أثر إضافة بديل "لا أعرف" في أسئلة الاختيار من متعدد على تقدير قدرة الفرد وخفض بارامترات التخمين وتحسين كفاءة الاختبار في ضوء نظرية الاستجابة للمفردة[1]

د/ أمين صبري نور الدين[2]

الأستاذ المساعد بقسم علم النفس التربوي – كلية التربية جامعة عين شمس

المستخلص:

يعتبر سلوك التخمين من الموضوعات المهمة في اختبارات الاختيار من متعدد، نظرا لاحتمال زيادة الخطأ المنتظم أو الخطأ العشوائي الذي يؤدي إلى تضخم درجات الطلاب. وتهدف الدراسة الحالية إلى التعرف على أثر إضافة بديل "**لا أعرف**" في أسئلة الاختيار من متعدد على خفض مؤشرات التخمين، وعلى تحسين بارامترات المفردات، وكفاءة الاختبار، وعلى تقدير القدرة، وذلك في ضوء نظرية الاستجابة للمفردة. تألفت عينة الدراسة من (٩٧) طالبا من الذكور، تراوحت أعمارهم (من ٢١ إلى ٢٢ سنة) بمتوسط عمري مقداره (٢١,٢٦)، وانحراف معياري (٠,٤٤)، تم تقسيمهم إلى مجموعتين أحدهما مجموعة ضابطة (٣٨ طالبا)، والأخرى تجريبية (٥٩ طالبا)، وتم التحقق من تكافؤ المجموعتين في القدرة التحصيلية، وطبق على كل منهم اختبار تحصيلي يتألف من (٢٠) سؤالا من نمط الاختيار من متعدد، ذي أربعة بدائل، وقد تمت إضافة بديل خامس، وهو بديل "**لا أعرف**" لأفراد المجموعة التجريبية يمنح لمن يختاره ($\frac{1}{4}$) درجة، تضاف إلى إجاباته الصحيحة للتحقق من خفض أثر التخمين. أظهرت النتائج أنه لا توجد فروق دالة بين متوسطي المجموعتين الضابطة والتجريبية في درجاتهم على الاختبار التحصيلي، إذ كانت قيمة "ت" (٩٥)= ٠,٤٤ ومستوى دلالة (٠,٦٦)، بالرغم أن المجموعة الضابطة (م=١٢,٥٨، ع=٣,٦٩) كانت أعلى من المجموعة التجريبية (م=١٢,٢٢، ع=٤,٠٩). كما بينت النتائج أيضا أن إضافة بديل لا "اعرف" قد خفّض من بارامترات التخمين لمفردات الاختبار، فالمجموعة التجريبية (م=٠,١٩، ع=٠,٠٣) كانت أعلى بصورة دالة من المجموعة الضابطة (م=٠,٢١، ع=٠,٠٢)، إذ كانت قيمة "ت" (٩٥)= ٣,٠٨، عند مستوى دلالة ٠,٠٠١. كما أسفرت النتائج أيضا أن إضافة بديل "لا اعرف" للمجموعة التجريبية واستخدامه في الاختبار قد حسّن من بارامترات التمييز، فالمجموعة التجريبية (م=١,٢٩، ع=٠,٥٥) كانت أعلى بصورة دالة من المجموعة الضابطة (م=١,٢٠، ع=٠,٢٨)، في متوسط التمييز، إذ كانت قيمة "ت"

---

[1] تم استلام البحث في ٢٠٢١/٧/٣ وتقرر صلاحيته للنشر في ٢٠٢١/٨/١٠

[2] ت: ٠١٠٠٦٣٠٤٧٨٦ Email: aminsabry2011@gmail.com

(٩٥)= ٢,٢١، عند مستوى دلالة ٠,٠٣، وفيما يتعلق بكفاءة الاختبار، كانت المجموعة التجريبية (م=٣,٧٩، ع=١,١٦) أعلى بصورة دالة من المجموعة الضابطة (م=٣,١٢، ع=٠,٠٨) في متوسط مقدار معلومات الاختبار، إذ كانت قيمة "ت" (٩٥)= ٢٤,٣٢، عند مستوى دلالة ٠,٠٠١ وكشفت النتائج أيضا أن المجموعة التجريبية (م=٠,٥٠، ع=٠,١٣) كانت أقل بصورة دالة من المجموعة الضابطة (م=٠,٥٩، ع=٠,١٧)، في متوسط الخطأ المعياري لتقدير القدرة، إذ كانت قيمة "ت" (٩٥)= ٣,١٣، عند مستوى دلالة ٠,٠٠٢ وقد خلصت الدراسة إلى بعض التضمينات التربوية، واقتراح بحوث ذات صلة.

مصطلحات الدراسة: أسئلة الاختيار من متعدد– التخمين – بديل "لا اعرف"– النموذج ثلاثي البارامتر – دالة المعلومات– تقدير القدرة – الخطأ المعياري لتقدير القدرة.

المقدمة:

هناك مجموعة من الأركان والعناصر التي تتضمنها العملية التعليمية، منها المعلم، والمتعلم، والوسيلة التعليمية، ولكن من أهمها على الإطلاق التقويم التربوي. فلا يمكن تصور تحقيق الأهداف، أو تطوير محتوى التعليم بدون الارتكاز على عملية التقويم. ومن ثم؛ فإن الإعداد الجيد لأدوات ووسائل التقويم التربوي من الأهمية بمكان لإنجاز أهداف التعلم، والتوصل إلى نتائج دقيقة. وتعتبر الاختبارات التحصيلية من أهم وسائل تحقيق التقويم التربوي، فهي الأداة الأساسية التي يستخدمها المعلمون في المدارس والمؤسسات التعليمية المختلفة، للتأكد من مدى تحقيق الأهداف التعليمية للطلاب، وقياس مستوى تحصيلهم الدراسي، وتشخيص نواحي الضعف لديهم (رجاء أبو علام، ١٩٨٧) (Nitko, & Brookhart, 2014) . لذا ينبغي صياغة الاختبارات التحصيلية سواء الموضوعية منها أو المقالية بإحكام، حتى تحقق الأهداف المرجوة منها، وتقدم نتائج موثوق بها.

وتتنوع الاختبارات التحصيلية الموضوعية فمنها: أسئلة الإكمال، ومنها البديلين أو الصواب والخطأ، والمطابقة، غير أنه من أشهر الاختبارات الموضوعية وأهمها هي أسئلة الاختيار من متعدد (Frey, 2018; Gierl, Bulut, Guo, & Zhang, 2017) . وتتميز أسئلة الاختيار من متعدد بعدد من المميزات، مثل غيرها من الاختبارات الموضوعية الأخرى، إذ يسهل تقدير درجاتها بسهولة، ولا تستغرق وقتا طويلا للإجابة عن أسئلتها، وهي أيضا موضوعية في تصحيحها (رجاء أبو علام، ١٩٨٧)؛ (Paek 2015) كما تتميز أسئلة الاختيار من متعدد تحديدا –بخلاف غيرها من الاختبارات الموضوعية الأخرى– بالمرونة في قياس عدد من المستويات

التحصيلية المختلفة، فمن الممكن صياغة أسئلة اختيار من متعدد لتقيس القدرة على المعرفة والتذكر، وإمكانية الفهم والاستيعاب، والقدرة على التطبيق في ذات الاختبار، مما يحقق تشخيص التحصيل الدراسي، وتتميز أيضا بقدرتها على التحكم في مدى التجانس بين البدائل، وإمكانية زيادة عدد البدائل، مما يتيح للمختبر التحكم بمستوى صعوبة الأسئلة (رجاء أبو علام، ١٩٨٧) (Parkes & Zimmaro 2016).

غير أنه من أهم المشكلات التي تعاني منها أسئلة الاختيار من متعدد هي قضية التخمين Guessing وذلك منذ منتصف القرن الماضي حتى الوقت الحالي (Angoff, 1989; Baldwin, 2020; Bond, Yan, & Heene, 2020; Carroll, 1945; Hejri, Mafinezhad, & Jalili, 2014 Keats 1952; Plumlee, 1952) . فهي من أهم مصادر خطأ القياس سواء خطأ القياس المنتظم، أو الخطأ العشوائي (Grant & McGrath, 2021; Paek, 2015) .

لقد بذلت جهود عديدة من جانب الباحثين لخفض أثر التخمين سواء أثناء إعداد الاختبار، مثل جعل البدائل متجانسة قدر الإمكان (Nitko, & Brookhart, 2014) ، أو إثناء تطبيق الاختبارات مثل التنبيه على المتعلمين قبيل إجراء الاختبار بعدم ممارسة لسلوك التخمين، ومحاولة اختيار الإجابة الصحيحة، دون إغفال أي سؤال بدون إجابة (صلاح علام ١٩٩١)، أو عند تقدير الدرجات عن طريق معادلات خطية بسيطة تعتمد على عدد الإجابات الصحيحة، والخاطئة، وأحيانا المتروكة أيضا، مثل استقطاع جزء من الدرجة الإجمالية للمتعلم لمظنة التخمين، سواء بمقدار عدد الإجابات الخاطئة، بما يسمى "**بطريقة العقاب**" (American Psychological Association, 2015) ، أو **بتخفيف العقاب** عن طريق افتراض أن عددا من الإجابات الخاطئة قد يكون تخمينا ذكيا (أحمد عودة، ٢٠١٤)، أو **بمكافأة** المتعلم لتركه الأسئلة دون الإجابة عليها (Crocker & Algina, 2008) .

لقد أجريت عدد كبير من الدراسات سواء من منطلق نظرية القياس التقليدية أو المستندة على نظرية الاستجابة للمفردة، وذلك لفحص أثر التخمين في أسئلة الاختيار من متعدد، فلقد أظهرت دراسة صلاح علام (١٩٩١) أن أداء الطلاب يتباين باختلاف الموقف الاختباري (جماعي المرجع/محكي المرجع)، كما أن استخدام تعليمات التصحيح من أثر التخمين قد قلل نسبيا من تخمين الطلاب، غير أنه لم يتمكنوا من استخدام المعلومات الجزئية لديهم، أو التخمين الذكي، مما خفض درجاتهم عن طلاب جماعي المرجع. كما أظهرت دراسة Abad, Olea, and (2001) Ponsoda أنه لا توجد فروق دالة في مؤشرات الصعوبة، والتمييز، والتخمين بين الاختبار ذي الأربعة بدائل والاختبار ذي الخمسة بدائل. كما خلصت دراسة Paek (2015) إلى أن التخمين

في أسئلة الاختيار من متعدد له أثره السلبي على ثبات الاختبارات وأنه كلما كانت المفردات أصعب؛ فإن احتمال التخمين يكون أعلى، ويقل معه معاملات التمييز. كما أوصت دراسة وسام الشرفا (٢٠١٦) باستخدام طريقة المكافأة والعقاب عند تصحيح أثر التخمين، وبينت دراسة حسين الزهراني (٢٠٢١) عدم وجود تأثير دال إحصائياً لعدد بدائل الاختبار بين الاختبارات ذات الثلاثة والأربعة والخمسة بدائل في تقدير قدرات الأفراد في ضوء النموذج الثلاثي البارامتر.

مشكلة الدراسة:

يتضح مما تقدم من نتائج الدراسات التي اهتمت بعزل أثر التخمين، أن هذه النتائج لم تفي بعزل أثر التخمين كما يرجى منها، فمعظم الإجراءات والأساليب المستخدمة لم تخفض من أثر التخمين بالصورة المتوقعة، فالإجراءات المستخدمة لتصحيح أثر التخمين لم تضبط سلوك المتعلمين بصورة كافية (Crocker & Algina, 2008) ، لأن مشكلة التخمين العشوائي أنه لا يكشف عن قدرة المتعلم تحديدا، ولكنه مزيج من قدرة المتعلم وصعوبة المفردات معا. وهناك متعلمون لا يلجؤون إلى التخمين على الإطلاق، فهم يحاولون دائما التوصل إلى الإجابة الصحيحة، بينما يلجأ البعض الآخر إلى ممارسة التخمين العشوائي عندما تكون لديه القدرة على ذلك (Bond et al., 2020) .

كما أن هذه الإجراءات أو تلك المعادلات لم تقدم بديلا مناسبا للمعرفة الجزئية، أو الطريقة التي يستخدمها المتعلم أثناء اختياره بديلا بعينه، فهذه الطرق لم تفسر سلوك المتعلمين في العالم الحقيقي (Crocker & Algina, 2008) . ومن ثم لم يحدث ارتفاع كبير في قيمة الثبات من أثر استخدام التصحيح، ولم يتحقق صدق الاختبار بصورة كافية، إلا إذا كان هناك تفاوت في سلوك التخمين بين المتعلمين، وأن يكون الاختبار متوسط الصعوبة، وأن يكون هناك خمسة بدائل للمفردة على الأقل، (Crocker & Algina, 2008) . حتى زيادة عدد البدائل كإجراء وقائي أثناء إعداد الاختبار، كما اقترحته بعض الدراسات (Ketterlin-Geller, 2005) ، وذلك لخفض التخمين وزيادة الثبات؛ لم يقلل من احتمالية التخمين، حتى لو تمت زيادة عدد البدائل إلى ستة بدائل؛ فإن احتمال التخمين غالبا قائم (Kubinger et al., 2010).

وفي المقابل فإن توجيه المتعلمين أثناء تطبيق الاختبار إلى ضرورة توخي الدقة، وانتقاء الصواب من البدائل، بدلا من اختيار الإجابات الصحيحة عشوائيا؛ فإن هذا التوجيه لا يتم الاستجابة له غالبا، فكثير من الطلاب لا ينتبهون للتعليمات المقدمة أثناء تطبيق اختبارات الاختيار من متعدد (Grant, & McGrath, 2021) . ومن ثم فإن الدرجة المحصلة له قد لا تكون هي

الدرجة الحقيقية للمتعلم (Lord, 1975).

ومن ناحية ثالثة؛ فإن استقطاع مقدار من درجات المتعلمين لافتراض التخمين قد لا يكون عادلا إلى حد كبير، كما أنه قد يحرم المتعلم المتفوق من التخمين الذكي، وليس العشوائي، لاختيار بديل قد يحمل معرفة جزئية للمفردة. بالإضافة إلى ذلك؛ فإن تنبيه المتعلمين بعدم الإجابة عن الأسئلة الصعبة بالنسبة لهم بدلا تخمينها، قد يحد عن سلوك المخاطرة باختيار الإجابة الصحيحة (صلاح علام، ١٩٩١)، بينما عندما يجيب المتعلمون عن جميع المفردات؛ فإنهم قد يحققوا درجة خام أعلى بالفعل، إا أنهم قد يدفعهم ذلك إلى التخمين العشوائي.

