مجلـــة العلـــوم التربـــوية والدراســـات الإنســـانيـــة
Humanities and Educational Sciences Journal
ISSN: 2709-0302 (online)
ISSN: 2617-5908 (print)

المقارنة بين الطريقة الجدية والارتباطية في حساب تمييز الفقرة في ضوء تباين حجم العينة وطول الاختبار(*)

بشائر بنت عبد الله بن محمد الغدير
أستاذ مساعد في علم النفس التربوي القياس والتقويم
بجامعة المستقبل المملكة العربية السعودية
bshayralghdyr@gmail.com

*) تاريخ تسليم البحث 25/10/2025 تاريخ قبوله للنشر 28/12/2025
*) موقع المجلة: http://hesj.org/ojs/index.php/hesj/index

المقارنة بين الطريقة الجدية والارتباطية في حساب تمييز الفقرة في ضوء تباين حجم العينة وطول الاختبار

بشائر بنت عبد الله بن محمد الغدير
أستاذ مساعد في علم النفس التربوي القياس والتقويم
بجامعة المستقبل المملكة العربية السعودية

الملخص

تتناول هذه الدراسة مقارنةً نظرية بين طريقتين شائعتين في حساب تمييز الفقرة ضمن النظرية الكلاسيكية في القياس: طريقة المجموعتين المتطرفتين والطريقة الارتباطية (ارتباط الفقرة بالمجموع). تتمثل مشكلة الدراسة في احتمال تضارب تقديرات التمييز باختلاف الطريقة المعتمدة، وفي حساسية هذه التقديرات لتغير حجم العينة وطول الاختبار، بما قد ينعكس على قرارات الإبقاء على الفقرات أو حذفها وعلى جودة الاختبار، وتهدف الدراسة إلى بناء إطار مفاهيمي يوضح الفروق الجوهرية بين الطريقتين، وتحليل أثر حجم العينة وطول الاختبار على استقرار التقدير، ودقته، والتحيزات المحتملة، وصولًا إلى إرشادات نظرية لاختيار المؤشر الأنسب وفق سياق القياس، واعتمدت الدراسة المنهج النظري التحليلي القائم على تحليل الأدبيات السيكومترية، ومناقشة الافتراضات التي تحكم مؤشرات التمييز في ضوء خصائص العينة والاختبار، وتناولت الدراسة محاور: مفهوم تمييز الفقرة، وعلاقته بالصعوبة والثبات والصدق، وصف الطريقتين وخصائصهما، ثم مقارنة أثر (ن) والطول في كل منهما، وتتمثل الاستنتاجات في أن الطريقة التقليدية أكثر مباشرة تفسيرًا، لكنها أقل استقرارًا عند العينات الصغيرة بسبب تقليص العينة وإهمال الوسط، بينما تميل الطريقة الارتباطية – خصوصًا المصححة – إلى اتساق أكبر مع الاتساق الداخلي، مع تأثر ملحوظ بطول الاختبار، وتجانس العينة، وبنية الأبعاد.

الكلمات المفتاحية: تمييز الفقرة، النظرية التقليدية للقياس، معامل تمييز، طريقة المجموعتين المتطرفتين، طريقة الارتباطية (ارتباط فقرة – مجموع)، حجم العينة، طول الاختبار، الثبات، الصدق.

Comparison between the syllogistic and correlational methods in calculating item discrimination in light of variations in sample size and test length

Bashaer Abdullah Mohammed ALGhadeer
Assistant Professor of Educational Psychology, Measurement, and Evaluation
Mustaqbal University in the Kingdom of Saudi Arabia

Abstract

This study presents a theoretical comparison between two common methods for calculating item discrimination within classical measurement theory: the extreme groups method and the correlational method (item-to-group correlation). The research problem lies in the potential for discrepancies in discrimination estimates depending on the method used, and in the sensitivity of these estimates to variations in sample size and test length. This sensitivity can impact decisions regarding item retention or exclusion, as well as test quality. The study aims to develop a conceptual framework that clarifies the fundamental differences between the two methods, analyzes the impact of sample size and test length on the stability, accuracy, and potential biases of the estimation, and ultimately provides theoretical guidelines for selecting the most appropriate indicator within the measurement context. The study employs a theoretical and analytical approach based on an analysis of psychometric literature and a discussion of the assumptions governing discrimination indicators in light of sample and test characteristics. The study addresses the following key areas: the concept of item discrimination and its relationship to difficulty, reliability, and validity; a description of the two methods and their characteristics; and a comparison of the impact of n and test length on each method. The expected conclusions are that the traditional method is more direct in its explanation but less stable at small samples due to sample reduction and neglect of the mean, while the correlational method - especially the corrected one - tends to have greater consistency with internal consistency, with a notable influence of test length, sample homogeneity and dimension structure.

Keyword: Item discrimination, Classical Test Theory, discrimination coefficient, Extreme Groups Method , Item-Total Correlation Method, sample size, test length, reliability, validity.

مقدمة الدراسة:

يشـــكّل القياس والتقويم ركيزةً مركزيةً في المنظومة التربوية والنفســـية؛ إذ تُبنى عليهما قراراتٌ تتعلّق بالتصـــنيف والانتقاء، والتشـــخيص، والتحســـين، كما يرتبطان مباشـــرةً بجودة الأحكام الصـــادرة عن نتائج الاختبارات، وتؤكد الأدبيات السيكومترية أن جودة الاختبار لا تُختزل في "وجود أسئلة" بقدر ما تتحدد بخصائص القياس الكامنة في فقراته وفي الدرجة الكلية التي تنتج عنها، مثل الثبات والصدق وملاءمة صعوبات الفقرات لأهداف الاختبار والفئة المســـتهدفة (Nunnally & Bernstein, 1994؛ علّام، 2000)، ومن هذا المنطلق يصـــبح بناء الاختبارات عمليةً منهجية تتجاوز الصـــياغة اللغوية إلى تحليلٍ دقيق للفقرات وفق مؤشـــرات كمية تتيح الحكم على أدائها، وتوجيه قرارات الإبقاء أو التعديل أو الحذف (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991).

