج ثنائي البارامتر two parameter logistic model لتدريج أسئلة الاختيار من متعدد، واستخدم نموذج التقدير الجزئي العام generalized partial credit model لتدريج الأسئلة مفتوحة النهاية. وتم المقارنة بين النماذج المختلفة اعتمادا على عدد الفقرات التي تطابقت مع كل نموذج، وقيم معالم النموذج المستخرجة من كل نموذج، والدالة المعلوماتية للاختبار.

معالم الفقرات

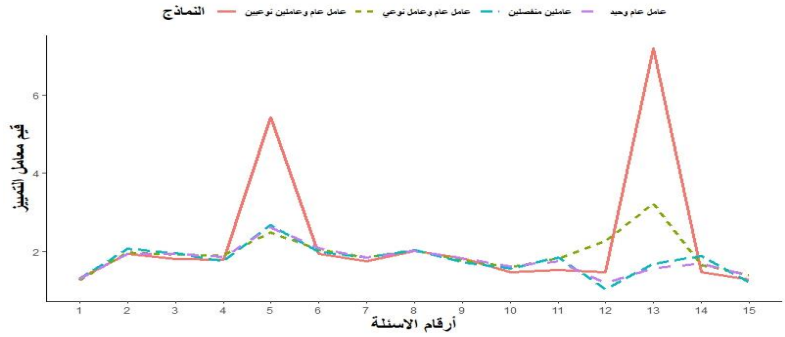
أحد الطرق الهامة للكشف عن الفروق بين النماذج المختلفة هي مقارنة قيم معالم الفقرات، ونظراً أن عدد القدرات المتضمنة في النموذج تؤثر في عدد المعالم أو البارامتر التي نحصل عليها إلا أن هناك دوال

تعطي قيم مجمعة لكل معلم من معالم النموذج، وقد قام الباحث بحساب مؤشرات الإحصاء الوصفي لمعالم

تمييز الفقرات وذلك لكل نموذج من نماذج المقترحة، وجاءت النتائج كما يوضحها جدول (١)

جدول (١) مؤشرات الإحصاء الوصفي لقيم معالم التمييز للفقرات

النماذج	المتوسط	الانحراف المعياري	القيمة العظمى	القيمة الصغرى
عامل عام	١,٧٨	٠,٣٥	١,١٩	٢,٦٣
عاملين منفصلين	١,٧٨	٠,٤	١,٠٤	٢,٦٧
عامل عام وعامل نوعي	١,٩٥	٠,٤٧	١,٢٧	٣,٢٢
عامل عام وعاملين نوعيين	٢,٢٩	١,٦٩	١,٢٩	٧,٢١



شكل (٢) تقديرات قيم معاملات التمييز عبر النماذج الأربعة

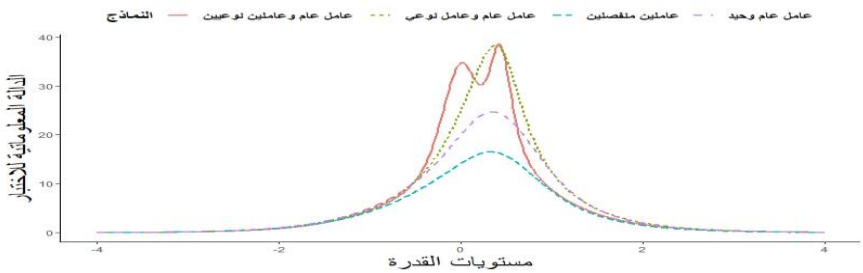
يتضح من جدول (١) وشكل (٢) تقارب قيم معاملات التمييز التي تم تقديرها باستخدام النماذج المختلفة، باستثناء فقرتين اعطي فيها نموذج العامل العام والعاملين النوعيين قيم متطرفة بعيدة عن القيم المتوقعة لمعاملات تمييز الفقرات، وقد يرجع ذلك إلى أن حجم العينة غير كاف لاستخدام نماذج نظرية الاستجابة للمفردة متعددة البدائل التي تحتوي على أكثر من عاملين، وهذه النقطة تحتاج إلى مزيد من البحوث، لمعرفة حجم العينة المناسب لاستخدام نماذج نظرية الاستجابة للمفردة متعددة الأبعاد.

