HNSJ Humanities & Natural Sciences Journal
ISSN: (e) 2709-0833
www.hnjournal.net

مجلة العلوم الإنسانية والطبيعية
مجلة علمية محكمة (التصنيف: NSP)
معامل التأثير العربي للعام 2024 = 0.44

عنوان البحث

أثر حجم العينة في تقدير الثبات بطرق الاختبار وإعادة الاختبار وألفا كرونباخ

د. عبدالاله محمد القرني[1]

[1] أستاذ القياس والتقويم والإحصاء المشارك، جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية.

HNSJ, 2025, 6(8); https://doi.org/10.53796/hnsj68/26

المعرف العلمي العربي للأبحاث: https://arsri.org/10000/68/26

تاريخ الاستقبال: 2025/07/07م تاريخ القبول: 2025/07/15م تاريخ النشر: 2025/08/01م

المستخلص

هدفت هذه الدراسة إلى تحليل أثر حجم العينة على تقديرات الثبات باستخدام طرق الاختبار وإعادة الاختبار ومعامل ألفا كرونباخ. تم تطبيق مقياس الانتماء الوطني على عينة مكونة من 500 مشارك، وأُجري التحليل باستخدام عينات فرعية بأحجام مختلفة تتراوح من 20 إلى 500. أظهرت النتائج أن العينات الصغيرة (أقل من 50) لم تعطي تقديرات مستقرة للثبات، بينما كانت التقديرات أكثر قوة واستقرارًا عندما تم استخدام 100 فرد أو أكثر. كما أظهرت فترات الثقة باستخدام Fisher دقة أعلى من التقديرات النقطية. توصي الدراسة باستخدام عينات لا تقل عن 100 مشارك لضمان الحصول على تقديرات ثبات دقيقة، بالإضافة إلى أهمية استخدام تقدير الثبات كفترة زمنية بدلاً من استخدامه كقيمة مفردة.

الكلمات المفتاحية: الثبات، حجم العينة، إعادة التطبيق، معامل كرونباخ ألفا.

RESEARCH TITLE

The Effect of Sample Size on the Estimation of Reliability Using Test-Retest Methods and Cronbach's Alpha

Abdulelah Mohammed Alqarni[1]

[1] Associate Professor of Measurement, Evaluation, and Statistics, King Abdulaziz University, Jeddah, Saudi Arabia.

HNSJ, 2025, 6(7); https://doi.org/10.53796/hnsj68/26
Arabic Scientific Research Identifier: https://arsri.org/10000/68/26

Received at 07/07/2025 **Accepted at 15/07/2025** **Published at 01/08/2025**

Abstract

This study aimed to analyze the effect of sample size on reliability estimates using test-retest methods and Cronbach's alpha. The "National Belonging" scale was applied to a sample of 500 participants, and the analysis was conducted using subsamples of varying sizes ranging from 20 to 500. The results revealed that small samples (less than 50) yielded unstable reliability estimates, while robust and stable estimates emerged when using 100 or more individuals. Fisher's confidence intervals also demonstrated higher accuracy than point estimates. The study recommends using samples of at least 100 participants to ensure accurate reliability estimates and emphasizes the importance of using reliability estimates as an interval rather than as a single value.

Key Words: Reliability, Sample Size, Test-Retest, Cronbach's Alpha.

مقدمة

يُعد تقدير الصدق و الثبات في أي بحث أمر بالغ الأهمية، ولتحقيق هدف أي البحث فإننا نُواجه عادةً مسألتين: الأولى: كيف نتأكد من أننا نقيس بالفعل ما نريد قياسه؟ والثانية: هل نحن متأكدون من أننا سنحصل على نفس النتيجة إذا كررنا القياس؟ يتعلق السؤال الأول بقضايا الصدق، بينما يتعلق السؤال الثاني بالثبات. ويُشار إلى هذين المفهومين بالخصائص السيكومترية.

يشير مصطلح الثبات في البحوث النفسية إلى اتساق دراسة بحثية أو اختبار قياس (2007,McLeod). إذا أمكن تكرار نتائج البحث باستمرار، فإنها تُعتبر ثابتة، وموثوقة. في أغلب الأحيان، قد لا يكون الحصول على نفس النتائج ممكنًا نظرًا لاختلاف المشاركين والمواقف. ومع ذلك، إذا وُجد ارتباط إيجابي قوي بين نتائج الاختبار نفسه، فهذا يُشير إلى الثبات والموثوقية (2017,Balkin).

تتعدد تعريفات الثبات في أدبيات القياس النفسي. ووفقًا لويلكينسون وروبرتسون (2006)، فإن الثبات فيما يتعلق بالبحث يعني إمكانية التكرار أو الاتساق. كما يُمكن تعريف الثبات بأنه الدرجة التي تُنتج بها أداة التقييم نتائج مستقرة ومتسقة (Meyer, 2010). من جانبه، Mellenbergh, (2011) أشار إلى أن الثبات هو اتساق الاختبار، أو مدى اتساق نتائجه حيث يُعد مقياسًا لدقة وموثوقية الاختبار. الثبات هو مدى إعطاء التجربة أو الاختبار أو أي إجراء قياس نفس النتيجة عند تكرار التجربة أو القياس.

وفقًا للمجلس الوطني للقياس في التعليم (NCME; 1999)، يُعرّف الثبات في الإحصاء والقياس النفسي بأنه الاتساق العام للمقياس. ويُقال إن المقياس يتمتع بثبات وموثوقية عالية إذا أسفر عن نتائج متشابهة في ظل ظروف متشابهة. وهي سمة لمجموعة من درجات الاختبار تتعلق بمقدار الخطأ العشوائي الناتج عن عملية القياس والذي قد يكون متضمنًا في الدرجات. الدرجات عالية الثبات والموثوقية دقيقة وقابلة للتكرار ومتسقة من اختبار لآخر. أي أنه إذا تكررت عملية الاختبار مع مجموعة من المُختَبِرين، فسيتم الحصول على نفس النتائج تقريبًا.

ووفقًا للمعايير التي وضعتها الجمعية الأمريكية لبحوث التربية (AERA) والجمعية الأمريكية لعلم النفس (APA) والمجلس الوطني للقياس في التعليم(NCME) ، 2014، يُشير الثبات إلى اتساق القياسات عند تكرار عملية الاختبار لفرد أو مجموعة من الأفراد.