ويأتي تحليل الفقرات بوصفه مرحلةً محورية في تقويم جودة الاختبار ضمن إطار النظرية الكلاسيكية في القياس، فهو يزوّد باني الاختبار بمؤشـــرات تُبيّن مدى ملاءمة كل فقرة من حيث الصـــعوبة، والتمييز، والاتســـاق مع البنية الكلية للاختبار، وتُعد هذه المؤشـــرات أدواتٍ لضـــبط جودة القياس قبل الانتقال إلى مرحلة التطبيق الواســـع أو اعتماد الاختبار في قرارات تربوية أو نفســـية عالية الأثر (علّام، 2000؛ عودة، 1985). كما أن تحليل الفقرات يدعم الاتســـاق الداخلي للاختبار عبر الكشـــف عن الفقرات التي تعمل "ضـــد" الاتجاه العام للقياس أو التي لا تضيف معلومات مفيدة للدرجة الكلية (Crocker & Algina, 1986).

ويتخذ "تمييز الفقرة" موقعًا خاصًـا بين مؤشـرات جودة الفقرة؛ لأنه يعكس قدرة الفقرة على التفريق بين ذوي الأداء المرتفع وذوي الأداء المنخفض في الســـمة أو التحصـــيل المقاس، وتذهب الأدبيات إلى أن الفقرة الجيدة هي التي ترتفع احتمالية الإجابة الصـــحيحة عنها لدى ذوي الدرجات الكلية الأعلى، وتنخفض لدى ذوي الدرجات الكلية الأدنى، بما يجعلها عنصـــرًا داعمًا للدقة والاتســـاق الداخلي، ومؤشـــرًا عمليًا على جودة الفقرة في ســـياق الاختبار ككل (Ebel & Frisbie, 1991؛ Lord & Novick, 1968). غير أن تقدير التمييز ليس واحدًا بالضـــرورة؛ إذ تتعدد طرق حســـابه داخل النظرية الكلاســـيكية في القياس، وأشـــهرها طريقة المجموعتين المتطرفتين، والطريقة الارتباطية المعتمدة على ارتباط الفقرة بالدرجة الكلية، وهو ما يفتح المجال لمســـاءلة الفروق النظرية بين الطريقتين، وشروط ملاءمتهما.

مشكلة الدراسة:

تتمثل مشـــكلة هذه الدراســـة في وجود تباينٍ نظري متوقع في تقديرات تمييز الفقرة تبعًا للطريقة المعتمدة في الحساب، بما قد يقود إلى أحكام مختلفة حول جودة الفقرات نفسها، فطريقة المجموعتين المتطرفتين تقوم على مقارنة أداء الفقرة بين مجموعتين تُشتقان من التوزيع الكلي للدرجات (غالبًا أعلى وأدنى نسبة محددة)، بينما تقوم الطريقة الارتبـاطيـة على تقـدير قوة العلاقـة بين أداء المفحوص على الفقرة ودرجتـه الكليـة (Kelley, 1939؛

Crocker & Algina, 1986). ويترتب على هذا الاختلاف البنائي أن كل طريقة قد "تلتقط" جانبًا مغايرًا من التمييز: الأولى تُبرز التباعد بين طرفي التوزيع، والثانية تُعبّر عن اتساق الفقرة عبر كامل مدى الدرجات.

وتزداد المشكلة تعقيدًا عند إدخال عاملين مؤثرين في خصائص التقدير: حجم العينة وطول الاختبار، فمن حيث حجم العينة، تشير الكتابات السيكومترية إلى أن مؤشرات الفقرات، وخاصة المؤشرات الارتباطية، تتأثر باستقرار التقدير الإحصائي؛ فالعينات الصغيرة تميل إلى إعطاء تقديرات أكثر تقلبًا وأوسع خطًا معياريًا، كما أن طريقة المجموعتين المتطرفتين تقلّص فعليًا حجم البيانات المستخدمة لأنها تعتمد على جزء من المفحوصين (عادةً 27% من الأعلى و27% من الأدنى)، ما قد يزيد حساسية التقدير لتذبذب العينات عندما تكون الأعداد محدودة (Kelley, 1939؛ Brennan, 1972). ومن حيث طول الاختبار، فإن الدرجة الكلية التي تدخل في حساب الارتباط ليست معطًى ثابت الخصائص، فهي تتأثر بثبات الاختبار وعدد فقراته، وتُشير الأدبيات إلى أن زيادة طول الاختبار غالبًا ما ترفع ثباته وفق منطق التنبؤ بالثبات، ما ينعكس على مؤشرات الاتساق الداخلي والارتباطات المعتمدة على الدرجة الكلية (Nunnally & Bernstein, 1994؛ علّام، 2000). كما أن قِصر الاختبار يجعل ارتباط الفقرة بالمجموع أكثر عرضة للتضخيم إذا لم تُصحح الدرجة الكلية باستبعاد الفقرة نفسها، وهو ما استدعى تطوير تصحيحات منهجية لارتباط الفقرة بالمجموع (Henrysson, 1963).

وعليه تتحدد المشكلة في الحاجة إلى فهمٍ نظري أدق: أي الطريقتين يُتوقع أن تكون أكثر اتساقًا وثباتًا، وأقل تحيزًا، عندما يتغير حجم العينة وطول الاختبار، وما الحدود التي ينبغي مراعاتها عند تفسير تمييز الفقرة استنادًا إلى كل طريقة.

أسئلة الدراسة:

1. ما الأسس النظرية لكل من طريقة المجموعتين المتطرفتين والطريقة الارتباطية في حساب تمييز الفقرة؟
2. كيف يُتوقع نظريًا أن يتأثر تقدير تمييز الفقرة بحجم العينة في كل طريقة؟
3. كيف يُتوقع نظريًا أن يتأثر تقدير التمييز بطول الاختبار؟
4. ما أوجه الاتفاق والاختلاف بين الطريقتين، ومتى تكون كل طريقة أنسب من منظور نظري؟

أهداف الدراسة: تهدف هذه الدراسة إلى بناء مقارنة مفاهيمية ونظرية دقيقة بين طريقتي تمييز الفقرة الأكثر شيوعًا في النظرية الكلاسيكية في القياس، وتحليل كيف يمكن لتباين حجم العينة وطول الاختبار أن يؤثر في خصائص التقدير مثل: الثبات والدقة والتحيز، كما تهدف إلى صياغة إطار إرشادي نظري يساعد الباحثين، وبناة الاختبارات على اختيار مؤشر التمييز الأكثر ملاءمة تبعًا لظروف بناء الاختبار وتطبيقه (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991).