الدالة المعلوماتية للاختبار

اعتمد الباحث عند المقارنة بين نماذج نظرية الاستجابة للمفردة الأحادية ومتعددة الأبعاد على قيم الدالة المعلوماتية للاختبار Test Information Functionالتي تعد مؤشر على جودة الاختبار في تقدير قدرات الأفراد، وقد جاءت النتائج كما يوضحها جدول (٢) وشكل (٣)

جدول (٢) المتوسط والانحراف المعياري والقيم العظمى والصغرى للدالة المعلوماتية للاختبار عبر النماذج المختلفة

النماذج	المتوسط	الانحراف المعياري	القيمة العظمى	القيمة الصغرى
عامل عام	٥,٢٢	٧,٢٦	٠,٠٣	٢٤,٦٥
عاملين منفصلين	٣,٨	٤,٩٨	٠,٠٣	١٦,٤٧
عامل عام وعامل نوعي	٦,٢٤	٩,٨٢	٠,٠٣	٣٨,٣٣
عامل عام وعاملين نوعيين	٥,٩٧	٩,٨٨	٠,٠٣	٣٨,٤٩



شكل (٣) الدالة المعلوماتية للاختبار عبر النماذج المختلفة

يتضح من جدول (٢) وشكل (٣) تفوق نموذج عامل عام مشترك وعامل نوعي للأسئلة مفتوحة النهاية على جميع النماذج حيث أن متوسط الدالة المعلوماتية لهذا النموذج أعلى من بقية النماذج، ويأتي في المرتبة الثانية نموذج العامل العام أو المشترك وعامل نوعي لأسئلة الاختيار من متعدد وعامل نوعي للأسئلة مفتوحة النهاية، وأن اقل النماذج دقة هو نموذج العاملين المنفصلين. ولدراسة دلالة الفروق بين المتوسطات استخدم الباحث تحليل التباين للقياسات المتكررة وجاءت النتائج كما يوضحها جدول (٣)

جدول (٣) تحليل التباين للقياسات المتكررة لتأثير النماذج على الدالة المعلوماتية للاختبار

تأثير	درجة حرية البسط	درجة حرية المقام	قيمة ف	مستوى الدلالة	مربع إيتا العام
النماذج	١,٣٩	١١٠٨,٩٧	١٣٠,٩٦٢	<٠,٠٠١	٠,٠١٣

يتضح من جدول (٣) وجود فروق ذات دلالة إحصائية بين متوسطات الدالة المعلوماتية للاختبار ترجع إلى تأثير الطريقة، ولتحديد مصدر تلك الفروق قام الباحث بإجراء مقارنات ثنائية بين الطرق باستخدام بينفروني، وجاءت النتائج كما يوضحها جدول (٤).

جدول (٤) المقارنات الثنائية بين النماذج المختلفة باستخدام تعديل بينفروني لمستوى الدلالة

المجموعة الأولى	المجموعة الثانية	قيمة ت	مستوى الدلالة	مستوى الدلالة المعدل
عامل عام وعاملين نوعيين	عامل عام وعامل نوعي	-٣,١١٦	٠,٠٠٢	٠,٠١١
عامل عام وعاملين نوعيين	عاملين منفصلين	١١,٤٩٧	٠,٠٠٠	٠,٠٠٠
عامل عام وعاملين نوعيين	عامل عام وحيد	٥,٨٠٣	٠,٠٠٠	٠,٠٠٠
عامل عام وعامل نوعي	عاملين منفصلين	١٣,٥٦٦	٠,٠٠٠	٠,٠٠٠
عامل عام وعامل نوعي	عامل عام وحيد	٩,٧٧٩	٠,٠٠٠	٠,٠٠٠
عاملين منفصلين	عامل عام وحيد	-١٧,٣١٨	٠,٠٠٠	٠,٠٠٠

يضح من جدول (٤) أن هناك فروق بين الطرق المختلفة في متوسطات الدالة المعلوماتية للاختبار، مما يدل على أن أسوأ الطرق هي عاملين منفصلين أي التدريج المنفصل لكل نوع