الثبات هو الحصول على نفس النتائج تقريباً عند تكرار القياس سواءً باستخدام الاستبيان أو الاختبار أو أياً من أدوات التقييم (Bolarinwa, 2015). باختصار، الثبات يُعبر عن استقرار أو اتساق الدرجات بمرور الوقت أو عبر المقدرين أو المقيمين (Miller, 2015). وتجدر الإشارة إلى أن نقص الثبات قد ينشأ عن الاختلافات بين المقيمين أو غموض في أدوات القياس أو عدم استقرار السمة التي يتم قياسها (Last, 2015). يرى نانلي (2016,Bardhosh et al) أن القياسات ثابتة إلى الحد الذي تكون فيه قابلة للتكرار وأن أي تأثير عشوائي يميل إلى جعل القياسات مختلفة من مناسبة إلى أخرى أو من ظرف إلى ظرف هو مصدر خطأ في القياس. ووفقًا لكلاين (2000) فإن الثبات له معنيان متميزان: يشير أحدهما إلى الاستقرار بمرور الوقت، والثاني إلى الاتساق الداخلي. فالثبات هو مؤشر على الاتساق، أي مؤشر على مدى استقرار درجة الاختبار أو البيانات عبر التطبيقات أو الوقت. يجب أن ينتج عن المقياس نتائج مماثلة أو متطابقة باستمرار إذا كان يقيس نفس الشيء.(Sawilowsky, 2000). يمكن أن يكون المقياس ثابتاً دون أن يكون صادقاً، ولكن لا يمكن أن يكون المقياس صادقاً دون أن يكون ثابتاً (2013,Erford).

يلعب معامل الارتباط دورًا هامًا في تحديد درجة الثبات، فيُعد معامل الارتباط + 1.0 علاقة إيجابية تامة، و– 1.0 علاقة سلبية تامة، بينما يشير معامل الارتباط 0.0 إلى عدم وجود علاقة ارتباطية بين المتغيرات. كلما اقترب معامل الارتباط من + 1.0، زاد ثبات الاختبار وارتفعت موثوقية النتائج. إذا كان المقياس ثابتاً تمامًا فه1 يعني أنه لا يُوجد خطأ في القياس، أي أن كل ما نلاحظه هو درجة حقيقية وصحيحة. لذلك، بالنسبة للمقياس الثابت تمامًا، تكون الموثوقية = 1. أما إذا كان المقياس غير ثابت تمامًا، فلا توجد درجة صحيحة، أي أن المقياس خاطئ تمامًا. في هذه الحالة، يكون الثبات والموثوقية = 0. تخبرنا قيمة تقدير الثبات بنسبة التباين في المقياس المنسوبة إلى الدرجة الحقيقية الصحيحة. يعني الثبات بقيمة 0.5 أن حوالي نصف تباين الدرجة المرصودة يُعزى إلى الصدق والنصف الآخر يُعزى إلى الخطأ. ووفقًا للجمعية الأمريكية لبحوث التربية(AERA) ، والجمعية الأمريكية لعلم النفس(APA) ، والمجلس الوطني للقياس في التعليم(2014) (NCME) ، فإن ثبات 0.8 تعني أن التباين يبلغ حوالي 80% من القدرة الحقيقية و20% من الخطأ. جميع إجراءات القياس تنطوي على أخطاء. ومع ذلك، فإن مقدار/درجة الخطأ هو ما يُشير إلى مدى ثبات القياس. عندما يكون مقدار الخطأ منخفضًا، يكون ثبات وموثوقية القياس عالية. وعلى العكس، عندما يكون مقدار الخطأ كبيرًا، يكون ثبات وموثوقية القياس منخفضة (Meyer, 2010 ;2013,Elford).

من المهم ملاحظة أن الثبات يُشير إلى النتيجة وليس إلى الاختبار نفسه, ويجب أن تكون العينات التي يُشتق منها معامل الثبات مُمثلةً للمجتمع الذي صُمم الاختبار من أجله، وأن تكون كبيرة بما يكفي لتكون ثابتة إحصائيًا (Leann, & Ken, 2012). ووفقًا لكلاين (2000)، فإن قيمة الثبات 0.70 هو الحد الأدنى لاختبار جيد، ويرجع ذلك ببساطة إلى أن الخطأ المعياري للقياس الذي يُمثل الانحراف المعياري المُقدّر للدرجات يزداد مع انخفاض الثبات.

بشكل عام، هناك أربعة أنواع رئيسية من الثبات: ثبات إعادة الاختبار، وثبات الصور المتكافئة، والاتساق الداخلي للثبات، وثبات المقيمين (Kaplan & Saccuzzo, 2005). في هذه الدراسة، سنتناول ثبات الاستقرار من خلال الاختبار وإعادة الاختبار، وثبات الاتساق الداخلي (ألفا كرونباخ).

ثبات اختبار وإعادة الاختبار (الاستقرار):

ثبات إعادة الاختبار ويُسمى أيضًا الاستقرار يُجيب على السؤال: "هل ستستقر الدرجات بمرور الوقت؟". يُشير ثبات إعادة الاختبار إلى الاستقرار الزمني لاختبار ما من جلسة قياس إلى أخرى، ويتمثل الإجراء في تطبيق الاختبار على مجموعة من المشاركين، ثم تطبيقه على نفس المشاركين لاحقًا. ويُحدد الارتباط بين الدرجات في الاختبارات المتطابقة التي أُجريت في أوقات مختلفة ثبات إعادة الاختبار عمليًا. ويرتكز استخدام إجراء إعادة الاختبار على افتراضين (Wells, 2003) هما:

- الافتراض الأول المطلوب هو أن الخاصية المقاسة لا تتغير على مدار الفترة الزمنية، ويُسمى ذلك "تأثير الاختبار" (Engel & Schutt, 2013)
- الافتراض الثاني هو أن الفترة الزمنية طويلة بما يكفي، لكنها قصيرة في الوقت نفسه، بحيث لا تؤثر ذاكرة المستجيبين عن إجراء الاختبار، في المرة الأولى، على درجاتهم في المرة الثانية والاختبارات اللاحقة، ويُسمى ذلك "تأثير الذاكرة."