أهمية الدراسة:

تنبع أهمية الدراسة من كونها تسد فجوة نظرية تتصل بقرارات شائعة في بناء الاختبارات؛ إذ إن اختيار مؤشر تمييز بعينه قد يقود إلى قرارات مختلفة بشأن الفقرات (حذفًا أو تعديلًا أو إبقاءً)، ومن ثم يؤثر في جودة الاختبار وخصائصه السيكومترية. وتفيد الدراسة:

1. الباحثين في القياس النفسي والتربوي من خلال توضيح الفروق النظرية وحدود الاستدلال لكل طريقة (Nunnally & Bernstein, 1994).
2. بناة الاختبارات عبر تقديم أساسٍ نظري لتفسير نتائج تحليل الفقرات في ضوء حجم العينة وطول الاختبار (علّام، 2000؛ عودة، 1985).
3. مقررات القياس والتقويم عبر تنظيم مفاهيم تمييز الفقرة وإبراز شروط تطبيق مؤشرات الفقرات بصورة منهجية (ملحم، 2002).

حدود الدراسة:

حدود موضوعية: تقتصر الدراسة على تمييز الفقرة ضمن النظرية الكلاسيكية في القياس، دون التوسع في مؤشرات نظرية الاستجابة للفقرة.

حدود منهجية: دراسة نظرية تحليلية تعتمد المقارنة المفاهيمية والحجاج السيكومتري دون تطبيقات ميدانية أو محاكاة عددية.

مصطلحات الدراسة:

ورد في هذه الدراسة عدد من المصطلحات يمكن تعريفها كما يلي:

الفقرة: وحدة قياس مفردة داخل الاختبار تمثل سؤالًا أو مثيرًا يقاس من خلاله أداء المفحوص (Ebel & Frisbie, 1991).

تمييز الفقرة: قدرة الفقرة على التفريق بين المفحوصين ذوي الدرجات الكلية المرتفعة والمنخفضة، بما يعكس اتساق أداء الفقرة مع الاتجاه العام للاختبار (Crocker & Algina, 1986).

صعوبة الفقرة: نسبة أو احتمال الإجابة الصحيحة عن الفقرة في عينة معينة ضمن الاختبارات الموضوعية، وتُفهم بوصفها مؤشرًا لوضع الفقرة على متصل الأداء (Ebel & Frisbie, 1991).

الدرجة الكلية: مجموع درجات المفحوص على فقرات الاختبار، وتمثل مؤشرًا مركبًا للأداء على السمة / التحصيل المقاس (علّام، 2000).

طريقة المجموعتين المتطرفتين (27%): أسلوب لتقدير تمييز الفقرة عبر مقارنة أداء الفقرة بين أعلى نسبة وأدنى نسبة من المفحوصين وفق الدرجة الكلية، وقد ارتبطت نسبة 27% بتبريرات سيكومترية كلاسيكية لتحسين التمايز بين المجموعتين (Kelley, 1939؛ Brennan, 1972).

الطريقة الارتباطية (ارتباط فقرة – مجموع): تقدير تمييز الفقرة عبر معامل ارتباط بين درجة الفقرة (ثنائية عادةً: صحيح / خطأ) والدرجة الكلية، ويشمل ذلك الارتباط الثنائي النقطي والارتباط الثنائي المتسلسل، مع إمكانية تصحيح الارتباط باستبعاد الفقرة من المجموع لتقليل التضخيم (Crocker & Algina, 1986؛ Henrysson, 1963).

حجم العينة: عدد المفحوصين الذين تُستخرج منهم مؤشرات الفقرات والاختبار، ويرتبط باستقرار تقدير المؤشرات ودقتها (Crocker & Algina, 1986).

طول الاختبار: عدد فقرات الاختبار، ويرتبط عادةً بثبات الدرجة الكلية وخصائص الاتساق الداخلي عند ثبات جودة الفقرات (Nunnally & Bernstein, 1994؛ علّام، 2000).

الثبات: درجة اتساق القياس عبر تكرار القياس أو عبر اتساق مكونات الاختبار، بوصفه شرطًا لازمًا لجودة الاستدلال بالدرجات (Nunnally & Bernstein, 1994).

الاتساق الداخلي: صورة من صور الثبات تشير إلى اتساق أداء الفقرات في قياس البناء نفسه، ويُستدل عليه من مؤشرات مثل: معاملات الاتساق والارتباطات المصححة بين الفقرة والمجموع (علّام، 2000؛ Henrysson, 1963).

الخلفية النظرية لتمييز الفقرة ضمن النظرية الكلاسيكية في القياس:

تحليل الفقرات: المفهوم والغايات:

يُقصد بتحليل الفقرات مجموعة الإجراءات النظرية والإحصائية التي تهدف إلى فحص أداء فقرات الاختبار بوصفها وحدات قياس مفردة، بما يسمح بتقدير مدى إسهام كل فقرة في وظيفة الاختبار الكلية، ولا ينحصر تحليل الفقرات في كونه "خطوة فنية" تالية لكتابة الفقرات، بل يمثل جزءًا من منطق بناء الاختبار نفسه؛ إذ يربط بين صياغة الفقرة ومحتواها من جهة، وبين خصائص الدرجة الكلية وقوة الاستدلال بها من جهة أخرى (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991). وتؤكد الأدبيات أن الاختبار قد يبدو متماسكًا شكليًا، لكنه قد يفتقر سيكومتريًا إلى الدقة إذا احتوى فقرات لا تميز بين الأفراد، أو فقرات شديدة السهولة / الصعوبة، أو فقرات تعمل في اتجاه مختلف عن البناء المقاس (علّام، 2000؛ Nunnally & Bernstein, 1994).