بناء على ما سبق؛ يطرح البحث الحالي أسلوبا جديدا ومختلفا، يعالج به بعض أوجه القصور التي تعاني منها الأساليب السابقة لخفض التخمين في أسئلة الاختيار من متعدد، وقد يكون أحد الحلول المثلى لخفض أثر التخمين، وذلك على ثلاثة مستويات: الأول أثناء إعداد الاختبار، ثم عند تطبيق الاختبار، وثالثا عند تقدير الدرجات. فأثناء إعداد الاختبار يتم إضافة بديل "**لا أعرف**" كبديل إضافي أخير لمفردات الاختبار، بغض النظر عن عدد البدائل التي يتضمنها الاختبار. وثانيا عند تطبيق الاختبار إذ يتم توجيه المتعلمين إلى عدم التخمين العشوائي للإجابة الصحيحة، وعوضا عن ذلك من الممكن اختيار بديل "**لا أعرف**" مع التأكيد على أن اختيار هذا البديل سيمنح صاحبه جزء من الدرجة (يتم تحديدها سلفا)، وذلك لتجنيب المتعلم محاولة التخمين العشوائي للإجابة الصحيحة، نظرا لأن التحذير من التخمين مطلقا ربما يقلل من سلوك المخاطرة لدى المتعلم، فعوضا عن ذلك يتم منحه جزء من الدرجة، والمستوى الثالث بعد الانتهاء من الاختبار، وعند تقدير الدرجات يتم منح المتعلم أجزاء من الدرجات التي وعد بها، وذلك بعدد اختياره بديل "**لا أعرف**" كمكافأة له لعدم التخمين، ويمكن أن يكون مقدار الدرجات هي $\frac{1}{4}$ درجة السؤال أو $\frac{1}{3}$ درجة السؤال. تتميز هذه الدراسة أيضا باستخدام نظرية الاستجابة للمفردة، وبالتحديد النموذج ثلاثي البارامتر وليس الأحادي، كما كانت دراسة شادية عبد العزيز وحسان مخلوف (٢٠١٧).

فنظرية القياس التقليدية تحسب مؤشرات الصعوبة والتمييز بالطرق التقليدية؛ ولكنها ليست مؤهلة لحساب مؤشرات التخمين. لأنها لا تفصل بين خصائص الأفراد وخصائص العينة، كما أنها لا تستطيع أن تحسب الأخطاء المعيارية لتقدير القدرة لكل فرد، وإنما تحسب الثبات للاختبار ككل (Azevedo, Oliveira, & Beites, 2019). وتقدم نظرية الاستجابة للمفردة Item Response Theory (IRT) من خلال النموذج ثلاثي البارامتر Three-Parameter Logistic Model (Bond et al. 2020) ، مؤشرات تخمين المفردات، كما تتميز في قدرتها

على الكشف عن صعوبة المفردة، وقوتها التمييزية بدقة أعلى من نظرية القياس التقليدية.

ومن ثم تحددت مشكلة الدراسة في محاولة للإجابة عن الأسئلة الثلاثة الآتية:

أسئلة الدراسة:

١. ما أثر إضافة بديل "لا أعرف" على تقدير قدرة الفرد؟
٢. ما أثر إضافة بديل "لا أعرف" على خفض بارامترات التخمين لمفردات الاختبار؟
٣. ما أثر إضافة بديل "لا اعرف" على تحسين بارامترات المفردات وكفاءة الاختبار؟

أهداف الدراسة:

تهدف الدراسة الحالية إلى الكشف عن أثر اختيار "**بديل لا أعرف**" في أسئلة الاختيار من متعدد على خفض مؤشر التخمين، وعلى تحسين بارامترات المفردات، وكفاءة الاختبار، وعلى تقدير القدرة وذلك في ضوء نظرية الاستجابة للمفردة.

أهمية الدراسة:

١. توجه الدراسة الحالية الاهتمام بسلوك التخمين لطلاب الجامعة.
٢. تحاول الدراسة الحالية إلقاء المزيد من الضوء حول ظاهرة التخمين فيما يتعلق بمدى الثقة بنتائج التحصيل الدراسي.
٣. تلفت هذه الدراسة القائمين على العملية التربوية بالجامعات عن أحد العوامل المهمة الظاهرة حول سلوك التخمين أثناء الإجابة عن أسئلة الاختيار من متعدد.
٤. تساعد نتائج الدراسة الراهنة على وضع أساليب جديدة لخفض أثر التخمين في الاختبارات الموضوعية.
٥. تدفع نتائج تلك الدراسة الناهضين بالعملية التربوية إلى الالتفات نحو تقليص مصادر أخطاء القياس بالنسبة للتخمين.

مصطلحات الدراسة:

أسئلة الاختيار من متعدد:

أسئلة الاختيار من متعدد Multiple-Choice Test هو نوع من أسئلة انتقاء الاستجابة، أو الاختبارات الموضوعية، تتألف من عدد من المفردات، المفردة منه تتكون من قسمين، القسم الأول يسمى الجذع أو المتن Stem، ويكون على شكل سؤال أو جملة ناقصة، والقسم الثاني يسمى بالبدائل أو الخيارات أو الإجابات المحتملة Alternatives، إحداها غالبا تكون الإجابة الصحيحة

Correct ، أما البدائل غير الصحيحة فتُسمى المشتتات أو المموهات Distractors) (Frey, 2018, P3371) . وفي الدراسة الحالية استخدمت أسئلة الاختيار من متعدد في صيغة العبارات الناقصة، يكملها المتعلم من خلال اختيار بديل صحيح من بين أربعة بدائل مطروحة له.

نظرية الاستجابة للمفردة

نظرية الاستجابة للمفردة Item Response Theory (IRT) نظرية جديدة نسبيا للقياس النفسي والتربوي، لمعالجة المفردات وتصميم الاختبارات. وتشترط هذه النظرية توافر مجموعة من الفروض حتى يمكن استخدامها، كما تشترط مطابقة بيانات الاختبار لأحد النماذج الاستجابة للمفردة المحددة للنظرية. وفي الدراسة الحالية تم تحليل البيانات باستخدام النموذج ثلاثي البارامتر. ولا تعتمد النظرية في تقدير درجة الأفراد على عدد الإجابات الصحيحة، كما هو الحال في نظرية الاختبار التقليدية، إنما تعتمد على نمط إجابة الفرد بناء على دالة رياضية مستندة على نظرية الاحتمالات تسمى دالة الاستجابة للمفردة (Hambleton et al., 1991) . وفي الدراسة الحالية تم تقدير القدرة باستخدام أسلوب الأرجحية القصوى Maximum Likelihood Estimation وقد تم حساب ذلك من خلال برنامج BILOG-MG 3 (Zimowski et al., 2003).

بارامترات المفردات:

بارامترات أو معلمات المفردات Item Parameters هي مجموعة من الخصائص التي تتسم بها كل مفردة في الاختبار، مثل الصعوبة Threshold، والقوة التمييزية Slope، والتخمين Asymptote، ويتم تقديرها وفقا لأحد أساليب تقدير القدرة، ويقوم برنامج BILOG-MG 3 بحساب بارامترات كل مفردة بالإضافة إلى متوسط الصعوبة والتمييز والتخمين للاختبار ككل.

بارامتر التخمين:

بارامتر التخمين Guessing Parameter هو مؤشر اختيار المتعلم العشوائي للبديل الصحيح بين البدائل المقدمة، بدون معرفة تامة بالبديل الصحيح، وإنما إما عن عشوائية محضة، أو بناء على معرفة جزئية غير مكتملة بالإجابة الصحيحة. (Crocker & Algina, 2008). ويتراوح المتوسط الجيد لبارامترات تخمين مفردات الاختبار في نظرية الاستجابة للمفردة من (٠ إلى ٠,٢٥) (Frey, 2018) وفي الدراسة الحالية تم تحليل البيانات ببرنامج BILOG-MG 3 لحساب مؤشر تخمين كل مفردة، كما تم حساب متوسط تخمين الاختبار ككل.

كفاءة الاختبار

كفاءة الاختبار Efficiency of Exam هي أحد مؤشرات دقة القياس العامة، وتقاس من خلال دالة المعلومات لكل مفردة، والخطأ المعياري لتقدير القدرة، وكذلك من خلال ثبات الاختبار بالنظرية التقليدية. والاختبار الكفء هو ذلك الاختبار الذي يعطي مقدار أكبر أو أعلى من المعلومات عن الفرد، ومقدار ثبات أعلى للاختبار، مع مقدار أقل من الخطأ المعياري (Frey, 2018). وفي الدراسة الحالية تم حساب ذلك ببرنامج BILOG-MG 3 .

الإطار النظري

تعتبر أسئلة الاختيار من متعدد من أشهر الاختبارات الموضوعية المستخدمة في الوقت الحالي (Gierl et al., 2017) ، إذ تتميز بعدد من السمات الجيدة، منها أن إعدادها لا يستغرق وقتا طويلا، ولا تتدخل الذاتية في تصحيحها، وقدرتها على قياس مستويات تحصيلية متعددة في ذات الاختبار، بما يسمح لواضع الاختبار بالتحكم مستوى صعوبتها (Parkes & Zimmaro 2016).

غير أن من اهم مشكلات أسئلة الاختيار من متعدد هي قضية التخمين (Bond et al., 2020; Hejri et al., 2014) . وبالرغم من أن احتمالية التخمين في أسئلة الاختيار من متعدد هي أقل من الاختبارات الموضوعية الأخرى؛ نظرا لإمكانية زيادة عدد البدائل، والقدرة على التحكم في تجانس البدائل؛ إلا أن مشكلة التخمين فيها تظل موجودة وقائمة، نظرا لأنها الأكثر استخداما من الاختبارات الموضوعية الأخرى. لذلك تعتبر مشكلة التخمين في أسئلة الاختيار من متعدد من أهم المشكلات التي اهتم بها مجال القياس النفسي والتربوي (صلاح علام ١٩٩١) (Hejri et al., 2014) ، وقد تم الالتفات إلى هذه المشكلة منذ الأربعينيات من القرن الماضي (Carroll, 1945; Keats 1952; Plumlee, 1952) ، حتى الوقت الحالي (Angoff, 1989; Baldwin, 2020; Daoust & Côté-Bouchard. 2021; Zimmerman & Williams, 2003) .

وتعد مشكلة التخمين من أهم مصادر خطأ القياس في أسئلة الاختيار من متعدد. فهناك نوعان من خطأ القياس: الخطأ المنتظم The Systematic Error والخطأ العشوائي The Random Error، فالخطأ المنتظم الذي يحدث بشكل متكرر وثابت عند تكرار عملية القياس، وذلك نتيجة لقصور في وسيلة القياس، أو أن يكون لدى الفرد المختبر سمة ثابتة تؤثر على أدائه مثل القلق، أو الخبرة المسبقة بالاختبار، أو عادة التخمين أو وجهة الاستجابة (صفوت فرج،

١٩٨٩). أما الخطأ العشوائي أو الخطأ غير المنتظم فهو مثل غموض التعليمات، أو صعوبة المفردات، أو نمط السؤال التي قد تدفع المتعلمين على التخمين العشوائي (صلاح علام، ٢٠١٥).

ومن ثم فإن التخمين في أسئلة الاختيار من متعدد عُرضة لأن يكون أحد مصادر الخطأ سواء المنتظمة أو غير المنتظمة معا. فعندما يتعلق الأمر بعادة التخمين عند المتعلم في اختبارات الاختيار من متعدد، أو القلق التي يساوره أثناء الموقف الاختباري، أو تدربه المسبق على الاختبار، فإن هذا يعد من مصادر الخطأ المنتظم. وعندما يتعلق الأمر بغموض تعليمات الاختبار، أو صعوبة المفردات، أو نمط المفردات التي تدفع المتعلم للتخمين؛ فإن هذا يعد أيضا من مصادر الخطأ العشوائي. ومن ثم تعتبر مشكلة التخمين من أهم مصادر خطأ القياس–التي يعبر عنها بالفرق بين الدرجة الملاحظة والدرجة الحقيقية– في أسئلة الاختيار من متعدد ، باعتباره من أكثر أنواع الاختبارات الموضوعية استخداما،.

إن عدم وجود تعليمات مسبقة للمتعلم بخصوص التعامل مع المفردات التي يصعب الإجابة عنها، ونزعة المتعلم السائدة نحو التخمين، وغموض التعليمات بخصوص تجنب التخمين، وغيرها من المصادر التي تعد من مصادر الخطأ سواء المنتظم أو العشوائي؛ كل هذه العوامل تؤثر على الدرجة الحقيقية للمتعلم، وتزيد الفجوة بينها وبين الدرجة الملاحظة، مما يؤدي إلى وجود أخطاء للقياس، كما تؤثر على صدق وثبات الاختبارات التحصيلية، وتشكك في النتائج المترتبة على ذلك، وتخفض الثقة في القرارات التربوية المترتبة عليها (أحمد عودة، ٢٠١٤؛ صلاح علام، ٢٠١٥) ، (Grant & McGrath, 2021; Lesage, Valcke, & Sabbe, 2013; Paek, 2015) .

والمقصود بالتخمين هو (احتمال) توصل المتعلم إلى الإجابة الصحيحة عشوائيا، أو من خلال المعرفة الجزئية. فالتخمين العشوائي Random Guessing يشير إلى سلوك المتعلم، عندما لا يعرف الإجابة الصحيحة على الإطلاق، فيختارها بصورة عشوائية، أو أن يكون لديه معرفة جزئية Partial Knowledge ، تساعده على استبعاد بعض البدائل الخاطئة، واستشفاف الإجابة الصحيحة، فالافتراض هو إما يعرف الإجابة الصحيحة أو يختارها عشوائيا من بين البدائلKnowledge or random guessing assumption (Baldwin, 2020) . أي أن أي إجابة خاطئة هي نتيجة للتخمين العشوائي الذي لم يصب البديل الصحيح (Crocker & Algina, 2008) .