يُعرف تقدير ثبات إعادة الاختبار أيضًا بمعامل الاستقرار (Cohen et al, 1996). يُوفر ارتباط إعادة الاختبار مؤشرًا على الاستقرار بمرور الوقت (Deniz, & 2013; Pedisic et al, 2014, Wong Ong & Kuek, 2012)

Alsaffar). بمعنى آخر، تكون الدرجات متسقة من المرة الأولى إلى الثانية. عند استخدام هذا النوع من الثبات، يجب توخي الحذر عند استخدام الاستبيانات أو المقاييس التي تقيس متغيرات يُحتمل تغيرها خلال فترة زمنية قصيرة، مثل الطاقة والسعادة والقلق بسبب تأثير النضج (Drost, 2011). بالنسبة لاختبارات التحصيل المعيارية المتطورة جيدًا، والتي تُجرى على فترات زمنية متقاربة نسبيًا، تتراوح تقديرات ثبات إعادة الاختبار بين 0.70 و 0.90 (Popham, 2000).

أسلوب ثبات الاختبار وإعادة الاختبار فيها العديد من القيود رغم جاذبيتها (Rosenthal & Rosnow, 1991). فعلى سبيل المثال، عندما تكون الفترة الفاصلة بين الاختبار الأول والثاني قصيرة جدًا، فقد يتذكر المستجيبون ما كان في الاختبار الأول وقد تتأثر إجاباتهم في الاختبار الثاني بالذاكرة. وبدلاً من ذلك، عندما تكون الفترة الفاصلة بين الاختبارين طويلة جدًا، فقد يحدث النضج. لاحظ Kaplan and Saccuzzo (2005) أن تقديرات ثبات الاختبار وإعادة الاختبار تقيم ثبات درجات الأداة عندما تُعطى الأداة في نقاط زمنية متعددة ومتتالية. وفي هذا الصدد كشفت دراسة Joppe (2000) عن مشكلة في طريقة الاختبار وإعادة الاختبار والتي يمكن أن تجعل الأداة، إلى حد ما، غير ثابتة، بسبب أن هذه الطريقة قد تجعل المستجيب حساسًا للموضوع، وبالتالي يؤثر ذلك على الاستجابات المقدمة. وبالمثل، لاحظ Crocker and Algina (1986) أنه عندما يجيب المستجيب على مجموعة من المفردات في الاختبار، فإن الدرجة التي يتم الحصول عليها تمثل عينة محدودة فقط من السلوك.

(Wong, Ong, & Kuek, 2012)

الاتساق الداخلي

يُجيب ثبات الاتساق الداخلي على السؤال التالي: "ما مدى جودة قياس كل بند للمحتوى أو البنية قيد الدراسة؟". تكمن جاذبية مؤشر الاتساق الداخلي للثبات في أنه يُقدّر بعد إجراء اختبار واحد فقط، وبالتالي، يتجنّب المشاكل المرتبطة بالاختبار على مدى فترات زمنية متعددة. (Wong, Ong, & Kuek, 2012). يشير تقدير ثبات الاتساق الداخلي إلى الارتباطات المتبادلة بين البنود على الأداة نفسها (Kaplan & Saccuzzo,2005). يُعدّ معامل ألفا لكرونباخ من أكثر الطرق استخدامًا لتقدير الاتساق الداخلي للثبات (Dimitrov,2002). ويُعدّ معامل ألفا الإجراء الأكثر استخدامًا لتقدير الثبات في البحوث التطبيقية. وكما ذكر Sijtsma (2009) ، فإن شعبيته كبيرة لدرجة أن كرونباخ (1951) تم الاستشهاد به كمرجع أكثر من المقالة التي تناولت اكتشاف الحلزون المزدوج للحمض النووي DNA الذي يُعد من أعظم الاكتشافات العلمية الحديثة. ومع ذلك، فإن حدودها معروفة جيدًا (Yang & Green,2011) ، ومن أهمها افتراضات الأخطاء غير المرتبطة، وتكافؤ تاو، والتوزيع الطبيعي.

تحديد حجم العينة في الثبات

شكّل تحديد حجم العينة مشكلةً رئيسيةً للباحثين وأخصائي القياس النفسي في دراسات الثبات. وتنوعت المناهج الحالية لتحديد حجم العينة في الدراسات النفسية، ولم تكن واضحةً ومباشرةً. وقد أدى ذلك إلى احتواء الأدبيات النفسية على مجموعة واسعة من المقالات التي تقترح أحجامًا مختلفة للعينات (Donner & Eliasziw; 1987, Eliasziw et al; 1994, Cocchetti; 1999, Charter; 1999, Mendoza, Stafford, & Stauffer, (2000, Bonett; 2002). وتُصنف هذه الدراسات إلى فئتين رئيسيتين: دراسات تستند إلى تجارب المؤلفين، ودراسات تستند إلى النظرية الإحصائية.

في الدراسات التي تستند إلى أحكام مستمدة من تجارب المؤلفين (1991,DeVellis;1992,Rea,& Parker;1993 Ferguson & Cox,)، تتباين توصيات حجم العينة تباينًا كبيرًا. دعا مؤلفون آخرون واقترحوا أن تتجاوز العينات 300

(Ware, et al ، 1997)، بينما افترض البعض أن عينات أصغر بكثير لا تتجاوز 30 فردًا (Rea,& Parker; Bonett & Wright, 19922014) قد تكون كافية. تشمل الفئة الثانية من توصيات حجم العينة الدراسات القائمة على النظرية الإحصائية (1987,Feldt et al, 1987; Donner & Eliasziw; 1994,Eliasziw, et al; 2002 Bonett,). تختلف هذه الدراسات في مناهج اختبار الثبات (Mendoza et al, 2000 ;1999,Charter) وتوصيات تتراوح فيها قيم n من n = 25 (1999,Cocchetti) إلى 400 لاختبار الثبات (Charter, 1999)