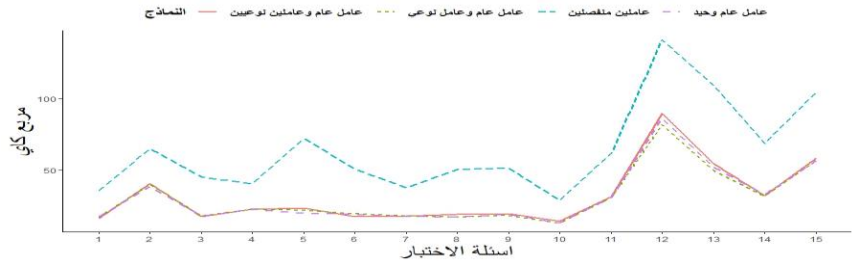
بصورة منفصلة عن النوع الأخر من الأسئلة، وأن طريقة العامل العام وعامل نوعي واحد هي أفضل من النموذج الأكثر تعقيدا وهو عامل عام وعاملين نوعيين، مما يدل على أن التعقيد الزائد في النموذج بإضافة عامل منفصل لأسئلة الاختيار من متعدد يؤدي إلى خفض دال في قيمة الدالة المعلوماتية للاختبار.

ثالثاً: جودة مطابقة الفقرات

للمقارنة بين النماذج المختلفة قام الباحث بمقارنة جودة مطابقة الفقرات لكل نموذج من النماذج المقترحة وذلك باستخدام مربع كاى، وقد قام الباحث بحساب المتوسطات، والانحرافات المعيارية والقيم الصغرى والعظمي لمربع كاي لكل نموذج من النماذج وجاءت النتائج كما يوضحها جدول (٥) وشكل (٤).

جدول (٥)المتوسطات والانحرافات المعيارية والقيم الصغرى والعظمى لقيم مربع كاي للنماذج المختلفة

النماذج	المتوسط	الانحراف المعياري	القيمة العظمى	القيمة الصغرى
عامل عام	٣٠,٧	٢٠,٢١	١٣,١٥	٨٦,٠٧
عاملين منفصلين	٦٤,٢	٣١,٤٨	٢٨,٩٦	١٤١,١٢
عامل عام وعامل نوعي	٣٠,٢٩	١٩,١٩	١٢,٩٥	٨١,٥٢
عامل عام وعاملين نوعيين	٣١,٥٥	٢١,١٧	١٤,٢٦	٨٩,٦٩



شكل (٤) قيم مربع كاي لأسئلة الاختبار لكل نموذج من النماذج المقترحة

يتضح من جدول (٥) وشكل (٤) تقارب قيم مربع كاى للنماذج ماعدا نموذج العاملين المنفصلين، حيث كانت قيم مربع كاى مرتفعة مقارنة ببقية النماذج مما يدل على سوء مطابقة الأسئلة في حالة استخدام النماذج المنفصلة، وللكشف عن دلالة الفروق بين تلك القيم استخدم الباحث تحليل التباين للقياسات المتكررة وجاءت النتائج كما يوضحها جدول (٦).

جدول (٦) تحليل التباين للقياسات المتكررة لتأثير النماذج على قيم مربع كاى

تأثير	درجة حرية البسط	درجة حرية المقام	قيمة ف	مستوى الدلالة	مربع إيتا العام
النماذج	١,٠٣	١٤,٤٤	٨٧,٧٠٣	٠,٠١>	٠,٢٨٨

يتضح من جدول (٦) وجود فروق ذات دلالة إحصائية عند مستوى ٠,٠١ ولتحديد مصدر تلك الفروق قام الباحث بإجراء مجموعة من المقارنات البعدية وجاءت النتيجة كما يوضحها جدول (٧).