وتتجسد الغايات الأساسية لتحليل الفقرات في ثلاث دوائر مترابطة:

أولًا: تحسين جودة الاختبار عبر تنقية الفقرات وتجويدها، فالفقرة الجيد لا يقتصر على صحة المحتوى، بل ينبغي أن يحقق أداءً قياسيًا مقبولًا من حيث مؤشرات الصعوبة والتمييز، وأن ينسجم مع بقية الفقرات في قياس البناء ذاته (Ebel & Frisbie, 1991).

ثانيًا: رفع الثبات؛ إذ إن الفقرات التي لا تميز أو التي تُدخل ضوضاء قياس عالية تميل إلى تقليل اتساق الدرجات، بينما يسهم اختيار فقرات أكثر تمييزًا في تقوية الاتساق الداخلي بوصفه صورة جوهرية من صور الثبات في الاختبارات التي تُصحح بطريقة المجموع (Nunnally & Bernstein, 1994؛ علّام، 2000).

ثالثًا: دعم الصدق، فالمؤشرات الكمية للفقرات تُعد أدوات مساندة للأدلة الخاصة بصدق البناء، لأنها تكشف ما إذا كانت الفقرات تتصرف وفق ما يتوقعه البناء النظري، وما إذا كانت تسهم فعليًا في قياس السمة المستهدفة بدل قياس عوامل عرضية (AERA et al., 2014؛ Lord & Novick, 1968).

تمييز الفقرة: المفهوم النفسي – القياسي:

يمثل تمييز الفقرة مفهومًا يتوسط بين جانبين: جانب نفسي يتعلق بكيفية استجابة الأفراد للفقرة وفق مستوى السمة، وجانب قياسي يتعلق بقدرة الفقرة على إنتاج فروق قابلة للرصد بين الأفراد، ومن ثم فإن تمييز الفقرة يُعرّف بوصفه قدرة الفقرة على الفصل بين ذوي الأداء المرتفع وذوي الأداء المنخفض على الدرجة الكلية، أو على البناء المقاس، بحيث ترتفع احتمالية الأداء الأفضل على الفقرة لدى ذوي الدرجة الكلية الأعلى مقارنة بذوي الدرجة الكلية الأدنى (Cronbach, 1951)، ويُعد هذا الفصل جوهرًا وظيفيًا للفقرة، فإذا لم تفرّق الفقرة بين الأفراد، فإنه لا يضيف معلومات قياسية نافعة للاختبار، حتى إن بدا مناسبًا من حيث المحتوى أو الصياغة.

وتتضح الطبيعة القياسية لتمييز الفقرة عندما يُنظر إليه كجزء من اتساق الفقرة مع البناء المقاس، فالفقرة المميز لا يكتفي بإحداث تباين في الاستجابات، وإنما يُفترض أن يكون تباينه "منظمًا" وفق البناء: أي أن الفروق في أداء الأفراد على الفقرة تعكس فروقًا في السمة المراد قياسها، لا فروقًا عشوائية أو فروقًا ناتجة عن عوامل ثانوية مثل: الغموض اللغوي، أو التحيز الثقافي، أو الاعتماد على مهارات غير مستهدفة (AERA et al., 2014؛ علّام، 2000). ولهذا يرتبط تمييز الفقرة بمفهوم البعدية: فكلما كان الاختبار أقرب إلى قياس بناء واحد متجانس، كان من المتوقع أن تظهر الفقرات الجيدة ارتباطًا واتساقًا أعلى مع الدرجة الكلية، وأن تُظهر قدرة أفضل على التفريق بين الأفراد على ذلك البناء (Nunnally & Bernstein, 1994).

كما أن التمييز يحمل بعدًا إجرائيًا يختلف باختلاف طريقة تقديره داخل النظرية الكلاسيكية في القياس. فطريقة المجموعتين المتطرفتين تنظر إلى التمييز بوصفه فرقًا في نسب الاستجابة بين طرفي التوزيع، وقد ارتبطت فكرة اختيار نسبة من الأعلى والأدنى بمنطق زيادة التباين بين المجموعتين لتحقيق حساسية أكبر للفروق (Kelley, 1939؛ Brennan, 1972). أما الطريقة الارتباطية فتنظر إلى التمييز بوصفه اتساقًا خطيًا بين أداء الفرد على الفقرة ودرجته على الاختبار، الأمر الذي يجعل التمييز وثيق الصلة بالاتساق الداخلي وبطول الاختبار وبكيفية تكوين الدرجة الكلية (Crocker & Algina, 1986؛ Henrysson, 1963). ويعني هذا أن مفهوم "تمييز الفقرة" ثابت من حيث المبدأ، لكنه قد يتجسد تقديريًا بصورة مختلفة بحسب منطق المؤشر المستخدم، ما يستدعي الانتباه إلى معنى التمييز الذي يقدمه كل مؤشر.

العلاقة بين تمييز الفقرة وصعوبة الفقرة:

تُعد العلاقة بين تمييز الفقرة وصعوبة الفقرة من أكثر العلاقات حضورًا في تحليل الفقرات ضمن النظرية الكلاسيكية في القياس، لأن الصعوبة تحدد مستوى الأداء الذي تستهدفه الفقرة، بينما يحدد التمييز مقدار المعلومات التفريقية التي تنتجها، وتبرز الإشكالية حين تبلغ الصعوبة حدًا شديد الارتفاع أو شديد الانخفاض؛ إذ تميل الفقرة في هاتين الحالتين إلى إضعاف التمييز لاعتبارات منطقية وإحصائية في آن واحد (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991).