أشار Kline (2000) إلى أنه يجب على الباحثين استخدام 100 مشارك على الأقل لكل عنصر على مقياسنا إذا كان تقدير الثبات ذا معنى. توجد الكثير من الاختلافات في الرأي حول تحديد حجم العينة في الأدبيات. يقترح بعض المؤلفين أن العينات الصغيرة التي تصل إلى ثلاثين (30) (Bonett, & Wright, 2014) يمكن أن تقيس الثبات، طالما أن عناصر المقياس بينها ارتباط قوي. وفي نفس السياق، أكد Nunnally & Bernstein (1994) أن الحد الأدنى لمعايير معاملات الثبات لألفا كرونباخ هو 0.80 إلى 0.30 لارتباطات العبارة بالمجموع، و0.30 لارتباطات العبارة بالعبارة، و0.80 لمعاملات الارتباط داخل الفئة. اقترح Kline (1986) حجم عينة أدنى يبلغ 300، ويتفق معه في ذلك كلاً من Nunnally & Bernstein (1994). وصف Segall (1994) حجم العينة البالغ 300 بأنه "صغير". وذكر Charter (1999) أن الحد الأدنى لحجم العينة البالغ 400 مطلوب لتقدير دقيق بما فيه الكفاية لمعامل ألفا للمجتمع. ورأى Charter (2003) أن معاملات ألفا قد تكون غير مستقرة مع أحجام العينات الصغيرة. واقترح Walker and Zhang (2004) حجم عينة أدنى يتراوح بين 125 و150 لحساب الثبات، بحيث يكون عدد الأفراد في العينة مساويًا على الأقل لعدد عناصر الاختبار. ومع ذلك، فقد ثار جدل حول الحد الأدنى لحجم العينة لمعامل ألفا للعينة نظرًا لصعوبة جمع البيانات في البحوث النفسية. ورغم أن تحديد حجم العينة اللازم لدراسات الثبات أمرٌ ذاتي إلى حد ما، إلا أنه يُوصى بحد أدنى يبلغ 400 فرد.

في دراسات الثبات، يستخدم مؤلفون وباحثون مختلفون أحجام عينات مختلفة. علاوة على ذلك، لا يوُجد تجانس في أحجام العينات المستخدمة. يلعب حجم العينة دورًا مهمًا في تقدير مستوى ثبات المقياس المستخدم للقياس.

إلى جانب معظم المؤشرات الإحصائية الأخرى، تحتوي الارتباطات على أخطاء معيارية، مما يُشير إلى مدى ثبات وموثوقية النتائج. ومع ذلك، يمكن القول أنه كلما زاد عدد المشاركين، قلّ الخطأ المعياري للإحصاءات (2013,Erford). وهذا يعني أنه من الضروري أن تُستمد تقديرات الثبات من عينة كبيرة بما يكفي لتقليل هذا الخطأ الإحصائي (AERA، APA، وNCME، 2014). في اختبار الثبات، غالبًا ما يكون تحديد حجم العينة المناسب أمرًا بالغ الأهمية (2013 Meyer, 2010 ;Erford,). إذا كان حجم العينة المستخدم صغيرًا جدًا، فإنه لا يمكن الحصول على الكثير من المعلومات من الاختبار، مما يحد من قدرة الفرد على استخلاص استنتاجات ذات معنى. من ناحية أخرى، إذا كان حجم العينة كبيرًا جدًا، فقد تتجاوز المعلومات التي تم الحصول عليها من خلال الاختبار ما هو مطلوب (AERA، APA، وNCME، 2014). وبالتالي، تكبد الباحثين تكاليف غير ضرورية. لكن في أغلب الأحيان، لا يتوفر لدى مطوري الاختبار موارد كثيرة أو وفرة للحصول على عدد العينات المطلوبة، بل يتعين عليهم إنشاء خطة اختبار تعتمد على الميزانية أو قيود الموارد الموضوعة للمشروع البحثي.

مشكلة البحث:

هناك اختلافٌ ملحوظٌ في الآراء في الأدبيات حول حجم العينة المناسب لتحديد ثبات أدوات البحث، فعلى سبيل المثال، أشار Kline (2000) إلى أن النصيحة الأساسية هي استخدام 100 مشارك على الأقل لكل بند على المقياس لضمان

دقة تقدير الثبات. ومن ناحية أخرى أكد كلاً من Bonnet and Wright (2014) أن العينات يجب ألا تتجاوز 30 عينة لتحديد الثبات، طالما أن بنود المقياس تتمتع بترابط قوي. علاوةً على ذلك، يستخدم العديد من الباحثين أحجام عينات مختلفة لتحديد تقديرات الثبات عند إجراء الدراسات البحثية، فيستخدم البعض 20 أو 30 أو 40 أو 50 أو 100 عينة حسب الحالة. ولكن –حسب علم الباحث –هناك نُدرة في الأبحاث العلمية التي أُجريت لتبرير استخدام هذه الأحجام، كما يستخدم بعض الباحثين أساليب مختلفة لتقدير معاملات الثبات في أنواعه المختلفة. فعلى سبيل المثال، يستخدم البعض اختبار إعادة الاختبار لأداة الاستبيان بدلاً من معامل ألفا كرونباخ الشائع. (Vacha-Haase & Thompson, 2010).

وعلى الرغم من أن موضوع الثبات قد حظي باهتمام كبير في الأدبيات، إلا أن الدراسات المتعلقة بمتطلبات حجم العينة لا تزال نادرة، لذلك، فإنه من الضروري فحص أثر أحجام عينات مختلفة على تقديرات معامل الثبات لطرق الثبات الأكثر استخداماً: طريقة إعادة الاختبار وطريقة معامل ألفا كرونباخ.

أسئلة البحث

تسعى الدراسة الحالية للإجابة عن الأسئلة الآتية:

1 .هل هناك فرق في تقدير ثبات إعادة الاختبار لأداة باستخدام أحجام عينات مختلفة (20، 30، 40، 50، 100، 150، 200، 300، 400، 500)؟

2 .هل هناك فرق في تقدير ثبات ألفا كرونباخ لأداة باستخدام أحجام عينات مختلفة (20، 30، 40، 50، 100، 150، 200، 300، 400، 500)؟

أهمية الدراسة

ستساعد نتائج الدراسة المهتمين بالقياس النفسي والمعلمين والباحثين على معرفة الحد الأدنى لحجم العينة عند إجراء دراسات الثبات، وهذا سيُساعد في التغلب على مشكلة اختيار حجم العينة المناسب لضمان قيم ثبات مقبولة. ستكون هذه الدراسة إلى جانب مثيلاتها بمثابة نافذة على المنهجية وحجم العينة اللازمين لإجراء دراسة ثبات. وفي السياق نفسه، ستساعد النتائج الباحثين والأخصائيين في القياس النفسي على تقدير نسبة التباين في قياساتهم، والتي تُعزى إلى النتيجة الحقيقية. أي أنها ستساعدهم على تحديد مقدار أو درجة الخطأ التي تُشير إلى مدى ثبات المقياس. فعندما يكون مقدار الخطأ منخفضًا، يكون ثبات المقياس عالي، وعلى العكس، عندما يكون مقدار الخطأ كبيرًا، يكون ثبات المقياس منخفض.