جدول (٧) المقارنات الثنائية بين النماذج المختلفة باستخدام تعديل بينفروني لمستوى الدلالة

المجموعة الأولى	المجموعة الثانية	قيمة ت	مستوى الدلالة	مستوى الدلالة المعدل
عامل عام وعاملين نوعيين	عامل عام وعامل نوعي	١,٩٨٦	٠,٠٦٧	٠,٤٠٢
عامل عام وعاملين نوعيين	عاملين منفصلين	-٩,٩١٣	٠,٠٠٠	٠,٠٠٠
عامل عام وعاملين نوعيين	عامل عام وحيد	١,٩٩٤	٠,٠٦٦	٠,٣٩٦
عامل عام وعامل نوعي	عاملين منفصلين	-٩,٠٧٣	٠,٠٠٠	٠,٠٠٠
عامل عام وعامل نوعي	عامل عام وحيد	-١,٠٣١	٠,٣٢٠	١,٠٠٠
عاملين منفصلين	عامل عام وحيد	٩,٤٦٣	٠,٠٠٠	٠,٠٠٠

يتضح من جدول (٧) أن مصدر الفروق في قيم مربع كاى ترجع إلى قيم مربع كاى لعاملين منفصلين، فبينما لا توجد فروق ذات دلة إحصائية بين النماذج الثلاثة في متوسط قيم مربع كاى، فإنه يوجد فروق ذات دلالة إحصائية بين متوسط قيم مربع كاى لنموذج عاملين منفصلين وكل نموذج من النماذج الثلاثة الأخرى وأن تلك الفروق في اتجاه نموذج العاملين المنفصلين أي أنه الأسوء من حيث جودة مطابقة الفقرات.

مناقشة النتائج

هدفت الدراسة إلى المقارنة بين عدة طرق مختلفة لتدريج الاختبارات مختلطة الفقرات، وقد اهتمت الدراسة بنوع واحد من الاختبارات مختلطة الفقرات وهي الاختبارات التي تجمع بين أسئلة الاختبار من متعدد والاسئلة مفتوحة النهاية أو المقالية، وقد تناولت اختبار مكون من ١٤ سؤال وهذا العدد قد يكون غير كاف وان واقع الاختبارات أطول من ذلك، وقد توصلت الدراسة إلى أن نموذج العامل العام وعامل نوعي واحد أفضل من النموذج الأكثر تعقيدا وهو عامل عام وعاملين نوعين والنماذج الابسط وهو نموذج العامل العام، إلا أن هناك نتيجة واضحة من تلك الدراسة تتمثل في أنه من الخطأ تدريج الأبعاد أو الأنواع المتخلفة من الاختبار بصورة منفصلة، فقد اظهر هذا الإجراء انه اسواء الإجراءات حتى في حالة الاختبارات القصيرة، ولهذا فإنه من الضروري أن يجرب الباحث أكثر من نموذج ليختار الأكثر مطابقة لطبيعة بيانات دراسته. لهذا فإن نتائج

تلك الدراسة يجب أن تأخذ وفقاً لمحددات معينة أهمها أن نتائج التحليل العاملي الاستكشافي أظهر وجود عامل عام سائد فسر ما يزيد عن ٤٠% من التباين، وأنه إذا اكان هناك اكثر من عامل أو كانت نسبة التباين المفسر بواسطة العامل الأول أقل من تلك النسبة– فربما تظهر النماذج الأخرى أداءً مختلفاً، كما أن نسبة أسئلة الاختيار من متعدد كانت ضعف نسبة الأسئلة مفتوحة النهاية وإن اختلاف تلك النسبة قد يؤدى إلى نتائج مختلفة– الأمر الذي فتح المجال لدراسات أخرى في ها المجال,

كما ان الدراسة الحالية استخدمت عينة حجمها ٧٣٨ وأن هذا الحجم في مجال نظرية الاستجابة للمفردة ليس بالحجم المناسب، فقد اوصت العديد من الدراسات إلى استخدام احجام عينات قد تصل إلى الألف في حال استخدام النموذج الثلاثي البارامتر، لهذا فإن احجام عينات مختلفة قدي يؤدى إلى نتائج مختلفة من حيث جودة مطابقة الفقرات، حيث أن نماذج نظرية الاستجابة للمفردة متعددة الأبعاد تتطلب أحجام عينات أكبر من ذلك، لهذا فإن الدراسة الحالية توصي بدراسات أخرى تتضمن احجام عينات مختلفة أكبر من حجم العينة المستخدم في تلك الدراسة، لأن حجم العينة من العوامل المؤثرة في دقة تقدير معالم الفقرات.