ويتوقف التخمين في اختبارات التحصيل من نمط الاختيار من متعدد على مجموعة من المتغيرات بعضها مرتبط ببناء الاختبار، وبعضها مرتبط بسلوك المتعلم. فمن الجوانب المربطة

ببناء الاختبار: طول الاختبار، وزمن الاختبار، ومدى صعوبته، وعدد البدائل (Haladyna, 2004, p217). ومن الجوانب المرتبطة بسلوك المتعلم مثل وجهة الاستجابة أي ميل الفرد لتفضيل بديل محدد للإجابة، وتبني استجابة محددة بناء على موقفه من الاختبار، وليس على ثقته في البديل الذي اختاره، وهذا يحدث عندما تكون التعليمات غامضة، أو المفردات شديدة الصعوبة (صفوت فرج، ١٩٨٩).

وفي نظرية القياس التقليدية Classical Test Theory ؛ تكون نسبة احتمالية التخمين في أسئلة الاختيار من متعدد هي مقلوب عدد البدائل، أي أن احتمالية التخمين= $\frac{1}{\text{عدد البدائل}}$. وعدد البدائل المعروفة تقليديا هي أربعة بدائل، كما تستخدم في اختبارات التقييم الدراسي الأمريكية Scholastic Assessment Test (SAT) واختبار القبول بالجامعات American College Testing (ACT) ومن ثم فاحتمال التخمين السائد هو (٠,٢٥) (Frey, 2018)، وإن كان من الممكن زيادة أو نقصان عدد البدائل بناء على المستوى التعليمي للمتعلمين. وإذا كان من المتاح في نظرية القياس التقليدية حساب مؤشرات الصعوبة والتمييز بالطرق التقليدية؛ إلا أن هذه النظرية ليست مؤهلة لحساب مؤشرات التخمين. فنظرية القياس التقليدية لا تستطيع أن تفصل خصائص الأفراد عن خصائص العينة، كما أنها لا تسطيع أن تحسب الأخطاء المعيارية لتقدير القدرة لكل فرد، وإنما تحسب الثبات للاختبار ككل (Azevedo, Oliveira, & Beites, 2019).

والنظرية المثلى لحساب ذلك هي نظرية الاستجابة للمفردة Item Response Theory (IRT) وتحديدا من خلال النموذج ثلاثي البارامتر Three-Parameter Logistic Model (Bond et al. 2020) ، فهي المناسبة لتصحيح بيانات الاختبار، والكشف عن مؤشرات التخمين لدى المفردات، فضلا عن تميز النظرية في الكشف عن صعوبة المفردة، وقوتها التمييزية بدقة أعلى من نظرية القياس التقليدية.

ولا يهدف بارامتر التخمين في النموذج الثلاثي إلى فحص الدوافع النفسية للتخمين لدى المتعلم، التي تتضمن مستواه التحصيلي وخصائصه الشخصية، واتخاذه نمط محدد للاستجابة على الأسئلة (شادية عبد العزيز وحسان مخلوف، ٢٠١٧) (Haladyna, 2004). ولكن يهدف فقط لتفسير أفضل لمعامل الصعوبة. فيستخدم هذا البارامتر مع بارامتري الصعوبة والتمييز في حساب قدرة المتعلم، فالمفردة ذات التمييز العالي والتخمين المنخفض تعطي مقدارا من المعلومات أعلى من المفردات ذات التمييز المنخفض والتخمين العالي (Hambleton, Swaminathan, & Rogers 1991).

ويفترض النموذج ثلاثي البارامتر في نظرية الاستجابة للمفردة أن هناك دالة رياضية تربط بين احتمالية الاستجابة الصحيحة للمفردة لفرد ذي قدرة ما، وبين البارامترات الثلاثة (الصعوبة والتمييز والتخمين)، وتسمى هذه الدالة الرياضية دالة الاستجابة للمفردة Item Response Function والتي يرمز لها بالرمز P(θ) أي احتمالية الاستجابة الصحيحة على المفردة. وذلك على افتراض أن P(θ) تزيد بزيادة القدرة (θ)، وتمثل الدالة في النموذج ثلاثي البارامتر بالمعادلة الآتية (Lord, 1980, p.12):

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}}$$

إذ ترمز (θ) لبارامتر القدرة، بينما ترمز (a ، b ، c) للبارامترات الثلاثة للمفردة ، فالبارامتر (a) هي القوة التمييزية للمفردة discriminating power ، وتشير إلى انحدار المنحنى المميز للمفردة عندما يكون البارامتر (b) أو صعوبة المفردة item difficulty تساوي قدرة المتعلم ، أما البارامتر (c) فيمثل الخط التقاربي الأسفل lower asymptote للمنحنى المميز للمفردة، وهو بارامتر التخمين Guessing Parameter ، أو ما يطلق عليه بارامتر مستوى الصدفة الزائفة Pseudo-Chance Level Parameter (Hambleton et al., 1991; Lord 1980) ، ويكون مثاليا عندما يتراوح من (٠ إلى ٠,٢٥) (Frey, 2018) .

كما تقدم نظرية الاستجابة للمفردة مؤشرات لكفاءة مفردات الاختبار ودقة القياس من خلال دالة المعلومات Information function ، التي يقصد بها مقدار ما تعطيه المفردة من معلومات عند مستويات القدرة المختلفة. ودالة المعلومات في نظرية الاستجابة للمفردة مشروطة على القدرة، ويتم حسابها عن طريق مقلوب تباين الخطأ المعياري لتقدير القدرة (Weiss & Yoes, 1994, p. 79). فيتم قياس كفاءة الاختبار من خلال ما تقدمه مفردات الاختبار من مقدار أعلى من المعلومات، مع أقل قدر من الخطأ المعياري لتقدير القدرة، بالإضافة إلى ثبات أعلى للاختبار. وحساب البارامترات أو المؤشرات المختلفة في نظرية الاستجابة للمفردة يحتاج إلى برامج وتطبيقات مخصصة لذلك، منها برنامج BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003) ، وهو المستخدم في الدراسة الحالية.

لقد تم اقتراح عدد من الإجراءات التي يمكن أن تتخذ للتقليل من أثر التخمين (Hejri et al., 2014)، منها إجراءات تتخذ قبل تطبيق الاختبارات أو أثناء إعداد الاختبار، ومنها إجراءات

أثناء تطبيق الاختبارات، وأخرى إجراءات بعد تطبيق الاختبار أو عند تقدير الدرجات.

فمن الاستراتيجيات التي تستخدم لخفض أثر التخمين قبل تطبيق الاختبارات، أو أثناء إعداد الاختبار هو جعل البدائل متجانسة قدر الإمكان، وزيادة عدد البدائل مع مراعاة المرحلة التعليمية للمتعلم، وطبيعة المفردة في حد ذاتها (Nitko, & Brookhart, 2014)، غير أن هذا الإجراء محدود نوعا ما، نظرا لأن بعض الأسئلة لا تسمح بزيادة عدد البدائل، كما أن طبيعة المرحلة التعليمية للمتعلم قد لا تسمح بزيادة في عدد البدائل، بالإضافة إلى ذلك؛ فإن دراسة Kubinger, et al., (2010) أظهرت أن زيادة عدد البدائل لم تقلل من أثر التخمين، حتى لو وصل عدد البدائل إلى ستة بدائل، مما يجعل هذا الإجراء قاصرا إلى حد كبير، ولا يضمن سلوك التخمين للمتعلم بصورة كافية.

ومن الإجراءات التي تتخذ أثناء تطبيق الاختبار هي التنبيه على المتعلمين قبيل إجراء الاختبار بعدم ممارسة لسلوك التخمين، ومحاولة اختيار الإجابة الصحيحة، دون إغفال أي سؤال بدون إجابة، وبذلك تكون درجة المتعلم هي مجموع المفردات التي أجاب عنها المتعلم إجابة صحيحة Number- Right Scoring (صلاح علام ١٩٩١) . غير أن Lord (1975) يرى أن عدد الإجابات الصحيحة التي يتم الحصول عليها لا تعبر عن الدرجة الحقيقية للمتعلم، نظرا لأن التنبيهات على المتعلم قد لا يتم الاستجابة لها تماما، أو ترك الإجابة عن السؤال إذا لم يكن المتعلم متأكدا من الإجابة، غير أن هذا الإجراء لا يمنع المتعلم من التخمين الذكي المبني على معرفة جزئية (Nitko, & Brookhart, 2014).

ومن الأساليب المستخدمة للتقليل من أثر التخمين، بعد إجراء الاختبار، أو أثناء تقدير الدرجات هي استقطاع مقدار من الدرجة الإجمالية للمتعلم لمظنة التخمين، أو لتصحيح أثر التخمين Correction For Guessing، وذلك بناء على عدد الإجابات الصحيحة، وعدد الإجابات الخاطئة للمتعلم، بالإضافة لعدد البدائل المستخدمة في الاختبار. وعندما يتم اللجوء إلى استقطاع مقدار من الدرجات لمظنة التخمين؛ يجب إعلام المتعلم بذلك، حتى يمكنه أن يتخذ قرارا بعدم التخمين. وقد تختلف الدرجة المصححة، إذا تم اعتبار المفردات المتروكة، أي المفردات التي لم يجب عنها المتعلم (Crocker & Algina, 2008).

وهناك أكثر من طريقة أو معادلة لذلك، لعل من أبسطها ما تسمى بـــ"**طريقة العقاب بسبب التخمين**"، وتحسب على النحو الآتي:

الدرجة المصححة من أثر التخمين = عدد الإجابات الصحيحة − $\frac{\text{عدد الإجابات الخاطئة}}{\text{عدد البدائل}-1}$ فإذا كان هناك اختبار من نمط الاختيار من متعدد مكون من (٥٠ سؤالا)، وكان عدد البدائل فيه هي (٤ بدائل)، أجاب فيه المتعلم عن (٣٨ سؤالا) إجابة صحيحة، وأخطأ في (٩ أسئلة)، بالإضافة إلى أن المتعلم قد ترك (٣ أسئلة) أخرى، لم يجب عنها بصورة صحيحة أو خاطئة، فتعتبر على أساس هذه الطريقة ضمن الإجابات الخاطئة، لتكون عدد الإجابات الخاطئة (١٢ سؤالا)،

ومن ثم تكون الدرجة المصححة من أثر التخمين = $38 - \frac{12}{1-4}$ = 34 ، وتفترض طريقة "**العقاب بسبب التخمين**" أن الإجابة الصحيحة التي اختارها المتعلم إما أنه يعرفها فيختارها من بين البدائل، أو أنه لم يكن يعرفها، ولكنه خمنها عشوائيا (American Psychological Association, 2015, p. 255) فالقاعدة الأساسية هي أن اختيار البديل الخاطئ هي نتيجة للتخمين العشوائي، إذ لم يختر المتعلم البديل الصحيح (Crocker & Algina, 2008) .

ولذلك تعتمد طريقة "**العقاب بسبب التخمين**" على تصحيح درجة الاختبار من خلال استقطاع مقدار من درجة الإجابات الصحيحة للمتعلم (في المثال السابق ٣٨ درجة)، إذ أنها تعتبر أن قدرا (مجهولا) من هذه الإجابات الصحيحة ربما يكون قد اختارها عشوائيا (فيعاقب) على ذلك بخصم بعض الدرجات (في المثال السابق ٤ درجات) وذلك بمقدار عدد الإجابات الخاطئة التي ارتكبها، كما أنها لا تأخذ في الاعتبار أن بعض الإجابات التي أخطأ فيها المتعلم قد تحمل معرفة جزئية بالإجابة الصحيحة، فيكافأ عليه. ومن ثم يؤخذ على هذه الطريقة أنها تبخس underestimate درجات المتعلم في التقدير، لأنها تخصم بعض الدرجات من الإجابات الصحيحة لاحتمال التخمين العشوائي، ولا تعتبر للمعرفة الجزئية للإجابات الخاطئة، فضلا عن اعتبار الإجابات المتروكة هي خاطئة أيضا (ميساء صمصم، ٢٠١٢).

وبالرغم من أن هذا الاجتزاء لبعض الدرجات ربما يؤدي إلى تحسين الخصائص السيكومترية للاختبارات الموضوعية (Crocker & Algina, 2008)؛ إلا أن هذا الإجراء يؤخذ عليه أنه لا يغير من ترتيب المتعلمين، أو تغيير مراكزهم بناء على تلك الدرجات. كما أنها لا تساعد على تصحيح أثر التخمين المبني على معلومات جزئية، إذ أن هذه الطريقة لا تأخذ في الاعتبار احتمال أن يكون المتعلم قد توصل إلى الإجابة الصحيحة من خلال معرفة جزئية، بل اختارها عشوائيا بالتخمين، ومن ثم لا تقترب كثيرا من الدرجات الحقيقية (Choppin, 1990; Lord 1975) . ويرى Nitko, and Brookhart (2014) أنه من المطلوب تشجيع اللجوء إلى

التخمين الذكي في حال الشك في الإجابة عن السؤال، بدلا من العقاب، عن طريق استقطاع بعض الدرجات. كما أن اجتزاء مقدار من الدرجات لمظنة التخمين ربما يحد من سلوك المخاطرة باختيار الإجابة الصحيحة Risk –Taking Behavior (صلاح علام، ١٩٩١) (Crocker & Algina, 2008).

والطريقة الثانية لتصحيح أثر التخمين، وهو نوع من تخفيف العقاب بسبب التخمين ويكون بالمعادلة الآتية: الدرجة المصححة من أثر التخمين = عدد الإجابات الصحيحة − $\frac{\text{عدد الإجابات الخاطئة}}{\mathbf{2}(\text{عدد البدائل}-\mathbf{1})}$ ، فإذا كان هناك اختبار من نمط الاختيار من متعدد مكون من (٥٠ سؤالا)، وكان عدد البدائل فيه هي (٤ بدائل)، أجاب فيه المتعلم عن (٣٨ سؤالا) إجابة صحيحة، وأخطأ في (١٢ سؤالا)، فتكون الدرجة المصححة من أثر التخمين $= 38 - \frac{12}{(1-4)2} = 36$ ، ومن الملاحظ أن هذه الطريقة قد رفعت الدرجات المصححة من (٣٤ درجة) كما في الطريقة الأولى، إلى (٣٦ درجة) في الطريقة الثانية، مفترضة أن ما يقوم به المتعلم من الإجابة لا يمكن أن عشوائيا تاما، أو تخمينا ذكيا تاما، وأن التخمين يتم مناصفة بين النوعين (أحمد عودة، ٢٠١٤)، ومن ثم فقد استقطعت من الدرجة الأصلية درجتين فقط، ولكن هذه الطريقة لم تأخذ في الاعتبار الإجابات المتروكة، ومن ثم لا تزال الطريقة الثانية تحد أيضا من سلوك المخاطرة عند المتعلم. ويتساءل Budescu, and Bar-Hillel, (1993) حول المغزى النفسي لخصم درجات أو عقاب المتعلم دون التعرف على سبب اختياره لبديل محدد.