عندما تكون الفقرة شديدة السهولة، فإن معظم الأفراد – بمن فيهم ذوو الأداء المنخفض – يجيبون عنها بصورة صحيحة، فتتقلص الفروق بين الأفراد على الفقرة، ويصبح من الصعب أن تقوم الفقرة بوظيفتها التفريقية، وبالمثل، عندما تكون الفقرة شديدة الصعوبة، فإن معظم الأفراد – بمن فيهم ذوو الأداء المرتفع – يفشلون فيها، فتتساوى الاستجابات تقريبًا، ويتضاءل التباين اللازم لإظهار فروق ذات معنى، وعلى مستوى المؤشرات، فإن ضعف التباين في الاستجابات يؤدي عادةً إلى انخفاض مؤشرات التمييز لأن التمييز يحتاج إلى قدر كافٍ من التباين في أداء الأفراد على الفقرة كي "يظهر" (Kuder & Richardson, 1937)

ومن ثم يتصور الأدب السيكومتري علاقةً نظرية تفيد بأن أعلى مستويات التمييز غالبًا ما تتحقق عند مستويات صعوبة متوسطة نسبيًا، حيث تتوزع الاستجابات بين الصحيح والخطأ بصورة تسمح بظهور الفروق بين ذوي الدرجات الكلية المرتفعة والمنخفضة (Ebel & Frisbie, 1991). ولا يعني ذلك أن كل بند متوسط الصعوبة سيكون مميزًا، أو أن الفقرات غير المتوسطة لا يمكن أن تكون مفيدة، فقد تُبنى اختبارات لأغراض تشخيصية أو محكية تستلزم وجود فقرات سهلة جدًا أو صعبة جدًا، إلا أن النقطة النظرية الأساسية هي أن حدود التمييز ليست مستقلة عن موقع الفقرة على متصل الصعوبة، وأن تفسير التمييز ينبغي أن يراعي هذا التداخل البنائي بين المؤشرين (علّام، 2000؛ Crocker & Algina, 1986).

تمييز الفقرة والثبات والصدق:

يرتبط تمييز الفقرة ارتباطًا وثيقًا بالثبات، لا بوصفه بديلًا عنه، بل بوصفه أحد المكونات التي تُسهم في تحقيق الاتساق الداخلي عندما يُفترض أن الاختبار يقيس بناءً واحدًا أو بنية متجانسة، فالاتساق الداخلي يتأثر بمدى ترابط الفقرات واشتراكها في قياس المصدر نفسه للتباين، والفقرات التي تميز عادةً ما تكون أكثر اتساقًا مع الدرجة الكلية، ومن ثم ترفع من تجانس الاختبار وتقلل من الضوضاء التي تُضعف الثبات (Spearman, 1910) وعلى العكس، فإن الفقرات ضعيفة التمييز قد تعمل كفقرات "حيادية" لا تضيف معلومات، أو كفقرات "مخالِفة" تخفض الاتساق إذا كانت تقيس مهارة أخرى أو تتضمن غموضًا يولد استجابات غير منظمة.

ويظهر هذا الارتباط بوضوح في المؤشرات الارتباطية التي تجعل التمييز جزءًا مباشرًا من منطق الاتساق الداخلي. غير أن هذا الترابط يستدعي حذرًا منهجيًا: فإذا كانت درجة الفقرة تدخل ضمن الدرجة الكلية التي يُحسب معها الارتباط، فإن الارتباط قد يتضخم بسبب "تداخل" الفقرة في المجموع، وهو ما دعا إلى تطوير صيغ تصحيحية تعتمد على ارتباط الفقرة مع "بقية الاختبار" بدل الاختبار كله (Henrysson, 1963). ويترتب

على ذلك أن الحديث عن تمييز الفقرة بوصفه دعمًا للثبات ينبغي أن يكون حديثًا مشروطًا بمراعاة طريقة الحساب، وتكوين الدرجة الكلية وطول الاختبار (Crocker & Algina, 1986).

أما من زاوية الصدق، فإن تمييز الفقرة يقدم إشارة نظرية مساندة لأدلة صدق البناء وصدق المحتوى، لكنه لا يكفي وحده لإثبات الصدق، فصدق المحتوى يُستمد أساسًا من تمثيل الفقرات لمجال المحتوى وأهدافه، غير أن الفقرات التي لا تميز قد تشير إلى خلل في مواءمة المحتوى لمستوى المفحوصين أو إلى ضعف في صياغة الفقرة، ما ينعكس سلبًا على صلاحية الاستدلال بالنتائج (AERA et al., 2014؛ علّام، 2000). وفي صدق البناء، تتوقع النظرية أن الفقرات التي تقيس البناء نفسه ستظهر نمطًا من التمييز والاتساق يدعم تفسير الدرجة الكلية باعتبارها مؤشرًا على ذلك البناء، بينما الفقرات التي تُظهر تمييزًا ضعيفًا أو غير مستقر قد توحي بأن الفقرة يقيس عاملًا آخر أو يتأثر بعوامل سياقية (Nunnally & Bernstein, 1994). وعليه، يُفهم تمييز الفقرة كقطعة ضمن "حزمة أدلة" للصدق، لا كحكم نهائي منفرد (AERA et al., 2014).

افتراضات النظرية الكلاسيكية المؤثرة على التمييز:

تقوم النظرية الكلاسيكية في القياس على تصور أساسه أن الدرجة المشاهدة تتكون من درجة حقيقية وخطأ قياس، وأن الخطأ – في صورته المثالية – عشوائي متوسطه صفري ولا يرتبط بالدرجة الحقيقية، ويترتب على هذا التصور أن مؤشرات الفقرات، ومنها التمييز، تتأثر بحجم الخطأ العشوائي: فكلما ارتفع الخطأ، تقل قدرة الدرجة الكلية على تمثيل البناء بدقة، وتضعف قدرة أي مؤشر يعتمد على الدرجة الكلية في إظهار علاقة منظمة بين أداء الفقرة ومستوى السمة (Brown, 1910) وبعبارة أخرى، لا يُتوقع لتمييز الفقرة أن يكون مستقرًا في سياق قياس شديد الضوضاء، لأن الضوضاء تقلل من انتظام الفروق التي يُفترض أن يلتقطها التمييز.