كما إن الدراسة الحالية لم تدرس دقة تقديرات القدرات المستخلصة من النتائج المختلفة والخطأ المعياري لتقديرات قدرات الأفراد، وخصائص توزيع تلك القدرات من حيث الاعتدالية، كل تلك العوامل يمكن أن يكون لها تأثير في الحجم على الكفاءة النسبية للنماذج المقترحة، لهذا توصي الدراسة الحالية بإجراء المزيد من الدراسات التي تأخذ في اعتبارها طبيعة توزيع القدرة للأفراد المستخدمين في عملية التدريج، وإ ضافة الخطأ المعياري لتقدير القدرة في تكوين صورة واضحة حول الكفاءة النسبية لتلك الطرق.

المستخلص من تلك الدراسة، هو توجيه الباحثين إلى إجراء تحليل عاملي استكشافي للاختبار وفي ضوء نتائج التحليل العاملي الاستكشافي وفي ضوء فهم البحث لمكونات المقياس المستخدم، يقوم الباحث باستخدام أكثر من نموذج ومقارنة جودة مطابقة الفقرات لكل نموذج من النماذج المستخدمة، وذلك باستخدام عدة مؤشرات، كما يجب أن يهتم الباحث بدقة المقياس من خلال المقارنة بين قيم الدالة المعلوماتية للاختبار لكل نموذج، وعدد الفقرات المطابقة لهذا النموذج، وجودة مطابقة الأفراد لكل نموذج لكي يصل إلى النموذج الأفضل والانسب لطبيع المقياس المستخدم.

المراجع

محمد حبشي حسين (٢٠٠٦). تكافؤ القياس بين النسختين العربية والانجليزية لاستبيان مؤشر أساليب التعلم في ضوء نظرية الاستجابة للمفردة. **دراسات نفسية** ١٦(٤)، ٥٣٧-٥٩١.

محمد حبشي حسين (٢٠١١). الخصائص السيكومترية لاستبانة إدارة الوقت لدى عينة من طلاب الجمعة في مصر والسعودية: دراسة تقويمية لنظرية القياس التقليدية ونظرية الاستجابة للمفردة. **المجلة التربوية جامعة الكويت** ٣٥٣-٤٠٤.

محمد حبشي حسين (٢٠١٨). الخصائص السيكومترية لمقياس اليقظة العقلية: مقارنة بين نظرية القياس التقليدية والنماذج الأحادية والمتعددة الأبعاد لنظرية الاستجابة للمفردة. **المجلة المصرية للدراسات النفسية، ٩٩**(٢٨)، ١٩-٧٦.

محمد حبشي حسين (٢٠١٩). تكافؤ/ثبات القياس في البحوث النفسية والتربوية مقارنة بين التحليل العاملي التوكيدي متعدد المجموعات ونظرية الاستجابة للمفردة. **المجلة المصرية للدراسات النفسية، ٢٩**(١٠٣)، ٢٦-٥٤.

محمد حبشي حسين، أحمد محمد عبد الخالق (٢٠١٩). الخصائص السيكومترية للقائمة العربية للعوامل الخمسة الكبرى للشخصية في إطار نظرية الاستجابة للمفردة. **المجلة المصرية للدراسات النفسية، ٢٩**(١٠٥)، ١-٣٢.

Alagoz, C. (2000). *Scoring tests with dichotomous and polytomous* (Unpublished master thesis). University of Georgia, Georgia.

Berger, M. P. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement, 22*(3), 248-258.

Ercikan, K., and Schwarz, R. (1995). Dimensionality of multiple-choice and constructed-response tests for different ability groups, in Paper Presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco (San Francisco, CA).

Ercikan K., Schwarz R. D., Julian M. W., Burket G. R., Weber M. M., Link V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *J. Educ. Meas. 35*, 137–154.

Friyatmi,F.; Mardapi, D., & Haryanto,. (2020). Assessing students' higher order thinking skills using multidimensional item response. Problems of Education in the 21st Century, Problems of Education in the 21st Century, 78(2), Continuous. presented at the April/2020.

Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment Research & Evaluation, 12*(18).