والطريقة الثالثة لتصحيح أثر التخمين كان قد وضعها Traub, Hambleton, and Singh (1969) من زمن بعيد وما زالت تستخدم حتى الوقت الحاضر (Crocker & Algina, 2008)، وتعتمد على **المكافأة لعدم المخاطرة** بزيادة الدرجات الأصلية، إذا ترك بعض المفردات دون الإجابة عنها بدلا، بدلا من المخاطرة بالتخمين، فهناك رغبة لدى بعض المتعلمين في حذف أو ترك بعض المفردات دون الإجابة عنها، وربما يكون بعضهم قد أرجأ الإجابة عن بعض المفردات إلى نهاية الاختبار، غير أنه نسي بعضها . وتنص المعادلة على ما يلي:

الدرجة المصححة من أثر التخمين = عدد الإجابات الصحيحة + $\frac{\text{عدد الإجابات المتروكة}}{\text{عدد البدائل}}$ فإذا كان هناك اختبار من نمط الاختيار من متعدد مكون من (٥٠ سؤالا)، وكان عدد البدائل فيه هي (٤ بدائل)، أجاب فيه المتعلم عن (٣٨ سؤالا) إجابة صحيحة، وترك (٣ أسئلة) ، لم يجب عنها بصورة صحيحة أو خاطئة فتكون عندئذ الدرجة المصححة من أثر التخمين $= 38 + \frac{3}{4} = 39$ تقريبا،

ومن الملاحظ أن هذه الطريقة قد كافأت المتعلم لعدم مخاطرته بالتخمين عن طريق زيادة طفيفة في درجاته الأصلية، فهي تمنح درجة إضافية تتناسب مع احتمال أن يختار الإجابة الصحيحة عشوائيا، ولكن يؤخذ على هذه الطريقة أيضا أنها مثل الطريقة الأولى التي تبخس تقدير المتعلم، بينما طريقة **المكافأة** تفرط أو تبالغ overestimate في تقدير المتعلم. ويرى Crocker and Algina (2008) أن هناك ارتباطا خطيا بين الطريقة الأولى والثالثة، كما أن تعليمات الاختبار لأي من هذه الإجراءات ربما تثير دوافع تبعات نفسية للمتعلمين، مما تؤدي إلى تقديم نتائج مختلفة لكل منهم.

ولذلك فإن الطريقة الرابعة قد جمعت بين الطريقة الأولى والثالثة، وتسمى طريقة المكافأة والعقاب في المعادلة الآتية (أحمد عودة، ٢٠١٤):

$$\text{الدرجة المصححة من أثر التخمين} = \text{عدد الإجابات الصحيحة} + \frac{\text{عدد المفردات المتروكة}}{\text{عدد البدائل}} - \frac{\text{عدد الإجابات الخاطئة}}{\text{عدد البدائل}-1}$$

فإذا كان هناك اختبار من نمط الاختيار من متعدد مكون من (٥٠ سؤالا)، وكان عدد البدائل فيه هي (٤ بدائل)، أجاب فيه المتعلم عن (٣٨ سؤالا) إجابة صحيحة، وأخطأ في (٩ أسئلة)، وترك (٣ أسئلة) أخرى، لم يجب عنها بصورة صحيحة أو خاطئة فتكون عندئذ الدرجة المصححة من أثر التخمين $38 + \frac{3}{4} - \frac{9}{1-4} =$ ٣٦ تقريبا.

الدراسات السابقة

لقد أجري عدد من الدراسات السابقة لفحص تصحيح أثر التخمين في اختبارات الاختيار من متعدد، ومن تلك الدراسات دراسة صلاح علام (١٩٩١) التي كشفت عن مدى تباين مستويات التخمين والمخاطرة والأداء، بتباين نوعية الموقف الاختباري (جماعي المرجع/محكي المرجع). وتألفت عينة الدراسة من (١٢٤ طالبا جامعيا) طبق على نصفهم اختبار جماعي المرجع، والنصف الآخر اختبار محكي الرجع، مكون من (٢٤ مفردة) من نمط الاختيار من متعدد، ذي خمسة بدائل في مقرر سيكولوجية التعلم. أظهرت النتائج أن أداء مجموعتي التجربة يتباين بتباين نوعية الموقف الاختباري، إذ كان متوسط أداء طلاب اختبار محكي المرجع أعلى نسبيا من أداء طلاب جماعي المرجع، بينما كان متوسط درجات طلاب مجموعة الموقف الاختباري جماعي المرجع يفوق طلاب محكي المرجع، فهم أكثر ميلا للمخاطرة من طلاب محكي المرجع. كما أظهرت الدراسة أن استخدام تعليمات التصحيح من أثر التخمين لطلاب الاختبار محكي المرجع قد قلل نسبيا من التخمين، ومن الميل للمخاطرة، غير أنها أفقدتهم استخدام المعلومات الجزئية لديهم مما

خفض درجاتهم عن طلاب جماعي المرجع.

وهدفت دراسة Abad et al., (2001) إلى التعرف على العدد الأمثل من البدائل، وأثر ذلك على بارامترات المفردة وعلى تقدير القدرة، وكانت أداة الدراسة اختبار في اللغة الإنجليزية مكون من ثلاثة نماذج: وثلاثة بدائل وأربعة بدائل وخمسة بدائل ، تم تطبيقه على (٤٥٢) طالبا من المرحلة الثانوية والجامعية، ومن خلال تحليل البيانات باستخدام النموذج الثلاثي المعلمة في نظرية الاستجابة للمفردة؛ أسفرت النتائج عن وجود فروق دالة في معاملات القدرة والصعوبة والتمييز لصالح النموذج ذي الثلاثة بدائل، وعدم وجود فروق دالة بين النموذجين ذي الأربعة بدائل والخمسة بدائل.

وقارنت دراسة Kubinger et al. (2010) بين ثلاثة اختبارات مكونة من (١٨٨ فقرة) ذات أنماط مختلفة من الأسئلة الأول: نمط الأسئلة المفتوحة، والثاني الاختيار من متعدد التقليدي، أي إجابة واحدة صحيحة وستة بدائل خاطئة، والثالث الاختيار من متعدد ذي الإجابتين الصحيحتين مع خمسة بدائل خاطئة، وقد طبقت الاختبارات الثلاثة على أكثر من (٧٠٠٠) طالبا وطالبة من الصف الثاني الإعدادي. وأظهرت النتائج أن الاختيار من متعدد التقليدي أي إجابة واحدة صحيحة، كان الأقل صعوبة في الاختبارات الثلاثة، بينما لم تكن هناك فروق دالة في الصعوبة بين النمط المفتوح والإجابتين الصحيحتين.

وقدمت دراسة Chiu (2010) أسلوبا لتصحيح التخمين في إطار النموذج ثلاثي البارامتر، من خلال إجراءات Mantel-Haenszel والانحدار اللوجستي، وذلك باستخدام تقريب Taylor وطرق الاحتمال الشرطي، استخدمت الدراسة بيانات محاكاة، وأظهرت النتائج كفاءة نظرية الاستجابة للمفردة في عدم فقدان الكثير من مقدار المعلومات، بسبب استخدام تقريب تايلور كما خفضت نسبة التحيز بمقدار يتراوح من ٦% إلى ١٠%.

وسعت دراسة ميساء صمصم (٢٠١٢) إلى المقارنة بين ثلاثة طرق لتصحيح اختبارات الاختيار من متعدد (الطريقة التقليدية، ومعادلة التصحيح لأثر التخمين، وطريقة الاختيار الجزئي) لضبط التخمين على مؤشر ثبات المفردات وصدق الاختبار. وتألفت العينة من (٩٠ طالبة) من الصف الخامس الابتدائي، طبق عليهن اختبار تحصيلي، وأظهرت الدراسة ارتفاع ثبات الاختبار عندما يكون التصحيح بالاختيار الجزئي، يليه التصحيح لأثر التخمين ثم الطريقة التقليدية.

وانصبت دراسة أحمد عودة (٢٠١٤) على تقديم صورة معدلة لتصحيح معامل صعوبة الفقرة لأثر التخمين في اختبار الاختيار من متعدد لجيلفورد لاستخدامها في الاختبارات التحصيلية

والعقلية، حيث تركز على التكامل بين التخمين القائم على المعرفة الجزئية، والفروق الفردية لدرجة المخاطرة في ضوء التعليمات ذات الصلة بالتخمين، وما يترتب عليها من إجراءات في التصحيح، وتحليل الدرجات وقد أرفقت الدراسة الحالية جداول لاستخراج معاملات الصعوبة القائمة على التخمين العشوائي مقابل معاملات الصعوبة المقدرة بالمعادلتين الأصلية والمعدلة .

وفحصت دراسة Paek (2015) تأثير التخمين على ثبات الاختبارات وفقا للنموذج الثلاثي البارامتر في نظرية الاستجابة للمفردة، وذلك بالتفاعل مع تفاوت خصائص المفردات الأخرى (الصعوبة والتمييز). وقد استخدمت الدراسة ثلاثة اختبارات لبيانات محاكاة (٢٠، ٤٠، ٦٠ مفردة)، موزعين على (٩٠٠٠ فرد). أظهرت الدراسة أنه كلما كانت المفردات أصعب؛ فإن احتمال التخمين يكون أعلى، ويقل معه معاملات التمييز. وعندما تكون الصعوبة مساوية لقدرة الأفراد؛ فإن التخمين يكون منخفضا، وأحيانا ليس له تأثير يذكر. كما كشفت الدراسة عن أنه عندما يرتفع التخمين إلى (٠,٤٢)؛ فإن الثبات ينخفض إلى (٠,٥٠)، بينما عندما ينخفض التخمين إلى (٠,٠٨)، أو يصل إلى (٠,٢٥)؛ فإن معاملات الثبات تجاوز قيمتها (٠,٨٥)، بغض النظر عن انخفاض القدرة (−٢) أو ارتفاعها (+٢).

وقارنت دراسة وسام الشرفا (٢٠١٦) أثر ثلاث طرق للتصحيح لضبط أثر التخمين (العقاب، والمكافأة، والمكافأة والعقاب) على مقدار التضخم في قيمة معامل ارتباط المفردات بالدرجة الكلية، تألفت عينة الدراسة من (٣٠٠ طالب وطالبة) من الصف الخامس الابتدائي، مقسمين إلى ثلاث مجموعات وفقا لطريقة التصحيح، طبق عليهم اختبار تحصيلي مكون من (٢٥ مفردة) ذي أربع بدائل. أسفرت النتائج أن أقل مقدار محسوب للتضخم في معامل ارتباط المفردة بالدرجة الكلية هو عند استخدام طريقة المكافاة والعقاب، بينما أعلى متوسط للتضخم كان لطريقة العقاب. وقد أوصت الدراسة باستخدام طريقة المكافأة والعقاب عند تصحيح أثر التخمين.

وسعت دراسة محمد اليامي (٢٠١٦) إلى التحقق من أثر طريقتين من طرق تصحيح التخمين: أثر التخمين والاختيار الجزئي، على صدق وثبات اختبار تحصيلي مكون من (٣٨ مفردة) من نوع الاختيار من متعدد، طبق على (٣٢٥ طالبا) من الصف الثاني الثانوي، تم تقسيمهم إلى مجموعتين بالتساوي، طبق على كل منهما إحدى الطريقتين. وأظهرت الدراسة لا تختلف معاملات الثبات والصدق اختلافا دالا بين الطريقتين، كما لم تختلف متوسطات قدرة الطلاب بين الطريقتين: أثر التخمين والاختيار الجزئي.

وهدفت دراسة شادية عبد العزيز وحسان مخلوف (٢٠١٧) إلى الكشف عن العلاقة بين

التخمين وخصائص كل من الفرد والمفردة، ومدى دقة طريقة تحليل البواقي المعيارية في الكشف عن التخمين، وذلك من خلال نسب ترجيح الفرد لصحة بدائل الإجابة، كمحك مرجعي لتحديد الاستجابات المخمنة. وتكونت عينة الدراسة من (٣٤٢ طالبا) جامعيا، طبق عليهم اختبار تحصيلي مكون من (٣٠ مفردة) من نمط الاختيار من متعدد، لكل مفردة أربعة بدائل منها بديل واحد صحيح. أظهرت النتائج أن هناك علاقة متوسطة قوية بين تخمين الإجابة وتقدير قدرة الفرد، كما توجد علاقة متوسطة بين الحث على التخمين وتقدير صعوبة المفردة، وأن طريقة تحليل البواقي المعيارية لم تكشف عن أغلب الاستجابات المخمنة، وبينت النتائج أيضا عدم كفاية محك التباعد بين تقديري القدرة والصعوبة للكشف عن التخمين.

وسعت دراسة عبد الله الرشيدي وآمال الزعبي (٢٠١٩) إلى تقصي أثر عدد البدائل في اختبار الاختيار من متعدد على الخصائص السيكومترية للاختبار التحصيلي بنماذج ثلاثة خمسة بدائل، وأربعة بدائل، وثلاثة بدائل ، وتكونت عينة الدراسة من (٦٤١) طالباً من طلاب المدارس الإعدادية، طبق عليهم اختبار تحصيلي من النماذج الثلاثة، وتوصلت نتائج الدراسة إلى وجود فروق دالة في بارامتر التمييز بين النماذج الثلاثة لصالح الاختبار ذي الخمسة بدائل، كما أسفرت النتائج أيضا عن وجود فروق دالة في بارامتر التخمين بين النماذج الثلاثة لصالح الثلاثة بدائل، بينما لا توجد فروق دالة في بارامتري الصعوبة وتقدير القدرة بين النماذج الثلاثة.