ستكون هذه الدراسة مفيدة أيضًا للباحثين وغيرهم من أصحاب المصلحة الذين قد يُواجهون صعوبات في اختيار الطرق المناسبة لتقدير تقديرات الثبات. وستُساعد هذه الدراسة جميع الباحثين وغيرهم من أصحاب المصلحة على الإبلاغ بدقة عن تقديرات الثبات في أي إصدارات ومستندات وأدلة علمية على سبيل المثال كتيبات الاختبارات، وأوراق المؤتمرات، والمقالات.

المنهجية

اعتمدت الدراسة على أسلوب البحث المسحي، وأُختيرت عينة تطوعية من المشاركين والمشاركات المستجيبين بلغت 500 مشاركاً ومشاركة. كانت أداة جمع البيانات هي مقياس الانتماء الوطني، وهو مقياس للانتماء الوطني طوره (القرني، 2019). وقد قام الباحث باختبار خصائصه السيكومترية. يتكون من قسمين. حيث تم استخدام القسم الأول لاستخلاص

المعلومات والبيانات الأساسية والديمغرافية، والتي تشمل جنسهم وفئاتهم العمرية. يتكون القسم من قائمة عبارات مكونة من ثلاثة وعشرين (23) عبارة مصممة لقياس مستوى الانتماء الوطني في ثلاثة أبعاد رئيسة وذلك باستخدام مقياس ليكرت ذو الأربع نقاط. تقع العبارات تحت خيارات الاستجابة = SA :موافق بشدة، = Aموافق، = Dغير موافق، = SDغير موافق بشدة. حيث تم تسجيل SD نقطة واحدة، = Dنقطتين، 3 = Aنقاط و 4 = SAنقاط. تم التحقق من صحة الأداة من قبل (القرني، 2019)، كما تم التحقق من صحتها من قبل خبراء في القياس والتقييم، حيث كان ثبات وصدق الأداة جزءًا من المسائل التي أثيرت في الدراسة.

قُدِّر معامل الثبات باستخدام معامل ارتباط بيرسون (r) للأداة التي أُعيد اختبارها، ومعامل كرونباخ ألفا α للأداة التي طُبِّقت مرة واحدة. استُخدمت فترة ثقة فيشر 95% لتحديد أيٍّ من أحجام العينات يُعطي نتيجة مستقرة، وحُدِّد عرض الفترات في ضوء أحجام العينات المختلفة.

النتائج:

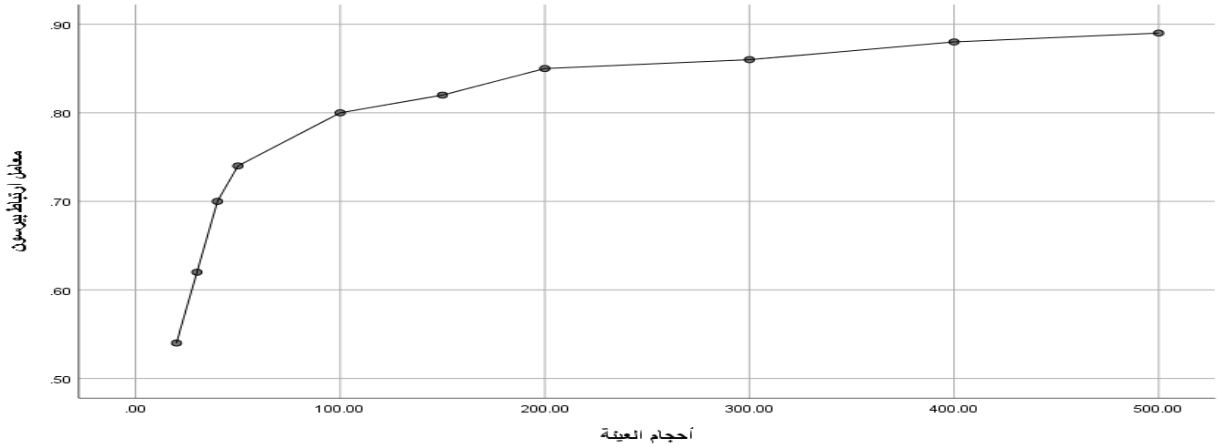
جدول (1) فترة الثقة 95% لتقديرات الثبات بطريقة إعادة الاختبار باستخدام تحويل فيشر:

العرض	الحد الأعلى لـ r	الحد الأدنى لـ r	الحد الأعلى لـ Z	الحد الأدنى لـ Z	الخطأ المعياري (σz)	قيمة فيشر (Zr)	معامل بيرسون (r)	حجم العينة
0.66	0.783	0.124	1.079	0.125	0.243	0.602	0.54	20
0.46	0.8	0.337	1.101	0.351	0.192	0.726	0.62	30
0.33	0.831	0.498	1.188	0.546	0.164	0.867	0.7	40
0.26	0.844	0.581	1.235	0.663	0.146	0.949	0.74	50
0.15	0.861	0.716	1.299	0.899	0.102	1.099	0.8	100
0.12	0.869	0.752	1.324	1.002	0.082	1.163	0.82	150
0.09	0.889	0.798	1.394	1.116	0.071	1.255	0.85	200
0.07	0.896	0.822	1.418	1.19	0.058	1.304	0.86	300
0.05	0.902	0.856	1.479	1.283	0.05	1.381	0.88	400
0.03	0.906	0.872	1.506	1.336	0.044	1.421	0.89	500

النتائج في الجدول أعلاه جدول رقم (1) توضح فترة الثقة 95٪ لتحليل ثبات أداة باستخدام طريقة إعادة الاختبار، وذلك عبر أحجام عينات مختلفة تبدأ من 20 وتصل إلى 500. يتضح من النتائج أن ازدياد حجم العينة يؤدي إلى تضييق فترة الثقة، مما يُشير إلى استقرار أكبر في تقديرات الثبات. وقد اعتُبر حجم العينة الذي يعطي أقل عرض للفاصل هو الأكثر اتساقاً واعتماداً في التقدير. هذا يدعم أهمية استخدام عينات كبيرة نسبيًا عند دراسة خصائص الثبات في الأدوات النفسية والتربوية.