ويتصل بذلك افتراض "أحادية البعد" بوصفه شرطًا تقريبيًا في كثير من تطبيقات النظرية الكلاسيكية في القياس، خاصة عندما يُراد تفسير الدرجة الكلية كمؤشر واحد لبناء واحد، فإذا كان الاختبار متعدد الأبعاد بصورة قوية، فإن الدرجة الكلية تصبح تجميعًا لمصادر تباين مختلفة (Gulliksen, 1950)، وقد يظهر بند ما ضعيف التمييز لا لأنه رديء، بل لأنه يقيس بُعدًا مختلفًا عن البعد الغالب في المجموع، كما قد يظهر بند عالي التمييز في بعد فرعي لا يتسق مع غرض القياس الأساسي، وهو ما يجعل تفسير التمييز دون فحص بنية الاختبار عرضة للالتباس (Crocker & Algina, 1986؛ AERA et al., 2014). لذا تنبّه الأدبيات إلى ضرورة ربط نتائج تحليل الفقرات بتصور نظري للبناء، وبخطة بناء الاختبار، لا الاكتفاء بمؤشرات منعزلة (علّام، 2000).

ومن الافتراضات المؤثرة كذلك خصائص العينة، وبخاصة تجانسها ومدى تباينها على السمة، فمؤشرات التمييز – وخاصة الارتباطية – تتأثر بتقييد مدى الدرجات: فإذا كانت العينة متجانسة جدًا (مثل: مجموعة منتقاة مسبقًا ذات مستوى متقارب)، تقل الفروق في الدرجة الكلية، ويضعف ظهور العلاقات بين الفقرات والدرجة الكلية، حتى لو كانت الفقرات جيدة في مجتمع أكثر تباينًا (Allen & Yen, 1979).

وفي المقابل، قد تُظهر العينة ذات التباين الواسع تقديرات تمييز أعلى بسبب وفرة الفروق التي يمكن للفقرة أن يلتقطها، كما أن طريقة المجموعتين المتطرفتين تتأثر أيضًا بخصائص التوزيع؛ لأن تحديد "الأعلى" و"الأدنى" يعتمد

على انتظام توزيع الدرجات وعلى دقة فرز الأفراد، وهي دقة تتأثر بالعينة وبطول الاختبار معًا (Kelley, 1939؛ Brennan, 1972). وعليه، فإن تمييز الفقرة في إطار النظرية الكلاسيكية لا يُقرأ بمعزل عن شروط القياس: مقدار الخطأ العشوائي، وبنية الاختبار من حيث البعدية، وخصائص العينة من حيث التباين والتجانس، وهي شروط تشكل الخلفية الضرورية لأي مقارنة لاحقة بين طرائق تقدير التمييز.

الطريقة التقليدية (المجموعتين المتطرفتين) في حساب تمييز الفقرة:

1- الفكرة الأساسية للطريقة التقليدية:

تقوم الطريقة التقليدية في تقدير تمييز الفقرة – المعروفة بطريقة المجموعتين المتطرفتين – على منطق بسيط مفاده أن الفقرة تكون "مميِّزة" إذا كان ذوو الأداء المرتفع على الاختبار يجيبون عنها بصورة صحيحة أكثر من ذوي الأداء المنخفض، ولتحويل هذا المنطق إلى إجراء قياسي، يُرتَّب المفحوصون وفق الدرجة الكلية للاختبار، ثم يُقسَّمون إلى مجموعتين: مجموعة عليا تمثل أعلى الدرجات، ومجموعة دنيا تمثل أدنى الدرجات، بعد ذلك تُقارن استجابات المجموعتين على الفقرة محل التحليل، ويُستخلص مؤشر يعكس حجم الفارق بينهما (Ebel & Frisbie, 1991؛ Crocker & Algina, 1986).

ويتأسس هذا الإجراء على افتراض ضمني مفاده أن الدرجة الكلية تُعد مؤشرًا كافيًا لمستوى المفحوص على البناء المقاس، ومن ثم فإن اختيار الطرفين يعظم التباين في مستوى السمة، ويجعل الفروق في الاستجابة للفقرة أكثر وضوحًا، ومع أن هذا التقسيم يُستخدم بصورة شائعة في تحليل الفقرات داخل إطار النظرية الكلاسيكية في القياس، إلا أن دلالته تعتمد على شروط عدة تتصل بثبات الدرجة الكلية، وبمدى تباين العينة، وبكيفية اختيار نسبة المجموعتين (Haladyna, 2004). وعليه، فإن الطريقة التقليدية ليست مجرد "قاعدة حسابية"، بل هي تصور إجرائي لمعنى التمييز بوصفه تباعدًا بين طرفين من توزيع الدرجات.

2- صيغ ومؤشرات التمييز في هذه الطريقة:

يُشتق مؤشر التمييز في الطريقة التقليدية عادةً من مقارنة نسب الإجابة الصحيحة في المجموعتين العليا والدنيا، ويُعبَّر عنه في أبسط صوره بأنه: نسبة الإجابة الصحيحة في المجموعة العليا ناقص نسبة الإجابة الصحيحة في المجموعة الدنيا، فإذا كانت النسبة في المجموعة العليا أكبر بكثير من النسبة في المجموعة الدنيا، دل ذلك على أن الفقرة تفصل بين ذوي المستوى المرتفع والمنخفض، وهو جوهر التمييز في هذا المنظور (Ebel & Frisbie, 1991). ويتميّز هذا المؤشر بسهولة تفسيره: فالقيمة الموجبة الكبيرة تعني تمييزًا أفضل، والقيمة القريبة من الصفر تشير إلى ضعف التمييز، أما القيمة السالبة – عندما تكون نسبة الإجابة الصحيحة في المجموعة الدنيا أعلى من العليا – فقد تشير إلى مشكلة جوهرية في الفقرة مثل الغموض أو وجود مفتاح إجابة غير صحيح أو أن الفقرة تقيس شيئًا مختلفًا عمّا يقيسه الاختبار (Crocker & Algina, 1986).