وكشفت دراسة حسين الزهراني (٢٠٢١) عن أثر عدد البدائل في اختبار الاختيار من متعدد على تقدير قدرة الفرد، ومعالم المفردات وفقاً للنظرية التقليدية والنموذج الثلاثي المعلمة. وتألفت عينة الدراسة من (٣٠٠٠ طالبا)، طبق عليهم اختبار تحصيلي (٣٩ مفردة) في مقرر الرياضيات للصف الثالث الثانوي، مكون من ثلاثة نماذج متكافئة تماما عدا عدد البدائل: (خمسة بدائل، وأربعة بدائل، وثلاثة بدائل)، وأظهرت النتائج عدم وجود تأثير دال إحصائيًا لعدد بدائل الاختبار بين نماذج الاختبار الثلاثة في تقدير قدرات الأفراد في ضوء النموذج الثلاثي البارامتر، بينما يوجد تأثير لعدد بدائل الاختبار في تقدير بارامترات الصعوبة والتمييز والتخمين بين النماذج الثلاثة لصالح النموذج ذي الخمسة بدائل، إذ كلما زادت عدد البدائل ازداد معامل الصعوبة، وارتفع معامل التمييز وانخفض مؤشر التخمين.

فروض الدراسة

يتضح مما تقدم من نتائج الدراسات السابقة أن الجهود التي بذلت لتقليص أثر التخمين في أسئلة الاختيار من متعدد في النظرية التقليدية لم تحرز تقدما كبيرا، إذ أن استقطاع مقدار من الدرجات لمظنة التخمين قد يكون غير عادل في أغلب الأحيان، كما أنه ربما يكون له آثار

تحصيلية و سلوكية أخرى مثل الحد من المعرفة الجزئية، والتخمين الذكي. ومن ثم يفترض البحث الحالي أن اختيار المتعلم لبديل "**لا أعرف**" عندما لا يعرف الإجابة الصحيحة، ومنحه جزء من الدرجة، كمكافأة له على "عدم المجازفة" في اختيار بديل آخر "مشكوك فيه" –قد يكون خاطئا– لا يؤثر على تقدير قدرته، بمعنى أنه لا يضخم inflating درجات المتعلم، عن درجاته الأصلية التي أجاب فيها إجابة صحيحة، ولكن يفترض أيضا أن استخدام بديل "**لا أعرف**" –حتى وإن منح أجزاء إضافية من الدرجات– يخفض من مؤشر التخمين لمفردات الاختبار، نظرا لأن هذا الإجراء يجنبه محاولة اختيار بديل قد يكون خاطئا، فيحصل عندئذ على جزء من "درجات مضمونة"، مما يقلل من فرص التخمين العشوائي. كما أن هذا الإجراء يجنب المتعلم ترك بعض المفردات دون الإجابة عليها، مما يعطيه الفرصة إما للإجابة الصحيحة –كما يراها– أو اختيار بديل "لا أعرف" فيجنبه التخمين العشوائي.

يفترض البحث الحالي أيضا أن خفض بارامتر التخمين يعظم من متوسطات بارامترات تمييز المفردات، بالإضافة إلى ذلك، فإنه يحسن من كفاءة الاختبار، كما تحسب بمقدار المعلومات التي يوفرها الاختبار ككل، ويقلص من الخطأ المعياري لتقدير القدرة، وذلك في ضوء نظرية الاستجابة للمفردة. ومن ثم توصلت الدراسة إلى الفروض الآتية:

الفرض الأول: لا يؤثر إضافة بديل "لا أعرف" على تقدير قدرة الفرد.

الفرض الثاني: يخفض إضافة بديل "لا أعرف" بارامترات التخمين لمفردات الاختبار.

الفرض الثالث: يحسن إضافة بديل "لا أعرف" بارامترات المفردات وكفاءة الاختبار.

المنهج والإجراءات:

المنهج المستخدم في الدراسة الحالية هو المنهج شبه التجريبي Quasi Experimental Method، إذ أن هناك مجموعتين من الطلاب، تم تعيين أحدهما كمجموعة تجريبية، والأخرى كمجموعة ضابطة، ونظرا لأن الباحث لا يمكنه إعادة توزيع الطلاب على المجموعتين، فكانت الدراسة شبه تجريبية.

عينة الدراسة:

تألفت عينة الدراسة من (٩٧) طالبا من الذكور، من المستوى الخامس في تخصص الدراسات الإسلامية، مما يدرسون مقرر علم النفس الشخصية، وجميعهم من الطلاب المنتظمين بالدراسة، بجامعة الجوف، بشمال المملكة العربية السعودية، تراوحت أعمارهم (من ٢١ إلى ٢٢ سنة) بمتوسط عمري مقداره (٢١,٢٦)، وانحراف معياري (٠,٤٤) ويوضح جدول (١) توزيع عينة الدراسة.

الجدول (١): توزيع عينة الدراسة

المجموعة	العدد	النسبة
الضابطة	٣٨	٣٩%
التجريبية	٥٩	٦١%
المجموع	٩٧	١٠٠%

الأدوات:

الاختبار التحصيلي:

لتحقيق هدف الدراسة؛ قام الباحث بتصميم اختبار تحصيلي في مقرر "**علم النفس الشخصية**" لطلاب الدراسات الإسلامية، للمجموعتين معا (الضابطة والتجريبية)، وكان هذا الاختبار جزء من الأعمال الفصلية، ويتألف من (٢٠) سؤالا من نمط الاختيار من متعدد، ذي أربعة بدائل، ويستحق عليه درجة من (٢٠ درجة)، بمتوسط معياري (١٠ درجات)، وقد تمت إضافة بديل خامس، وهو بديل "**لا أعرف**" للمجموعة التجريبية للتحقق من خفض أثر التخمين. وقد تم تحديد زمن تطبيق الاختبار لكل من الشعبتين (١٢ دقيقة)، وهو متوسط زمن الإجابة عن أسئلة الاختبار.

التحقق من الخصائص السيكومترية للاختبار:

صدق وثبات الاختبار:

قام الباحث بالتحقق من صدق الاختبار بعدة طرق، إذ قام في البداية بالتحقق من صدق الاختبار عن طريق عرضه على مجموعة من الأساتذة الذين يقومون بتدريس نفس المقرر في الشُعب (المجموعات) الأخرى، أو ممن سبق لهم تدريس المقرر من قبل. وقد تمت الاستجابة لملاحظات الزملاء من حيث الصياغة والأسلوب، وقد اتفقوا على جودة صياغة الأسئلة، وشموليتها للمحتوى المحدد، كما تم التحقق من صدق المحتوى عن طريق مطابقة عدد الأسئلة، وتوزيعها على الموضوعات المختلفة، ومدى تحقيقها للأهداف التعليمية المحددة، ولمستويات التحصيل المتدرجة، وقد تبين أن الاختبار مطابق لجدول المواصفات، ويعكس الأهداف التعليمية المصاغة، كما أن أسئلة الاختبار تقيس مستويات التحصيل المختلفة على النحو الآتي: مستوى التذكر (٧ أسئلة) تمثل (٣٥%) من الاختبار، وهي الأسئلة ذات الأرقام: ٤، ٥، ٦، ٧، ٨، ٩، ١٣، ومستوى الفهم (٨ أسئلة)، بنسبة (٤٠%) من الاختبار، وهي الأسئلة ذات الأرقام: ١، ٢، ٣، ١٦، ١٧، ١٨، ١٩، ٢٠، أما الأسئلة التي تقيس التطبيق فكانت (٥ أسئلة) بنسبة (٢٥%)، وهي الأسئلة ذات الأرقام: ١٠، ١١، ١٢، ١٤، ١٥ . وهذه النسب تمثل جدول المواصفات للموضوعات المحددة. كما تم حساب ثبات الاختبار التحصيلي عن طريق أسلوب معامل ألفا

كرونباخ وبلغت قيمته (٠,٧٩)، ويعرض الملحق (أ)، (ب) نموذجا أسئلة الاختبار لمقرر علم النفس الشخصية للمجموعتين، كما يعرض الجدول (٢-أ، ب، ج) جدول المواصفات والأوزان النسبية وأعداد الأسئلة للاختبار التحصيلي.

الجدول (٢-أ): المواصفات لاختبار الأعمال الفصلية

المحتوى	التذكر	الفهم	التطبيق	الأهمية
الدرس الأول	أن يصف العصابية	أن يستنتج مذهب السلوكية أن يستنتج طبيعة العادات السلوكية	أن يضع مثالا لانفعالات الرضيع	%٢٠
الدرس الثاني	أن يعرف المنخفضين في العصابية	أن يضع مثالا لتوجه السلوكية أن يتعرف على طبيعة انفعال الحزن	أن يستنتج الانفعالات في مرحلة الطفولة	%٢٠
الدرس الثالث	أن يعرف المرتفعين في الانبساط أن يصف الطيبة في الشخصية	أن يستخلص مميزات اختبارات الشخصية	أن يستنتج انفعالات المراهق	%٢٠
الدرس الرابع	أن يعرف الانفتاح على الخبرة	أن يستخلص عيوب مقاييس الشخصية أن يستخلص مميزات الشخصية	أن يستنبط أساليب التحكم في الانفعالات	%٢٠
الدرس الخامس	أن يصف المرتفعين في يقظة الضمير أن يعدد العوامل التي تؤثر في الانفعالات	أن يتعرف على عيوب مقاييس الشخصية	أن يكشف العلاقة بين القلق والتعلم	%٢٠
الأهمية النسبية	%٣٥	%٤٠	%٢٥	%١٠٠

الجدول (٢-ب): الأوزان النسبية لاختبار الأعمال الفصلية

المستوى / المحتوى	التذكر	الفهم	التطبيق	و ن للمحتوى
الدرس الأول	%٧	%٨	%٥	%٢٠
الدرس الثاني	%٧	%٨	%٥	%٢٠
الدرس الثالث	%٧	%٨	%٥	%٢٠
الدرس الرابع	%٧	%٨	%٥	%٢٠
الدرس الخامس	%٧	%٨	%٥	%٢٠
و ن للمستوى	%٣٥	%٤٠	%٢٥	%١٠٠

الجدول (٢-ج): أعداد الأسئلة لاختبار الأعمال الفصلية

المستوى / المحتوى	التذكر	الفهم	التطبيق	و ن للمحتوى
الدرس الأول	١	٢	١	٤
الدرس الثاني	١	٢	١	٤
الدرس الثالث	٢	١	١	٤
الدرس الرابع	١	٢	١	٤
الدرس الخامس	٢	١	١	٤
و ن للمستوى	٧	٨	٥	٢٠

إجراءات الدراسة:

تم تطبيق اختبار علم النفس الشخصية على مجموعتين من طلاب الدراسات الإسلامية ممن يدرس لهم الباحث، وقد تم تعيين أحدهما كمجموعة ضابطة، وعددهم (٣٨ طالبا)، والأخرى باعتبارها مجموعة تجريبية، وعددهم (٥٩ طالبا). وقد تم تطبيق أداة الدراسة وهو اختبار علم

النفس الشخصية على المجموعتين، حسب توقيت اختبار كل مجموعة. ولا يختلف الاختبار في المجموعة التجريبية عن المجموعة الضابطة إلا في إضافة بديل خامس لكل سؤال، بالإضافة إلى البدائل الأربعة، وهو بديل "لا أعرف" مع تنبيه طلاب هذه المجموعة أنه من لا يعرف الإجابة الصحيحة بدلا من اختيار بديل بصورة عشوائية قد يكون خاطئا؛ فإنه من الممكن اختيار البديل الخامس "**لا أعرف**" فيضمن لنفسه ($\frac{1}{4}$) الدرجة في الاختبار، تضاف على إجاباته الصحيحة، وذلك التعرف على أثر اختيار بديل "**لا أعرف**" على خفض تخمين الاختبار، أما المجموعة الضابطة فقد تلقوا نفس الاختبار تماما، ولكن مع إبقاء البدائل الأربعة فقط، أي بدون بديل "لا أعرف" .

الأساليب الإحصائية المستخدمة:

تم استخدام الإحصاءات الوصفية: المتوسطات، والانحرافات المعيارية، والإحصاءات الاستدلالية: اختبار "ت" للمقارنة بين المجموعتين المستقلتين T-Test For Independent Samples ، والمجموعتين المرتبطتين، T-Test For Paired- Samples ، وقد أجريت هذه التحليلات الإحصائية باستخدام برنامج SPSS Ver. 26 (IBM Corp. 2019) . وتم حساب الثبات بطريقة ألفا– كرونباخ، وطريقة التجزئة النصفية، وتبنت الدراسة الحالية المحكات التي أوصى بها Nunnally (1978)، وأكدها أيضا DeVellis (2016, p 136) و (Azevedo, et al., 2019, p 330) أن معامل ألفا إذا كانت قيمته (٠,٧٠)، يعتبر مقبولا، وإذا تراوح الثبات من (٠,٧٠) إلى (٠,٨٠) يكون ثباتا معتبرا، ومن (٠,٨٠) إلى (٠,٩٠)؛ يكون جيد جدا، وعندما يتجاوز الثبات أكثر من (٠,٩٠) يكون ثباتا ممتازا. وقد تم الأخذ في الاعتبار بهذه المحكات.

كما أجريت التحليلات الإحصائية لنظرية الاستجابة للمفردة باستخدام برنامج BILOG-MG 3 (Zimowski et al., 2003). وتم تحديد النموذج ثلاثي البارامتر 3- Parameter Logistic Model وتقدير الأرجحية القصوى Maximum Likelihood Estimation لتحليل البيانات، مع اتخاذ مدى واسع لتقدير القدرة (من –٤ إلى +٤)، ومتوسط التمييز للاختبار لا يقل عن (٠,٨٠)، ومتوسط التخمين لا يزيد عن (٠,٢٥) وفقا لما قرره Hambleton et al. (1991) . كما تتحقق كفاءة الاختبار من خلال دالة معلومات الاختبار، والخطأ المعياري لتقدير القدرة، فالاختبار الأفضل هو الذي يعطي قيمة متوسط معلومات average information أعلى، وأقل خطأ معياري لتقدير القدرة.

النتائج ومناقشتها:

أولاً: الإحصاءات الوصفية للنتائج:

أظهرت النتائج كما يوضحها جدول (٣) أن طلاب المجموعة الضابطة قد حصلوا على درجات تراوحت من (٣ إلى ١٨ درجة)، بمتوسط (١٢,٦ درجة)، وانحراف معياري مقداره (٦,٦٩)، وقد كان معامل الثبات بطريقة ألفا كرونباخ للمجموعة الضابطة (٠,٧٦)، وبطريقة التجزئة النصفية كان (٠,٨١). أما طلاب المجموعة التجريبية فقد أحرزوا درجات في الاختبار التحصيلي تراوحت من (٤ إلى ١٩ درجة)، بمتوسط (١٢,٢ درجة)، وانحراف معياري بمقدار (٤,٠٩)، كما بلغت قيمة معامل ثبات بطريقة ألفا كرونباخ للمجموعة التجريبية (٠,٨١)، وبطريقة التجزئة النصفية (٠,٨٤)، وتعتبر هذه قيم ذات اعتبار (Azevedo et al., 2019) .