تشير النتائج في جدول (1) إلى وجود علاقة طردية واضحة بين حجم العينة وقيمة معامل الثبات باستخدام معامل بيرسون. أظهرت النتائج أنه مع تزايد حجم العينة من 20 إلى 500، ارتفعت قيمة معامل بيرسون تدريجيًا من 0.54 إلى

0.89. هذا النمط يدل على أن تقديرات الثبات تصبح أكثر دقة واتساقًا مع ازدياد عدد المفحوصين، مما يعزز من موثوقية الأداة المستخدمة في القياس. كما أن عرض فترة الثقة يتناقص تدريجيًا، مما يشير إلى تزايد الثقة في دقة التقدير. وعليه، فإن اختيار حجم عينة مناسب وكبير يعد عنصرًا أساسيًا في دراسات الثبات والموثوقية

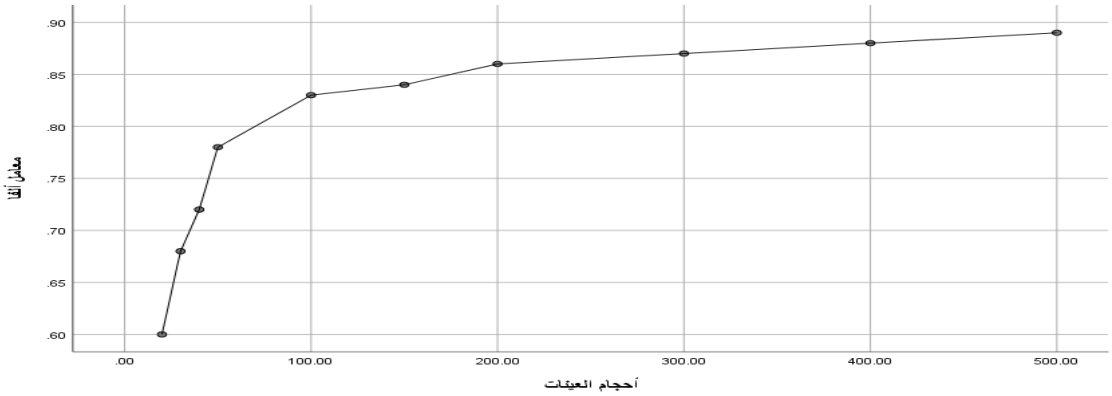


الشكل ١: يوضح فترة ثقة فيشر ٩٥٪ لتقديرات ثبات الاختبار وإعادة الاختبار.

جدول رقم (2) فترة الثقة 95٪ لتقديرات الثبات باستخدام ألفا كرونباخ

العرض	الحد الأعلى لـ α	الحد الأدنى لـ α	الحد الأعلى لـ Z	الحد الأدنى لـ Z	الخطأ المعياري (σz)	قيمة فيشر (Zα)	ألفا كرونباخ (α)	حجم العينة
0.6	0.828	0.23	1.175	0.475	0.243	0.7	0.6	20
0.4	0.841	0.44	1.224	0.472	0.192	0.848	0.68	30
0.31	0.873	0.566	1.366	0.722	0.164	1.045	0.72	40
0.21	0.885	0.675	1.385	0.813	0.146	1.099	0.78	50
0.12	0.88	0.757	1.387	0.99	0.102	1.188	0.83	100
0.1	0.882	0.786	1.382	1.061	0.082	1.221	0.84	150
0.08	0.89	0.812	1.419	1.149	0.071	1.284	0.86	200
0.06	0.891	0.829	1.421	1.194	0.058	1.307	0.87	300
0.05	0.892	0.844	1.433	1.235	0.05	1.334	0.88	400
0.04	0.893	0.856	1.445	1.275	0.043	1.36	0.89	500

يتضح من الجدول أعلاه أن معامل الثبات ألفا كرونباخ يزداد تدريجيًا مع تزايد حجم العينة، مما يعكس تحسن دقة التقدير وموثوقية الأداة. كما أن عرض فترة الثقة يتناقص بشكل ملحوظ، مما يشير إلى ازدياد الاتساق والثبات في التقدير الإحصائي. وتُعد هذه النتائج مؤشرًا مهمًا على أن استخدام عينات أكبر في البحوث التربوية والنفسية من شأنه أن يُسهم في الحصول على تقديرات أكثر دقة وثباتًا لمعامل الثبات، مما يعزز من صلاحية الأداة المستخدمة.



الشكل 2: يوضح فترة ثقة فيشر ٩٥٪ لتقديرات ثبات الاختبار وإعادة الاختبار

مناقشة النتائج

كشفت الدراسة أن أحجام العينات (20 و30) باستخدام إحصاءات إعادة الاختبار لم تكن ثابتة. ورغم ثبات حجم العينة (40 و50)، إلا أن الحد الأدنى كان خارج نطاق الثبات المقبولة وهو 0.70 لاختبار إعادة الاختبار (Kline , 2000). وازداد ثبات الأداة قوةً عندما كان حجم العينة 100 على الأقل. وتتوافق هذه النتيجة مع دراسة Leann و Ken (2012) التي أكدت على أن العينات التي يُشتق منها معامل الثبات يجب أن تكون كبيرة بما يكفي لتكون ثابتة إحصائيًا. كما تتفق هذه النتيجة مع دراسة Kline (2000) التي أشارت إلى أن النصيحة القياسية والأساسية هي استخدام 100 مشارك على الأقل على المقياس لضمان دقة تقدير الثبات. وفي المقابل لا تتفق النتيجة الحالية مع رأي Bonnet & Wright (2014) اللذين أكدا أن العينات يجب أن تكون صغيرة مثل ثلاثين (30) لتحديد الثبات طالما أن عناصر المقياس لها ارتباط قوي، ورأي ريا وباركر (1992) اللذين افترضا أن العينات الأصغر التي لا تزيد عن 30 موضوعًا قد تكون كافية لتقدير الثبات بطريقة الاختبار وإعادة الاختبار.