غير أن السؤال المنهجي الحاسم هنا يتعلق بنسبة المجموعتين: لماذا تُستخدم نسبة 27% بصورة متكررة في الأدبيات؟ تُنسب هذه النسبة إلى مبرر سيكومتري كلاسيكي مفاده أن اختيار طرفين بحجم يقارب 27% لكل

طرف يُحقق توازنًا بين هدفين متعارضين: (1) تعظيم الفروق المتوقعة بين الطرفين من خلال اختيار مفحوصين شديدي الاختلاف في الدرجة الكلية، و(2) الحفاظ على حجم كافٍ لكل مجموعة بما يقلل من تقلب التقدير الإحصائي الذي ينجم عن صِغر العدد (Kelley, 1939). فكلما صغرت النسبة كثيرًا زادت "حدة التطرف"، لكن على حساب العدد، وكلما كبرت النسبة اقترب الطرفان من الوسط وقلّت الحساسية للفروق. ومن هذا المنطلق ظهرت نسبة 27% كحل وسط شائع الاستخدام في التحليل التقليدي (Downing & Haladyna, 2006).

وتجدر الإشارة إلى أن هذا التبرير لا يعني أن نسبة 27% قاعدة ثابتة لا تُراجع؛ إذ قد تُستخدم نسب أخرى تبعًا لحجم العينة وطبيعة الاختبار وغرضه، كما أن بعض المعالجات طوّرت صيغًا تعميمية لمؤشر المجموعتين تقلل من الاعتماد الصارم على نسبة بعينها، وتتعامل مع الصعوبة بطريقة تجعل المؤشر أقل حساسية لبعض خصائص الفقرة (Brennan, 1972). لكن يبقى المبدأ في الطريقة التقليدية واحدًا: تقدير التمييز بوصفه فرقًا ملحوظًا في الأداء على الفقرة بين طرفين من توزيع الدرجات الكلية.

3- خصائص الطريقة التقليدية:

تتسم الطريقة التقليدية بعدة خصائص تجعلها شائعة في البيئات التربوية وفي بناء الاختبارات الصفية، أول هذه الخصائص بساطة الفهم، فهي تُحوِّل مفهوم التمييز إلى مقارنة مباشرة بين مجموعتين واضحتين، بما يسمح للباحث أو المعلم بتفسير النتائج دون تعقيد إحصائي كبير، كما أن سهولة التفسير تُعد ميزة تطبيقية مهمة؛ لأن قيمة المؤشر يمكن ربطها مباشرةً بوظيفة الفقرة في التفريق بين مرتفعي ومنخفضي التحصيل على الاختبار (Ebel & Frisbie, 1991).

ومن مزاياها أيضًا أنها تُبرز الفروق "المرئية" في الطرفين؛ فإذا كان هدف التحليل تربويًا سريعًا – مثل انتقاء فقرات مناسبة لاختبار صفي – فإن مقارنة الطرفين قد تمنح صورة أولية مفيدة حول الفقرات التي لا تُحدث فرقًا بين المفحوصين، وفي هذا السياق، تُعد الطريقة التقليدية متسقة مع كثير من ممارسات التقويم التكويني التي تتطلب قرارات سريعة حول تحسين الاختبار (علّام، 2000).

لكن هذه الطريقة تحمل حدودًا منهجية مهمة، أبرزها فقدان معلومات الوسط؛ إذ تُهمل استجابات نسبة كبيرة من المفحوصين تقع في منتصف التوزيع، رغم أنها قد تحمل دلالات مهمة حول كيفية تدرج الأداء عبر مستويات السمة، وبسبب هذا الإهمال قد تبدو فقرة ما ضعيفة التمييز لأنها لا تُحدث فرقًا حادًا بين الطرفين، رغم أنها قد تعمل بصورة جيدة عبر الوسط أو عند مستويات متوسطة من الصعوبة (Crocker & Algina, 1986). كما أن الطريقة حساسة لطريقة "الفرز"؛ إذ إن أي تغير في معيار التقسيم أو في نسبة الطرفين قد يقود إلى تغير في قيمة المؤشر، ما يطرح سؤال الاستقرار عبر العينات (Thorndike & Thorndike-Christ, 2010).

ومن حدودها كذلك تأثرها بالعينات الصغيرة؛ لأن تقسيم العينة إلى طرفين يقلص فعليًا عدد المفحوصين الذين يُستخدمون في الحساب، فإذا كانت العينة الأصلية محدودة، تصبح المجموعتان أصغر، ويزداد أثر التقلب العشوائي في نسب الإجابة الصحيحة، فتتذبذب تقديرات التمييز بصورة أكبر (Brennan, 1972؛ Nunnally & Bernstein, 1994). وتزداد هذه المشكلة عندما تكون الفقرة نفسها شديدة السهولة أو شديدة الصعوبة، إذ تقل الحالات "المفيدة" لإظهار الفروق بين المجموعتين.

4- مصادر التحيز والخطأ في التقدير:

تتعدد مصادر التحيز أو الخطأ في تقدير التمييز بالطريقة التقليدية، ويمكن تنظيمها في ثلاثة محاور: اختيار نسبة المتطرفين، شكل توزيع الدرجات، وخصائص العينة.

أولًا: اختيار نسبة المتطرفين يمثل قرارًا منهجيًا مؤثرًا، فالنسبة الأصغر قد تزيد التباين بين الطرفين لكن تقلل حجم كل مجموعة، والنسبة الأكبر تزيد الحجم لكنها تقرّب الطرفين من الوسط، هذا يعني أن المؤشر يتأرجح بين الحساسية للفروق وبين الاستقرار الإحصائي، وأن أي اختيار غير ملائم لحجم العينة قد ينتج تقديرًا مضخمًا أو منقوصًا (Kelley, 1939؛ Brennan, 1972). كما أن تجاهل هذه المفاضلة قد يقود إلى مقارنة غير عادلة بين اختبارات أو عينات مختلفة، فمؤشر التمييز المستخرج من مجموعتين صغيرتين جدًا لا يُتوقع أن يمتلك الاستقرار نفسه المستخرج من مجموعتين أكبر.