الجدول (٣): الإحصاءات الوصفية لدرجات الاختبار التحصيلي لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	ن	م	ع	حدود الدرجات	الثبات ألفا
الضابطة	٣٨	١٢,٥٨	٦,٦٩	٣-١٨	٠,٧٦
التجريبية	٥٩	١٢,٢٢	٤,٠٩	٤-١٩	٠,٨١

وتمثل مدى هذه الدرجات تنوعا واسعا ليقيس الاختبار التحصيلي مستويات متفاوتة من قدرات الطلاب، كما يعتبر متوسط الدرجات أعلى قليلا من المتوسط المعياري وفقا لما تمت الإشارة إليه في توصيف أداة الاختبار. كما تعتبر قيم معاملات ثبات ألفا للاختبارين ثباتا جيدا (Azevedo et al., 2019; DeVellis, 2016; Nunnally, 1978) . كما يعرض جدول (٤) قيم بارامترات الصعوبة والتمييز والتخمين والمعلومات لكل من المجموعة الضابطة والتجريبية، إذ تبين هذه القيم مدى مناسبتها للنموذج ثلاثي البارامتر، وفقا للمحكات والمؤشرات التي تمت الإشارة إليها في الموضع الخاص بالتحليلات الإحصائية. وبصورة عامة تعتبر هذه الإحصاءات مناسبة لتؤهلها للإجابة عن أسئلة الاختبار والتحقق من صحة الفروض.

الجدول (٤): قيم بارامترات الصعوبة والتمييز والتخمين والمعلومات لكل من المجموعة الضابطة والتجريبية

المفردة	المجموعة الضابطة					المجموعة التجريبية				
	التمييز	الصعوبة	التخمين	المعلومات القصوى	متوسط المعلومات	التمييز	الصعوبة	التخمين	المعلومات القصوى	متوسط المعلومات
1	0.92	-2.43	0.25	0.13	0.06	1.001	-2.017	0.235	0.16	0.08
2	1.01	-0.65	0.21	0.17	0.13	1.515	-0.356	0.189	0.40	0.26
3	1.30	-0.98	0.21	0.28	0.18	0.883	-0.960	0.223	0.13	0.10
4	1.18	-1.85	0.21	0.23	0.11	1.070	-1.673	0.226	0.19	0.11
5	1.66	-0.50	0.18	0.49	0.30	1.676	0.271	0.176	0.50	0.29
6	1.38	-0.47	0.19	0.33	0.23	1.255	-0.643	0.207	0.26	0.19
7	0.91	0.64	0.18	0.15	0.10	1.409	0.438	0.136	0.38	0.24
8	1.18	-0.10	0.18	0.24	0.18	3.081	0.223	0.116	1.90	0.75
9	1.45	-0.31	0.21	0.35	0.24	1.500	-0.596	0.183	0.40	0.26
10	0.84	0.80	0.17	0.13	0.09	1.077	0.601	0.160	0.21	0.14
11	1.73	-0.08	0.17	0.54	0.33	0.935	0.916	0.141	0.17	0.11
12	0.88	-0.42	0.22	0.13	0.11	0.853	-0.838	0.233	0.12	0.10
13	1.32	-0.64	0.20	0.30	0.21	1.263	-1.067	0.210	0.27	0.17
14	1.28	0.72	0.17	0.30	0.17	1.214	0.555	0.189	0.26	0.16
15	0.75	-0.90	0.24	0.09	0.08	0.609	-0.147	0.219	0.06	0.05
16	1.36	-0.72	0.18	0.33	0.22	1.004	-1.176	0.221	0.16	0.12
17	1.35	-0.46	0.20	0.31	0.22	2.016	0.052	0.177	0.72	0.40
18	1.16	-0.78	0.22	0.22	0.16	1.061	-0.417	0.211	0.19	0.15
19	1.21	-0.51	0.26	0.2	0.18	0.985	0.483	0.185	0.17	0.12
20	1.19	-0.55	0.18	0.28	0.16	1.285	0.220	0.200	0.3	0.18

ثانيا التحقق من صحة الفروض

الفرض الأول:

ينص الفرض الأول على "**لا يؤثر إضافة بديل "لا أعرف" على تقدير قدرة الفرد**"، للتحقق من صحة هذا الفرض؛ تم حساب اختبار "ت" لدلالة الفرق بين متوسطي المجموعة الضابطة والتجريبية في درجاتهم على الاختبار التحصيلي. ويعرض جدول (٥) نتائج ذلك.

الجدول (٥): اختبار "ت" لدلالة الفرق بين متوسطي درجات التحصيل للمجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	حجم العينة	المتوسط	الانحراف المعياري	درجات الحرية	قيمة ت	الدلالة
الضابطة	٣٨	١٢,٥٨	٣,٦٩	٩٥	٠,٤٤	٠,٦٦
التجريبية	٥٩	١٢,٢٢	٤,٠٩			

يبين جدول (٥) السابق أنه لا توجد فروق دالة بين متوسطي المجموعتين الضابطة والتجريبية في درجاتهم على الاختبار التحصيلي. فطلاب المجموعة التجريبية تلقوا نفس صورة

الاختبار كما المجموعة الضابطة، وتم تحديد نفس المقدار الزمني لاختبار كل من المجموعتين، كما طبق على المجموعتين نفس عدد الأسئلة، غير أنه اختلفت نسخة اختبار المجموعة التجريبية فقط، في إضافة بديل خامس وهو "**لا أعرف**"، إذا اختار الطالب هذا البديل –مشيرا إلى عدم معرفته باختيار البديل الصحيح– تمنح له مكافأة (٠,٢٥) من الدرجة، بدلا من اختيار البديل الصحيح عشوائيا، وذلك بهدف خفض التخمين.

كما تم حساب دلالة الفرق بين متوسطي المجموعتين التجريبية والضابطة، في الاختبار التحصيلي، من خلال حساب تقدير القدرة باستخدام الأرجحية القصوى Maximum Likelihood ، في النموذج ثلاثي البارامتر لنظرية الاستجابة للمفردة، وذلك للتأكيد على النتيجة السابقة، فأظهرت النتائج أن المجموعة الضابطة (م=–٠,٥١، ع=٠,٨٠) كانوا أقل من المجموعة التجريبية (م=–٠,٣٣، ع=٠,٨٢) في تقدير القدرة في التحصيل الدراسي؛ ولكن لم تكن هناك فروق دالة بين المجموعتين، إذ كانت قيمة "ت" (٩٥)= ١,٠٧، عند مستوى دلالة (٠,٢٨).

لقد استفاد من اختيار "لا أعرف" حوالي نصف عدد المشاركين في المجموعة التجريبية – بدرجات متفاوتة– على النحو الذي سيشار إليه لاحقا، ومع ذلك لم تتأثر درجاتهم بصورة دالة عن المجموعة الضابطة، فلم تتضخم درجات المجموعة التجريبية من جراء زيادة المكافاة على درجاتهم الأصلية، كما لم تبخس درجاتهم لعدم معرفتهم بالإجابة الصحيحة عن المجموعة الضابطة أيضا.

وبالرغم من إتاحة الفرصة لاستفادة طلاب المجموعة التجريبية من اختيار "لا أعرف"؛ وبالرغم من استفادة نصف طلاب المجموعة التجريبية من هذه الخاصية (٥١%)؛ إلا أن هذه الاستفادة بالرغم من قوة ميزتها في التقليل من أثر التخمين، ومع ذلك لم ترتفع درجاتهم ارتفاعا دالا، فلم تكن هناك فروق دالة بين المجموعتين في درجات الاختبار التحصيلي.

وتشير النتائج إلى أنه لم يستفد جميع طلاب المجموعة التجريبية الذين منحوا اختيار "لا اعرف" من هذه الخاصية، بل إن نصفهم فقط الذي استخدم هذا الاختيار بمقادير متفاوتة، إن هذه النتيجة التي توصلت إليها الدراسة الحالية تعني منح الثقة للقائمين على الاختبارات، في إمكانية تخصيص مكافأة بمقدار (٠,٢٥) من الدرجة للطالب الذي اختار بديل "**لا اعرف**" لكل سؤال لا يعرفه، دون أن تتأثر درجاته الكلية، وذلك تجنبا للتخمين باختيار الإجابة الصحيحة عشوائيا، مما يؤثر على الخصائص السيكومترية للاختبار، وهذا يبعث على الاطمئنان في استخدام هذا الاختيار دونما شك في تضخم درجات الطالب أو بخسها.

ومن ثم تمت الإجابة عن السؤال الأول، وتحقق الفرض الأول من الدراسة، أن وضع بديل إضافي كبديل خامس على سبيل المثال، لبدائل اختبار الاختيار من متعدد، لا يؤثر على تقدير قدرة الطالب، أو درجاته في الاختبار التحصيلي، كما أنه لا توجد فروق بين الاختبار المكون من أربعة بدائل أو خمسة بدائل. وتتفق هذه النتيجة مع ما توصلت إليه دراسة Abad et al. (2001) ، ودراسة عبد الله الرشيدي، آمال الزعبي (٢٠١٩) أنه لا توجد فروق دالة بين الأربعة بدائل، والخمسة بدائل في تقدير القدرة.

ولكن يتبقى أسئلة حول الحدود العليا للمكافأة التي يمكن أن يمنحها إضافة هذا البديل؟ لقد حسبت الدراسة الحالية أيضا إمكانية إضافة مكافأة بمقدار (٠,٣٣) من الدرجة، فأوضحت النتائج أيضا أنه لم تكن هناك فروق دالة بين المجموعتين، إذ كانت "ت" (٩٥)= ٠,١٥، عند مستوى دلالة (٠,٨٨)، بالرغم أن المجموعة التجريبية (م=١٢,٧٠، ع=٣,٨٦) كانوا أعلى من المجموعة الضابطة (م=١٢,٥٨، ع=٣,٦٩) في درجاتهم على التحصيل الدراسي، مما يعني أن منح (٠,٣٣) من الدرجة لا يؤثر في تقدير قدرة الطالب.

وتطرح هذه النتيجة تساؤلات فيما يتعلق "بسقف المكافأة" التي يمكن أن تمنح للطالب في اختيار بديل "**لا أعرف**"، والعدد المسموح به لاختيار هذا البديل. لقد كان عدد أسئلة الاختبار الحالي هو (٢٠ سؤالا)، ولكن ماذا عن التفاوت في أعداد الأسئلة للاختبارات الأخرى، وما أثر ذلك على تقدير قدرة الطالب؟ وهل يختلف هذا الأثر باختلاف عدد الأسئلة المخصصة لكل اختبار؟ تحتاج هذه التساؤلات إلى دراسات أخرى للإجابة عنها.

الفرض الثاني:

ينص الفرض الثاني على "**يخفض إضافة بديل "لا أعرف" بارامترات التخمين لمفردات الاختبار**".

للتحقق من صحة هذا الفرض؛ قام الباحث بحساب متوسطي بارامتر التخمين Asymptote parameter ، في النموذج ثلاثي البارامتر في نظرية الاستجابة للمفردة لكل من مفردات المجموعتين التجريبية والضابطة، وحساب دلالة الفرق بين المتوسطين، ويعرض جدول (٦) نتائج هذه الدلالة.

ا

لجدول (٦): اختبار "ت" لدلالة الفرق بين متوسطي مؤشرات التخمين لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	حجم العينة	المتوسط	الانحراف المعياري	درجات الحرية	قيمة ت	الدلالة
الضابطة	٣٨	٠,٢١١	٠,٠٢٣	٩٥	٣,١٨	٠,٠٠١
التجريبية	٥٩	٠,١٩١	٠,٠٣٤			

يبين جدول (٦) السابق أن هناك فروق دالة بين المجموعتين التجريبية والضابطة في متوسط مؤشرات التخمين، إذ أن المجموعة التجريبية التي استفادت من خاصية "**لا أعرف**" قد حازت على متوسط مؤشرات التخمين أقل بصورة دالة من المجموعة الضابطة، وكلما كان متوسط مؤشرات التخمين أقل؛ كان ذلك أفضل كما أشير إلى ذلك في الموضع الخاص بالأساليب الإحصائية. ويعزى هذا الفرق الدال إلى استفادة طلاب المجموعة التجريبية من خاصية لا اعرف في اختبارهم، إذ أن الطلاب الذين اختاروا البديل الخامس "لا اعرف" حصلوا على مكافأة لم تؤثر في درجاتهم، وفي ذات الوقت لم يختاروا بديلا عشوائيا قد يخصم من درجاتهم، ففضلوا "القليل المضمون عن الكثير غير المضمون"، وهذا الأمر ساعد كثيرا في خفض مؤشرات التخمين لدى مفردات المجموعة التجريبية. وبذلك تمت الإجابة عن السؤال الثاني، وتحقق الفرض الثاني من الدراسة، أن إضافة بديل "**لا أعرف**" يخفض بارامترات التخمين لمفردات الاختبار

وتفسر هذه النتيجة على أن المتعلمين في المجموعة التجريبية تحديدا إما اختاروا الإجابة الصحيحة، أو استفادوا من بديل "**لا اعرف**" باختيارهم إياه، عندما لا يعرفون الإجابة الصحيحة، فتفادوا التخمين العشوائي إلى حد كبير، وربما مارسوا بعض التخمين الذكي، فكانت النتيجة هي انخفاض بارامتر التخمين لديهم مقارنة بالمجموعة التجريبية. وهذا التفسير يؤيد ما ذهب إليه صلاح علام (١٩٩١) و(1990) Choppin و(2018) Frey من أن اللجوء إلى التخمين الذكي يتطلب مكافأة عليه للمتعلم، وليس العقاب من خلال استقطاع بضع درجات من درجته الإجمالية، وهو ما تحقق عندما منح المتعلم ربع درجة مضمونة إذا لم تشكك في الإجابة الصحيحة.