وكشفت الدراسة أيضًا أن أحجام العينات 20 و 30 باستخدام إحصائيات ألفا كرونباخ لم تكن ثابتة. كما أن حجم العينة 40 و 50، على الرغم من ثباتة فقد كان الحد الأدنى خارج معاملات الثبات المقبولة 0.80 فأكثر لألفا كرونباخ Nunnally & Bernstein (1994). أصبح ثبات الأداة أقوى عندما كان حجم العينة 100 على الأقل، وهذه النتيجة تتوافق مع AERA و APAوNCME، (2014) و Erford، (2013) الذين أشاروا إلى أنه كلما زاد عدد الأشخاص، قل الخطأ المعياري للإحصاء مما يعني أنه من الضروري أن يتم اشتقاق تقديرات ثبات من عينة كبيرة بما يكفي لتقليل هذا الخطأ الإحصائي. كما تتوافق هذه النتيجة نسبياً مع دراسة Kline (1986) و Nunnally & Bernstein (1994) التي اقترحت حجم عينة أدنى يبلغ 300. أطلق سيجال (1994) على حجم عينة يبلغ 300 اسم "صغير". كما أشار Charter (1999) أن حجم عينة أدنى يبلغ 400 مطلوب لتقدير دقيق بدرجة كافية لـ معامل ألفا لمجتمع الدراسة. أشار Charter (2003) أيضًا إلى أن معاملات ألفا قد تكون غير مستقرة مع انخفاض أحجام العينات. اقترح Walker and Zhang (2004) حجم عينة أدنى يتراوح بين 125 و150 لحساب الثبات، بحيث لا يقل عدد الأفراد في العينة عن عدد عناصر الاختبار. كما اقترح Charter (1999) حجم عينة يبلغ 400 لاختبار الثبات. في حين اختلفت هذه النتيجة مع نتائج دراسات Feldt et al (1987)، Donner & Eliasziw (1987)، Eliasziw et al (1994)، Bonett (2002)، و Charter (1999)، و Mendoza et al (2000)، و Cocchetti (1999) الذين أوصوا بأحجام عينات تبدأ من 25 بحيث n =25.

قد يكون الاختلاف في نتائج هذه الدراسة ناتجًا عن استخدام القيم المرصودة ميدانيًا، فمعظم النتائج الواردة في الأدبيات العلمية كانت إما من التجربة الشخصية أو من النظرية الإحصائية. للأسف فإن الكثير من الأدلة التجريبية يأتي من بيانات مُحاكاة، لذا، تُعدّ توصياتهم غير مكتملة، لأن بيانات المُحاكاة تنطوي على قيود مهمة مُقارنةً بالبيانات المُلاحظة الواقعية. فهي تستند إلى نماذج إحصائية أو حاسوبية مُختارة مُسبقًا، لا يُمكنها سوى المقاربة للبيانات المُلاحظة، ولها مُعاملات قابلة للتحكم بشكل مُصطنع، وغالبًا ما تُولّد لتعكس عينات مُوزّعة عشوائيًا. وهذا يُحدّ من الاستدلالات التي يُمكن استخلاصها من تحليل البيانات المُحاكاة، ويستلزم جمع البيانات المُلاحظة لضمان مصداقيتها.

ومن النتائج الأخرى التي توصلت إليها الدراسة الحالية أن تقديرات ثبات كلٍّ من اختبار إعادة الاختبار وتقديرات ثبات كرونباخ بدأت تتقارب من حجم عينة يبلغ 100 (انظر الشكلين 1 و2). وهذا يعني، بالتالي، أنه لدراسة ثبات مقبول، يجب استخدام مائة فرد على الأقل.

كما كشفت نتائج الدراسة أن تقدير الفاصل الزمني أعطى تقديرًا للثبات أفضل من تقدير النقطة لجميع العينات، فعلى سبيل المثال، بالنسبة لاختبار إعادة الاختبار، أعطت عينة من 40 فردًا مؤشر ثبات قدره 0.70 كتقدير نقطي، لكن تقدير الفاصل الزمني أعطى تقديرًا للثبات قدره 0.498، كان الحد الأدنى خارج نطاق مؤشر الثبات المقبول (≤= 0.70). يتوافق هذا مع نتائج دراسات (AERA وAPA وNCMEو2014)، والتي دعت إلى استخدام تقديرات الثبات كتقديرات فترة مقارنةً بتقدير النقطة المستخدم سابقًا.

الخلاصة

بناءً على نتائج هذه الدراسة، خلصت إلى الاستنتاجات التالية. أظهرت النتائج وجود عدد من الاختلافات في تحديد حجم العينة في دراسات الثبات. ولم يكن استخدام أحجام عينات تتراوح بين عشرين وثلاثين (30) عينة مبررًا. ويُعزى ذلك إلى أن دراسات أخرى اقترحت حدًا أدنى من عشرين وثلاثين عينة استخدمت بيانات محاكاة، مقارنةً بالبيانات المرصودة المستخدمة في هذه الدراسة.

كلما زاد عدد العينات، قلّ الخطأ المعياري للإحصاء. ولتقليل هذا الخطأ الإحصائي، يجب استخلاص تقديرات الثبات من عينة كبيرة بما يكفي. وقد أظهرت نتائج الدراسة أن استخدام أحجام عينات تتراوح بين عشرين وثلاثين (20) عينة لدراسات الثبات غير مبرر. كما أظهرت أنه للحصول على ثبات مقبول، يجب ألا يقل حجم العينة عن مئة(100) .

التوصيات

يُعد ثبات أي أداة القياس أمراً مهماً في الأبحاث، ويلعب تحديد حجم العينة دورًا بالغ الأهمية في تقدير الثبات. فكلما زاد حجم العينة، زاد الثبات وانخفض الخطأ الكامن في الأداة. وبناءً على ذلك، وُضعت التوصيات التالية:

1. ينبغي دائمًا استخدام القيم المُلاحظة أو المُختبرة ميدانيًا في تقدير ثبات أي أداة قياس.
2. للحصول على تقدير ثبات عالٍ، ينبغي استخدام مائة (100) فرد على الأقل.
3. لا ينبغي استخدام الثبات كتقدير نقطة، بل الأفضل استخدامه كتقدير فترة.