Humphreys, L.G. (1986). Describing 1he elephant. In R.J. Sternberg & D.K. Detterman (Eds.), What is intelligence? (97-100) Norwood, NJ: Ablex.

Hui-Fang Chen & Kuan-Yu Jin (2020): The Impact of Item Feature and Response Preference in a Mixed-Format Design, *Multivariate Behavioral Research*, 1-15.

Humphreys, L.G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B.B. Wolman (Ed.), Handbook of intelligence: Theories, measurements, and applications (201-224). New York: Wiley.

Hussein, M. H. & Abdel-Khalek, A. M. (2021). Developing a Revised Version of the Arabic Big Five Personality Inventory Using Item Response Theory. *Making Quarterly, 62*:2 327-343

Kim, S., & Lee, W. (2004). IRT scale linking methods for mixed-format tests. ACT research report series. Iowa City, IA.

Kim S., Walker M. E., McHale F. (2010). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *J. Educ. Meas*. 47, 186–201

Kim, S. Y., & Lee, W. C. (2018). Simple-Structure MIRT True-Score Equating for Mixed-Format Tests. Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 5), 127.

Kim, S., Walker, M. E., & McHale, F. (2008). *Equating of mixed-format tests in large scale assessments* (ETS Research Rep. No. RR-08-26). Princeton, NJ: ETS.

Kuechler W., Simkin M. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decis. Sci. J. Innovative Educ. 8*, 55–73.

Lee, W., & Ansley, T. N. (2007). Assessing IRT Model-Data Fit for mixed format tests. *Journal of Applied Psychology*, 92(2), 23–50.

Lee, G., & Lee, W. C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29*(3), 224-241.

Osbourn, H.G. (2000). Coefficient α and related internal consistency reliability coefficients. *Psychological Methods, 5*, 343-355.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*(2), 111-120.

Quansah, F., Ankomah, F., Hagan, J. E. Jr., Srem-Sai, M., Frimpong, J. B., Sambah, F., et al. (2022). Psychometric properties of the cultural mix

coping inventory for stressful situations using physical education teachers: a multidimensional item response theory analysis. *BMC Psychology 10*(209):1-12

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van derLinden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 271-286). New York: Springer-Verlag.

Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer Science Business Media.

Skyes, R. C. & Yen, W. M. (2000). The Scaling of Mixed-Item-Format Tests With the One-Parameter and Two-Parameter Partial Credit. *Journal of Educational Measurement , 37*(3), 211-244.

Supriyati, Y.; Falani, I. & Maulana, S. (2021). The information function of mixed-format test of physics learning outcomes measurement. AIP Conference Proceedings

Thissen, D., Wainer, H. ,and Wang, X.(1994).Are tests comprising both multiple Choice and free-response items necessarily less unidimensional than multiple- Choice tests? An analysis of two tests. J. Educ. Meas. 31,113–123.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research, 30*(1), 72–91.

Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α, Revelle's, β and McDonalds ω: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 1-11.

Calibrating Mixed Item Format Tests using Item Response Theory: Comparing Unidimensional versus Multidimensional Models

Dr. Mohamed Habashy Hussein

College of Education Alexandria University

Abstract

There is approximately full agreement among experts in measurement and evaluation that both multiple choice questions and assay questions have strengths and weaknesses and combining both types in single test increases the test accuracy in measuring the target trait from the logic that each type complements the other. However, combining both types into single test raise several questions one of them is these questions measure the same or similar ability or abilities or each type of measure different ability from the abilities are measured by the other types of items? based on this question, another question was raised: are the unidimensional item response modes valid to calibrate this type of test or these types of test require multidimensional item response theory models. Therefore, the current study aimed to compare the relative accuracy of the unidimensional and multidimensional models in calibrating mixed item format tests. To achieve this gool, a test for measuring the math achievement among the six grade students and the test contains 15 questions dividing into 10 multiple choice questions and five assay questions; the sample consists of 738 students from the six grade primary schools, 400 female students, 338 male students. The two-parameter logistic model was used to calibrate the multiple choice questions and generalized partial credit model was used to calibrate the open ended questions using Parscale, and mirt package from R was used to calibrate the multidimensional models. The results indicated that multidimensional models outperform the unidimensional models based on test information function and item goodness of fit for the data.