والتفسير الثاني لانخفاض بارامتر التخمين للمجموعة التجريبية مقارنة بالمجموعة الضابطة، يتمثل في سلوك المخاطرة Risk– Taking أو أسماه Crocker and Algina (2008) "الاستعداد للمقامرة" willingness to gamble ويرى Chiu (2010) أن المتعلمين الأكثر عرضة للمخاطرة هم الأكثر استعداداً للتخمين. بل إن بعض الباحثين (Betts, Elder, Hartley, & Trueman, 2009) يرون أن الإجراءات المستخدمة في اختبارات الاختيار من متعدد تقيس سلوك المخاطرة أكثر من قياسها لمعلومات محددة.

فسلوك المخاطرة هو شكل من أشكال استشفاف الاختبار testwiseness التي يمكن أن تؤثر بشكل مباشر على درجات الممتحنين. فالمتعلمون في المجموعة الضابطة لم يكن لديهم خيار آخر غير محاولة اختيار الإجابة الصحيحة ولو بصورة عشوائية، مما أوقعهم في البدائل الخطأ، على أمل أن تلعب الصدفة لديهم دورا في صالحهم، فسلوك المخاطرة لديهم كان مرتفعا، وكان متوسط مؤشرات التخمين لديهم مرتفعة أيضا بصورة دالة مقارنة بالمجموعة التجريبية، التي يحاول فيها المتعلمون معرفة الإجابة الصحيحة، أو اختيار بديل "**لا اعرف**"، مما قلّص لديهم الاستعداد للمخاطرة، فربما يكون تخمينهم كان قائما على المعرفة الجزئية كما ذكر ذلك عودة (٢٠١٤) وهشام الخولي (٢٠١٨)، فنتيجة لذلك كان بارامتر التخمين لديهم أقل بصورة دالة، وربما كان قلق الاختبار لديهم أيضا منخفضا، لقناعتهم أنهم سيحصلون على أجزاء من الدرجات، عندما لا يعرفون الإجابة الصحيحة باختيارهم بديل "لا اعرف" فانخفض لديهم مؤشر التخمين. فهم يوازنون بين البديل الذي يرونه (يحتمل) أن يكون صحيحا، وبين اختيار بديل مضمون بقيمة أقل وهو بديل "لا اعرف"، فهذه الموازنة على الأرجح ساهمت في خفض مؤشرات التخمين للمجموعة التجريبية. أما من لم يستخدم هذا البديل؛ فربما استخدموا التخمين العشوائي، وليس التخمين الذكي، ولذلك انتقصت درجاتهم مقارنة بمن استخدموا هذا البديل الإضافي.

التفسير الثالث لانخفاض بارامتر التخمين للمجموعة التجريبية بصورة دالة مقارنة بالمجموعة الضابطة هو المكافأة كما يدركها المتعلم، وليس العقاب. فلقد وجد Avila and Torrubia (2004) في تحليلاتهما البعدية على (١٩ دراسة سابقة) أن العقوبات بالمعادلات من خلال استقطاع الدرجات من الممكن أن تؤثر على إجابات الطلاب، حتى لو كانوا على دراية وفهم بمغزى التعليمات. فربما يكون إدراك المتعلم حصوله على مكافأة إضافية لاختياره بديل "لا اعرف"، حتى وإن لم يعي إلى أن هذا الإجراء، "يسحبه" من دائرة التخمين العشوائي إلى التخمين الذكي؛ إلا أنه خفض من بارامتر التخمين للمجموعة التجريبية.

غير أن هذه التفسيرات تحتاج إلى مزيد من الدراسات حول دور كل من قلق الاختبار، واستشفاف الاختبار، وسلوك المخاطرة إذ تتطلب تطبيق مقاييس لهذه المتغيرات مصاحبة للاختبار التحصيلي وهو ما لم يتحقق في الدراسة الحالية.

الفرض الثالث:

ينص الفرض الثالث على "**يحسن إضافة بديل "لا أعرف" بارامترات المفردات وكفاءة الاختبار**" وللتحقق من صحة هذا الفرض؛ تم حساب دلالة الفرق بين متوسطي مؤشرات الصعوبة،

والتمييز، ودالة معلومات الاختبار، والخطأ المعياري لتقدير القدرة، لكل من مفردات المجموعة التجريبية والمجموعة الضابطة. وتعرض الجداول (٧-١٠) دلالة الفرق بين المجموعتين في تلك المؤشرات.

الجدول (٧): اختبار "ت" لدلالة الفرق بين متوسطي مؤشرات الصعوبة لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	ن	م	ع	د. ح.	ت	الدلالة
الضابطة	٣٨	٠,١٢٦٦	١,٤٧١٨	٩٥	٠,١٦	٠,٨٨
التجريبية	٥٩	٠,١٧٠٥	١,٢٦٨١			

الجدول (٨): اختبار "ت" لدلالة الفرق بين متوسطي مؤشرات التمييز لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	ن	م	ع	د. ح.	ت	الدلالة
الضابطة	٣٨	١,٢٠٣	٠,٢٧٧	٩٥	٢,٢١	٠,٠٣
التجريبية	٥٩	١,٢٨٥	٠,٠٥٤٥			

الجدول (٩): اختبار "ت" لدلالة الفرق بين متوسطي مؤشرات دالة المعلومات لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	ن	م	ع	د. ح.	ت	الدلالة
الضابطة	٣٨	٣,١١٧	٠,٠٧٧	٩٥	٢٤,٣٢	٠,٠٠١
التجريبية	٥٩	٣,٧٩٣٤	٠,١٥٩٣			

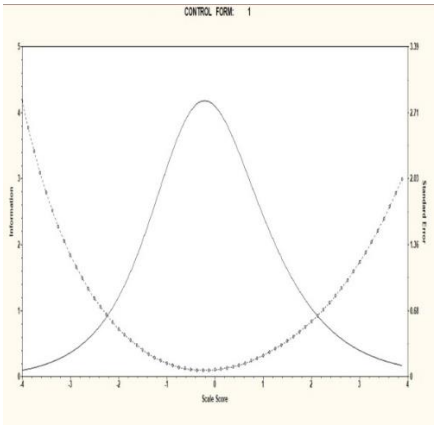
الجدول (١٠): اختبار "ت" لدلالة الفرق بين متوسطي مؤشرات الخطأ المعياري لتقدير القدرة لكل من المجموعتين الضابطة والتجريبية

الإحصاءات / المجموعة	ن	م	ع	د. ح.	ت	الدلالة
الضابطة	٣٨	٠,٥٩٣	٠,١٦٧	٩٥	٣,١٣١٧	٠,٠٠٢
التجريبية	٥٩	٠,٤٩٨٢	٠,١٣			

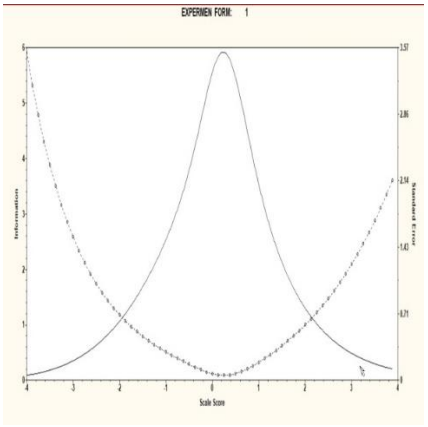
يبين جدول (٧) السابق أنه لم تكن هناك فروق دالة بين المجموعتين في متوسطي مؤشرات الصعوبة Threshold في ضوء نظرية الاستجابة للمفردة. كما يبين جدول (٨) السابق، أن هناك فروقا دالة بين المجموعتين الضابطة والتجريبية في متوسطي مؤشرات تمييز المفردات Slope لصالح المجموعة التجريبية. وتظهر النتائج أيضا -كما يعرضها جدول (٩)- أن متوسط مؤشرات دالة معلومات الاختبار للمجموعة التجريبية كانت أعلى بصورة دالة من متوسط معلومات المجموعة الضابطة، ويعرض جدول (١٠) السابق أن الخطأ المعياري لتقدير القدرة كان أقل بصورة دالة لدى المجموعة التجريبية عن المجموعة الضابطة.

ومن ثم؛ تظهر النتائج –في مجملها– أنه لم تكن هناك فروق دالة بين المجموعتين في بارامتر الصعوبة، بينما كانت هناك فروق دالة في بارامترات التمييز، ودالة معلومات الاختبار، والخطأ المعياري لتقدير القدرة، وقبل ذلك بارامتر التخمين، بين المجموعتين الضابطة والتجريبية. فقد كانت المجموعة التجريبية هي الأفضل بصورة دالة من المجموعة الضابطة في متوسط تلك البارامترات. ومن ثم فقد تمت الإجابة عن السؤال الثالث، كما تحقق الفرض الثالث في الدراسة الحالية.

وتعزى هذه النتيجة إلى إضافة بديل "**لا أعرف**" لدى أفراد المجموعة التجريبية، والذي استخدمه عدد منهم عندما لا يعرفون الإجابة الصحيحة، مما خفض من مؤشرات تخمين المفردة بصورة دالة، كما اتضحت من نتائج الفرض الثاني، وهذا بدوره قد حسّن من تمييز المفردات لدى طلاب المجموعة التجريبية، بالإضافة إلى ذلك؛ فإن انخفاض مؤشرات التخمين لدى المجموعة التجريبية قد حسّن من أداء الاختبار أيضا، كما تجلى من متوسط معلومات الاختبار، إذ كانت أعلى بصورة دالة من متوسط المعلومات للمجموعة الضابطة، وانخفض بالتبعية مؤشرات الخطأ المعياري لتقدير القدرة، لدى المجموعة التجريبية عن المجموعة الضابطة وبصورة دالة أيضا كما يوضح ذلك شكلا (١) و (٢).



شكل (١): مقدار معلومات الاختبار والخطأ المعياري للقياس للمجموعة الضابطة



شكل (٢): مقدار معلومات الاختبار والخطأ المعياري للقياس للمجموعة التجريبية

لقد كانت قيمة ثبات ألفا كرونباخ للمجموعة الضابطة (٠,٧٦)، بينما كان ثبات المجموعة التجريبية بطريقة ألفا كرونباخ هو (٠,٨١). إن انخفاض التخمين –كما تبين في نتائج الفرض الثاني– قد حسّن من كفاءة الاختبار بصفة عامة. وتتفق هذه النتيجة مع ما أثبته Paek (2015)

أنه فعندما تكون الصعوبة مساوية لقدرة الفرد في النموذج ثلاثي البارامتر –كما هو الحال في الدراسة الحالية– فإن بارامتر التخمين ينخفض، ومن ثم تتحسن بارامترات تمييز المفردات، وكذلك دالة معلومات الاختبار، والخطأ المعياري لتقدير القدرة. ويؤكد Hambleton et al. (1991) أن المفردات ذات التخمين المنخفض والتمييز العالي تقدم مقدارا من المعلومات أعلى من المفردات ذات التخمين المرتفع والتمييز المنخفض.

وبالرغم أن المجموعة التجريبية قد أتيح لها الاستفادة من بديل "**لا أعرف**"؛ إلا أن عدد الأفراد الذين استفادوا من هذا البديل (٣٠ من ٥٩ متعلما) من المجموعة التجريبية، أي نصف عددهم فقط بنسبة (٥١%)، وذلك بتكرارات متفاوتة، تراوحت من مرة واحدة (أي اختيار الخاصية لسؤال واحد فقط)، نفذها (١٣ متعلما)، إلى ثماني مرات كحد أقصى قام بها متعلمان، وكان مجموع الاستفادات لجميع المتعلمين (٨٩ استفادة)، ويوضح جدول (١١) الإحصاءات الوصفية للاستفادة من هذا البديل.

الجدول (١١): الإحصاءات الوصفية للمستفيدين من بديل "لا أعرف" في المجموعة التجريبية

العدد	تكرار الاستفادات	النسبة
١٣	١	%٤٣,٣
٣	٢	%١٠
٦	٣	%٢٠
١	٤	%٣,٣
٢	٥	%٦,٧
٣	٦	%١٠
٢	٨	%٦,٧
٣٠	٢٩	%١٠٠

تثير هذه النتيجة سؤالا حول مستوى تحصيل الطلاب الذي يستفيدون من هذه الخاصية. لقد أوضحت النتائج الحالية أنه توجد فروق دالة بين متوسطي درجات الذين استفادوا من خاصية "**لا أعرف**" قبل منحهم المكافأة، وغير المستفيدين من الخاصية، من ذات المجموعة التجريبية، فالمستفيدون من استخدام بديل "**لا أعرف**" (م=١١,٠٣، ع=٣,٥٨) كانوا أقل بصورة دالة من الطلاب غير المستخدمين للبديل (م= ١٣,٤٥، ع=٤,٠٣) في درجاتهم في التحصيل الدراسي. إذ كانت "ت" (٥٧)= ٢,٣٦، عند مستوى دلالة (٠,٠٢). أي أن الأضعف تحصيليا هو المستفيد من خاصية اختيار "لا عرف"، بالرغم أنها لم تنقل مستوى الطلاب من الضعيف إلى الأعلى، أو المتوسط إلى المرتفع.

ويمكن تفسير ذلك على أن سلوك المتعلمين ليس متسقا حتى مع وجود تعليمات واضحة للاختبار، فمشكلة التخمين تكمن في أنه لا يكشف عن قدرة المتعلم تحديدا، ولكنه مزيج من قدرة

المتعلم وصعوبة المفردات معا ،كما يشير إلى ذلك Bond et al. (2020) ، إذ يرى أن هناك من المتعلمين من لا يخمن على الإطلاق، فهم يحاولون دائما التوصل إلى الإجابة الصحيحة، بينما يلجأ البعض الآخر إلى ممارسة التخمين عندما تكون لديه القدرة على ذلك (Bond et al., 2020) .

لقد أسفرت دراسة Grant and McGrath (2021) التي أجريت للتعرف على تأثير التعليمات المقدمة من المدرسين أثناء عملية التعلم، وأثناء تطبيق اختبارات الاختيار من متعدد أن كثيرا من الطلاب لا ينتبهون للتعليمات المقدمة، وأنه من الممكن للطلاب التعرف على التعليمات دون تنبيه من المعلم.