REFERENCES

1. Alqarni, A. (2020). Developing a measure of national affiliation on a sample of Saudi society in the light of some demographic variables. *Journal of Educational and Psychological Sciences*, Umm Al-Qura University.
2. American Educational Research Association (AERA). (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
4. Balkin, R. S. (2017). Evaluating evidence regarding relationships with criteria. *Measurement and Evaluation in Counseling and Development, 50*, 264–269.
5. Bardhoshi, G., Erford, B. T., Duncan, K., Dummett, B., Falco, M., Deferio, K., & Kraft, J. (2016). Choosing assessment instruments for posttraumatic stress disorder screening and outcome research. *Journal of Counseling & Development, 94*, 184–194.
6. Bolarinwa, O. A. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science research. *Niger Postgraduate Medical Journal, 22*, 195–201.
7. Bonett, D. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational Behavioural Statistics, 27*, 335–340.
8. Bonett, D. G., & Wright, T. A. (2014). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size. *Journal of Organizational Behaviour, 36*(1).
9. Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*, 559–566.
10. Charter, R. A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology, 130*, 117–129.
11. Cocchetti, D. (1999). Sample size requirements for increasing the precision of reliability estimates: Problems and proposed solutions. *Journal of Clinical Experimental Neuropsychology, 21*, 567–570.
12. Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honour: An "experimental ethnography." *Journal of Personality and Social Psychology, 70*, 945–960.
13. Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt Brace Jovanovich College Publishers: Philadelphia.
14. Deniz, M. S., & Alsaffar, A. A. (2013). Assessing the validity and reliability of a questionnaire on dietary fibre-related knowledge in a Turkish student population. *Journal of Health Population and Nutrition, 31*, 497–503.
15. Devellis, R. F. (1991). *Scale Development: Theory and Applications*, Applied Social Research Methods Series, 26. Sage: Newbury Park.
16. Dimitrov, D. M. (2002). Error variance of Rasch measurement with logistic ability distributions. Paper presented at the meeting of the American Educational Research Association, New Orleans, Louisiana.

17. Donner, A., & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Journal of Statistical Medicine, 6*, 441–448.

18. Drost, E. A. (2011). Validity and reliability in social science research. *Educational Research Perspective, 38*, 105–123.

19. Eliasziw, M., Young, S., Woodbury, M., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and interrater reliability: Using goniometric measurements as an example. *Journal of Physical Therapy, 74*, 777–788.

20. Engel, R. J., & Schutt, R. K. (2013). *Measurement. The Practice of Research in Social Work* (3rd ed.). Sage Publications, Inc.

21. Erford, B. T. (2013). *Assessment for Counsellors* (2nd ed.). Belmont, CA: Cengage Wadsworth.

22. Erford, B. T., Johnson, E., & Bardhoshi, G. (2016). Meta-analysis of the English version of the Beck Depression Inventory–Second edition. *Measurement and Evaluation in Counseling and Development, 49*, 3–33.

23. Feldt, L., Woodruff, D., & Sailh, F. (1987). Statistical inference for coefficient alpha. *Journal of Applied Psychological Measure, 11*, 93–103.

24. Ferguson, E., & Cox, T. (1993). *Exploratory Factor Analysis: A User's Guide*. International Journal of Selection and Assessment, 1, 84–94.

25. Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

26. Joppe, M. (2000). The Research Process. Retrieved February 25, 1998, from http://www.ryerson.ca/~mjoppe/rp.htm

27. Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological Testing: Principles, Applications, and Issues* (6th ed.). Belmont, CA: Thomson Wadsworth.

28. Kline, P. (1986). *A Handbook of Test Construction: Introduction to Psychometric Design*. New York: Methune & Company.

29. Kline, R. B. (2000). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioural Research*. Washington, DC: American Psychological Association.

30. Last, J. M. (2015). *A Dictionary of Epidemiology* (4th ed.). New York: Oxford University Press.

31. Leann, J. T., & Ken, J. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter arrow estimation via narrow confidence intervals. *The British Journal of Mathematics and Statistical Psychology, 65*, 371–401.

32. McLeod, S. A. (2007). What is Reliability? Retrieved on June 27, 2017, from www.simplypsychology.org/reliability.html.

33. Mellenbergh, G. J. (2011). *A Conceptual Introduction to Psychometrics*. The Hague, Netherlands: Eleven International.

34. Mendoza, J., Stafford, K., & Stauffer, J. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Journal of Psychological Methods, 5*, 356–369.

35. Meyer, P. (2010). *Reliability: Understanding Statistics Measurement*. New York, NY:

Oxford University Press.

36. Miller, M. J. (2015). *Graduate Research Methods*. Available from: <="" a=""> [Last accessed on 2015 Oct 10].
37. National Council on Measurement and Evaluation in Education (1999).
38. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill: New York.
39. Pedisic, Z., Bennie, J. A., Timperio, A. F., Crawford, D. A., Dunstan, D. W., & Bauman, A. E. (2014). Workplace sitting breaks questionnaire (SITBRQ): An assessment of concurrent validity and test-retest reliability. *BMC Public Health, 14*, 1249.
40. Popham, W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (3rd ed.). Needham, MA: Allyn and Bacon.
41. Rea, L., & Parker, R. (1992). *Designing and Conducting Survey Research: A Comprehensive Guide*. Jossey-Bass, San Francisco.
42. Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis* (2nd ed.). McGraw-Hill Publishing Company.
43. Sawilowsky, S. S. (2000). Psychometrics versus data metrics: Comments on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Journal of Educational and Psychological Measurement, 60*, 157–173.
44. Segall, D. O. (1994). The reliability of linearly equated tests. *Psychometrika, 59*, 361–375.
45. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.
46. Vacha-Haase, T., & Thompson, B. (2010). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*, 159–168.
47. Wells, C. S. (2003). Reliability and Validity; Available from: <="" a="">. [Last accessed on 2015 Dec 09].
48. Wilkinson, G. S., & Robertson, G. J. (2006). *Manual for the Wide-Range Achievement Test (WRAT–4)*. Los Angeles, CA: Western Psychological Services.
49. Wong, K. L., Ong, S. F., & Kuek, T. Y. (2012). Constructing a survey questionnaire to collect data on service quality of business academics. *European Journal of Social Science, 29*, 209–221.