الخلاصة والتضمينات:

بينت الدراسة الحالية أن إضافة بديل "**لا أعرف**" لبدائل اختبارات الاختيار من متعدد يخفض مؤشر التخمين لدى الطلاب، ويحسن من الخصائص السيكومترية للاختبارات، مثل تمييز المفردات ومقدار المعلومات، ومعامل الثبات، بالإضافة إلى أنه يخفض من الخطأ المعياري لتقدير القدرة، وذلك في ضوء تحليل البيانات بالنظرية التقليدية، ونظرية الاستجابة للمفردة معًا. لقد تحققت أهداف الدراسة الحالية، إذ أظهرت النتائج أن استخدام هذا البديل مع بدائل الاختيار من متعدد الأخرى، يحسّن من كفاءة الاختبار، وفي ذات الوقت لا يؤثر على تقدير القدرة من حيث تضخيم درجات المتعلم.

إن مكافأة المتعلم بمنحه أجزاء من الدرجات –في الدراسة الحالية (٠,٢٥) من الدرجة– يقلل من فرصته من التخمين العشوائي مما يعمل على خفض بارامترات التخمين للاختبار، ومن ثم يتحسن بارامتر التمييز، وبالتالي تزيد دالة معلومات الاختبار، ويتقلص الخطأ المعيار لتقدير القدرة، مع ارتفاع قيمة معاملات الثبات. إن نتائج الدراسة الحالية تسلط الضوء على ضرورة الاهتمام بقضية التخمين العشوائي للمتعلمين في الاختبارات الجامعية المختلفة، وكيفية معالجتها بعيدا عن معادلات استقطاع درجات لتصحيح التخمين، لما لها من آثار جانبية ذات أبعاد تربوية ونفسية.

لقد استخدمت الدراسة الحالية (ربع الدرجة) إذا اختار بديل "لا اعرف"، لكن من غير الواضح تحديد "السقف المناسب" للمكافأة، ولذلك ربما دراسات أخرى لاحقة تحدد أجزاء أخرى من الدرجات مثل (٠,٣٣) أو (٠,٥٠) بحيث لا تؤثر على تقدير القدرة، وتقلص مؤشرات التخمين أكثر.

لقد أجريت الدراسة الراهنة على مجموعة صغيرة إلى حد ما (٩٧ طالبا) مقسمين إلى مجموعتين (ضابطة وتجريبية). إنه من المأمول في دراسات لاحقة أن تتم إعادة هذه الدراسة على عينات كبيرة نسبيا (أكبر من ٢٠٠ طالبا) حتى تكون النتائج أكثر وضوحا وتباينا.

تثير نتائج الدراسة الحالية التساؤلات حول العوامل التي دفعت نصف عدد المتعلمين من عدم الاستفادة من هذا البديل، وما هي الدوافع التي تدفعهم إلى الاستفادة، أو تمنعهم منها، وماهي المتغيرات النفسية التي يمكن أن ترتبط بمدى الاستفادة من استخدام هذا البديل مثل استشفاف الاختبار، والميل للمخاطرة، وقلق الامتحان وغيرها من المتغيرات ذات الصلة. ربما يكون عدم ثقة المتعلمين في مدى منح هذه الدرجات له دور في إحجامهم عن الاستفادة من بديل لا "أعرف"، أو ربما عدم فهمهم للتعليمات المقدمة، فقد أسفرت دراسة Grant and McGrath (2021) التي أجريت للتعرف على تأثير التعليمات المقدمة من المدرسين على عملية التعلم وتطبيق اختبارات الاختيار من متعدد، أن كثيرا من الطلاب لا ينتبهون للتعليمات المقدمة، وأن من الممكن للطلاب التعرف على التعليمات دون تنبيه من المعلم. إنه من المرجو في دراسات لاحقة الإجابة عن تلك التساؤلات.

لقد أجريت هذه الدراسة على الطلاب الذكور، وربما لو تم تطبيقها على الإناث لكشفت عن نتائج أخرى مقارنة بالذكور في استخدامهن لهذا البديل. إن من قيود الدراسة الحالية أن عدد مفردات الاختبار كان (٢٠ مفردة)، وهو عدد محدود إلى حد ما، إنه من المرجو في دراسات لاحقة يتم إضافة بديل "لا اعرف" على اختبارات تحصيلية من نمط الاختيار من متعدد تتألف من (٥٠ مفردة) أو أكثر حتى تتضح النتائج أكثر.

تستحق معالجة سلوك التخمين في أسئلة الاختيار من متعدد الالتفات من جانب القائمين على الاختبارات التحصيلية، وذلك لبناء الثقة في الدرجات الممنوحة للطلاب، واتخاذ قرارات صحيحة بالنسبة لهم.

المراجع العربية

أحمد سليمان عودة (٢٠١٤). تصحيح معاملات صعوبة الفقرات لأثر التخمين في أسئلة الاختيار من متعدد: صورة معدلة لمعادلة جيلفورد. *مجلة الدراسات التربوية والنفسية: جامعة السلطان قابوس، مجلد ٨*، ع٢، ٢٤٨-٢٥٧.

حسين عبد الله الزهراني (٢٠٢١). *أثر عدد البدائل لاختبار الاختيار من متعدد على تقدير قدرة الفرد ومعالمِ المفردة في ضوء كل من النظرية التقليدية والنموذج ثلاثي المعلمة في نظرية الاستجابة للمفردة*، رسالة دكتوراه غير منشورة، قسم علم النفس، كلية التربية، جامعة الملك سعود.

رجاء محمود أبو علام (١٩٨٧). *قياس وتقويم التحصيل الدراسي*، الكويت، دار القلم.

شادية عبدالعزيز منتصر، وحسان مخلوف خلاف (٢٠١٧). الكشف عن التخمين باستخدام نموذج راش وعلاقته بخصائص الفرد والمفردة. *دراسات تربوية واجتماعية: جامعة حلوان – كلية التربية*، مج٢٣ , ع٣ ، ١٢٢٣ – ١٣٠٨.

صفوت فرج (١٩٨٩). *القياس النفسي*، القاهرة، الأنجلو المصرية.

صلاح الدين محمود علام (١٩٩١). أثر المواقف الاختبارية جماعية المرجع، ومحكية المرجع في مستويات التخمين وسلوك المخاطرة والأداء في علم النفس: دراسة تجريبية. *علم النفس: الهيئة المصرية العامة للكتاب، س ٥، ع ١٨–١٩*، ٩٨–١٠٧.

صلاح الدين محمود علام (٢٠١٥). *القياس والتقويم التربوي والنفسي: أساسياته وتطبيقاته وتوجهاته المعاصرة*، القاهرة، دار الفكر العربي.

عبدالله مسند الرشيدي، وآمال أحمد الزعبي (٢٠١٩). أثر عدد البدائل في اختبار الاختيار من متعدد في الخصائص السيكومترية للاختبار التحصيلي وفقراته وفق النموذجين الثلاثي والرباعي المعلمة في نظرية الاستجابة للفقرة. *إربد للبحوث والدراسات الإنسانية: جامعة إربد الأهلية – عمادة البحث العلمي، مج٢١ ،ع٢*، ٤٢–٨٣.

محمد حسين فهيد اليامي (٢٠١٦). بعض طرق تصحيح التخمين وأثرها على خاصيتي الصدق والثبات لاختبار الاختيار من متعدد. *عالم التربية: المؤسسة العربية للاستشارات العلمية وتنمية الموارد البشرية، س١٧*، ع٥٣، ١–٣١.

ميساء محمود أحمد صمصم (٢٠١٢). *أثر طرق التصحيح لضبط التخمين في الاختبارات الاختيار من متعدد على كل من مؤشر ثبات الفقرة ومؤشر صدقها*، رسالة ماجستير غير منشورة، جامعة مؤتة، الأردن.

هشام محمد الخولي (٢٠١٨). أثر برنامج تدريبي لمهارات الحكمة الاختبارية على الأداء والتخمين والمخاطرة والحذر في مواقف اختبارية معيارية المرجع ومحكية المرجع لدى طلاب الجامعة. *مجلة كلية التربية: جامعة بنها – كلية التربية، مج٢٩*، ع١١٥ ، ٢٨٢-٣٤٦.

وسام عطية الشرفا (٢٠١٦). *أثر طريقة التصحيح لضبط أثر التخمين على مقدار التضخم المحسوب في قيمة معامل ارتباط الفقرة*، رسالة ماجستير غير منشورة. جامعة مؤتة، الأردن.

المراجع الأجنبية

Abad, F. J., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, *13*(1), 152-158.

American Psychological Association. (2015). *APA dictionary of psychology*. (2th[ed].), Washington, DC: American Psychological Association.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26*, 323-336.

Avila, C., & Torrubia, R. (2004). Personality, expectations, and response strategies in multiple-choice question examinations in university students: A test of Gray‘s hypothesis. *European Journal of Personality, 18*, 45-59.

Azevedo, J. M., Oliveira, E., & Beites, P. (2019). Using Learning Analytics to Evaluate the Quality of Multiple-Choice Questions: A Perspective with Classical Test Theory and Item Response Theory. *International Journal of Information and Learning Technology, 36*, 322-341.

Baldwin, P. (2020). A problem with the bookmark procedure's correction for guessing. *Educational Measurement: Issues and Practice, 40*, 7-15.

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education, 34*(1), 1–15.

Bond, T., Yan, Z. & Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th Ed.). New York: Routledge.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30*(4), 277–291.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1-19.

Chiu, T. (2010). *Correction for guessing in the framework of the 3PL item response theory* (Order No. 3418414). Available from ProQuest Dissertations & Theses Global. (751246969).

Choppin, B. H. (1990). Correction for guessing. In H. J. Walberg and G. D. Haertel, (Eds.), *The international encyclopedia of educational evaluation* (pp. 345-348). Oxford: Pergamon Press.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York: Cengage Learning.

Daoust, M K., Côté-Bouchard, C. (2021). Epistemic Consequentialism, Veritism, and Scoring Rules. *Erkenn* https://doi.org/10.1007/s10670-021-00426-5

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Newbury Park, CA: Sage.

Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications, Inc.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082–1116.

Grant, N. K., & McGrath, A. L. (2021). Effects of PowerPoint slides on attendance and learning: If you share it, they will (still) come. *Scholarship of Teaching and Learning in Psychology*. Advance online publication. https://doi.org/10.1037/stl0000241

Haladyna, T. M. (2004). *Developing and Validating Multiple-choice Test Items* (3rd ed.). Routledge.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Item response theory: Principles and applications* (2nd ed.). Boston: Kluwer-Nijhoff.

Hejri, S. M., Mafinezhad, M. K., and Jalili, M. (2014). Guessing in Multiple Choice Questions: Challenges and Strategies. *Iranian Journal of Medical Education, 14*, 594-604.

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp. [Computer software]

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *The Journal of Technology, Learning and Assessment, 4*(2), 1–23.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal Of Selection And Assessment*, *18*(1), 111-115**.**

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation, 39*, 188-193.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*(1), 7–11.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students*. (6th[ed].), Boston, MA: Pearson/Allyn & Bacon.

Nunnally, J. C. (1978). *Psychometric theory* (2nd [ed].). New York: McGraw-Hill.

Paek I. (2015). An Investigation of the Impact of Guessing on Coefficient α and Reliability. *Applied psychological measurement*, *39*(4), 264–277.

Parkes, J., & Zimmaro, D. (2016). *Learning and Assessing with Multiple-Choice Questions in College Classrooms* (1st ed.). Routledge.

Plumlee, L. B. (1952). The effect of difficulty and chance success on item-test correlation and on test reliability. *Psychometrika, 17*, 69-85.

Traub, R. E., Hambleton, R. K., & Singh, D. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29,* 847-862.

Weiss, D. & Yoes, M. (1994). Item response theory. In R. K. Hambleton, & J. N. Zaal. (eds.), *Advances in educational and psychological testing : theory and applications.* (pp. 69-95) Boston: Kluwer academic.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*, 357-371.

Zimowski, M. E, Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Chicago: Scientific Software.

The effect of adding an alternative "I don't know" in Multiple-Choice Questions on Estimating Ability of the individuals, minimizing Guessing Parameters and improving Test Efficiency in based on the Item Response Theory

Amin Sabry Nour-Eddin
Associate Professor of Educational Psychology
Faculty of Education, Ain Shams University

Abstract:

Guessing behavior is an important topic on multiple choice tests, due to greater the potential for systematic error or bias which that inflates observed test scores. The purpose of this research is to determine the effect of adding an alternative "I don't know" in multiple-choice questions on reducing guessing parameters, improving item parameters, efficiency of exam, and estimating of ability, based on item response theory. The study sample consisted of 97 male students, their ages ranged from 21 to 22 years with an average age of 21.26, and a standard deviation 0.44, they were divided into two groups, one of them is a control group 38 students, and the other is an experimental group 59 students. The equivalence of the two groups in achievement ability was verified. The instrument is an achievement test consisted of 20 multiple-choice items, with 4 alternatives each. Each item was scored as correct or incorrect. A fifth alternative "I don't know" has been added, for the exam of the experimental group. It is given to whoever chooses it (1/4) a score, to be added to his total score to verify the reduction of the guessing effect. The results showed that there was no significant difference for estimation ability, $t(95)=0.44$ $p= 0.66$ despite the control group (M = 12.58, SD = 3.69) attaining higher score than the experimental group (M = 12.22, SD = 4.09). The results also showed that adding an "I don't know" alternative reduced the guessing parameters for the test items. The experimental group who used an "I don't know" alternative (M = 0.19, SD = 0.03) was significantly higher than the control group (M = 0.21, SD = 0.02). "t" (95) = 3.08, $p= 0.001$. The results also revealed that adding the "I don't know" alternative to the experimental group and using it in the test improved the discrimination parameters. The experimental group (M = 1.29, SD = 0.55) was significantly higher than the control group (M = 1.20, SD = 0.28) in the mean of discrimination, "t" (95) = 2.21, p= 0.03. The results showed that the efficiency of the exam for the experimental group (M = 3.79, SD = 1.16) was significantly higher than the control group (M = 3.12, SD = 0.08) in the average amount of test information, "t" (95) = 24.32, $p= 0.001$. The results also revealed that the experimental group (M = 0.50, SD = 0.13)

was significantly less than the control group (M = 0.59, SD = 0.17), in the mean standard error of estimating ability, "t" (95) = 3.13, p= 0.002. The study concluded with some educational implications, and suggested related research.

The keywords: Multiple-Choice Test - Guessing - "I don't know" alternative- Three-Parameter Logistic Model -Information function- estimation of ability- Standard Error of Estimation