ثانيًا: توزيع الدرجات الكلية قد يشكل مصدرًا خفيًا للخطأ، فإذا كان توزيع الدرجات منحرفًا بشدة (مثل: اختبار سهل جدًا نتج عنه تراكم الدرجات في الطرف الأعلى)، فإن "المجموعة العليا" قد لا تختلف كثيرًا عن "الوسط"، وقد تصبح المجموعة الدنيا صغيرة في تنوعها أو محدودة في تشتتها، وفي هذه الحالة قد تتقلص الفروق بين المجموعتين على نحو يقلل تمييز الفقرة، حتى لو كانت الفقرة جيدة ضمن اختبار أكثر ملاءمة لمستوى المفحوصين (Lord & Novick, 1968). كذلك، إذا كان الاختبار شديد الصعوبة وتراكمت الدرجات في الطرف الأدنى، تتعرض العملية لخلل مشابه لكن باتجاه معاكس، ومن ثم، فإن المؤشر التقليدي لا يُقرأ بمعزل عن مدى ملاءمة صعوبة الاختبار عامةً لمستوى العينة.

ثالثًا: تجانس العينة وتقييد المدى يؤثران في جميع مؤشرات تحليل الفقرات، والطريقة التقليدية ليست استثناءً، فالعينة المتجانسة جدًا تُنتج فروقًا أقل في الدرجة الكلية، ما يجعل عملية "التطرف" أقل معنى؛ إذ قد يصبح الفرق بين المجموعتين العليا والدنيا فرقًا محدودًا في البناء المقاس، فتضعف قدرة الفقرة على إظهار تباعد واضح بينهما (Nunnally & Bernstein, 1994). ويضاف إلى ذلك أثر خطأ القياس العشوائي: إذا كانت الدرجة الكلية منخفضة الثبات، فإن تصنيف المفحوصين إلى عليا ودنيا يصبح أقل دقة، فتدخل حالات من "سوء التصنيف" تضعف الفروق الحقيقية بين المجموعتين، وتخفض تمييز الفقرة تقديريًا (Anastasi & Urbina, 1997). وهذا يوضح أن الطريقة التقليدية تعتمد ضمنيًا على جودة الدرجة الكلية بوصفها معيارًا للفرز، وأن أي قصور في هذا المعيار ينعكس على مؤشر التمييز.

5- متى تكون الطريقة التقليدية ملائمة نظريًا؟

يمكن القول إن ملاءمة الطريقة التقليدية تتحدد نظريًا عندما تكون الحاجة قائمة إلى مؤشر واضح التفسير ومباشر الاستدلال، وحين تكون ظروف القياس تسمح باستقرار معقول في تقدير الفروق بين المجموعتين، ففي الاختبارات القصيرة جدًا قد تكون بعض المؤشرات الارتباطية أقل استقرارًا لأن الدرجة الكلية تكون محدودة المحتوى وأكثر عرضة للضوضاء، بينما قد يوفّر أسلوب المقارنة بين طرفين قراءة أولية لمدى "اختلاف" الأداء على الفقرة بين مرتفعي ومنخفضي الدرجة الكلية، بشرط ألا تكون العينة صغيرة للغاية وبشرط ملاءمة صعوبة الاختبار (Lord & Novick, 1968 ؛Ebel & Frisbie, 1991).

كما تكون الطريقة مناسبة نسبيًا عندما تتوفر عينة متوسطة أو كبيرة تضمن حجمًا كافيًا لكل من المجموعتين بعد التقسيم، بحيث تقل التقلبات العشوائية في نسب الإجابة الصحيحة، وفي التطبيقات التربوية ذات الطابع السريع – مثل: بناء اختبار صفي أو اختبار مرحلي داخلي – قد تُعد الطريقة التقليدية خيارًا عمليًا من منظور نظري "إجرائي"، لأنها تترجم مفهوم التمييز إلى مقارنة سهلة، وتساعد في تنقية الفقرات التي لا تفصل بين الطرفين (علّام، 2000).

ومع ذلك، ينبغي التنبيه إلى أن استخدام الطريقة التقليدية يصبح أقل ملاءمة – من منظور نظري – عندما يكون الهدف بناء اختبار يُراد له اتساق داخلي مرتفع واستدلالات أدق عبر مدى واسع من الدرجات؛ لأن إهمال وسط التوزيع قد يحجب معلومات مهمة عن كيفية عمل الفقرة عبر مستويات السمة المختلفة (Crocker & Algina, 1986). كذلك، في سياقات القرارات عالية الأثر، يتوقع أن تكون الحاجة أكبر إلى تقديرات أكثر استقرارًا وأقل اعتمادًا على قرار "تقسيم" قد يختلف باختلاف العينة، وعليه، فإن ملاءمة الطريقة التقليدية ليست مطلقة، بل تُفهم بوصفها اختيارًا مقيدًا بشروط حجم العينة، وملاءمة الاختبار، وبهدف التحليل ذاته، وبالوعي بمصادر التحيز المحتملة التي قد تؤثر في تقدير تمييز الفقرة.

الطريقة الارتباطية (ارتباط فقرة – مجموع) في حساب تمييز الفقرة:

1- الفكرة الأساسية للطريقة الارتباطية:

تنطلق الطريقة الارتباطية في تقدير تمييز الفقرة من تصورٍ مختلف عن تصور المجموعتين المتطرفتين؛ إذ لا تُعرِّف التمييز بوصفه "فارقًا" بين طرفين من توزيع الدرجات، بل بوصفه قوة العلاقة المنتظمة بين أداء الفرد على الفقرة وأدائه على الاختبار ككل، فكلما كانت الفقرة تقيس البناء نفسه الذي تقيسه بقية فقرات الاختبار، يُتوقع أن يتجه أداؤها في الاتجاه ذاته الذي تتجه إليه الدرجة الكلية: من يحصل على درجة كلية أعلى تكون لديه احتمالية أعلى للإجابة الصحيحة أو الأداء الأفضل على الفقرة، والعكس صحيح (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991).

ويُفهم "الارتباط" هنا بوصفه تلخيصًا لمدى اتساق الفقرة مع المحك الداخلي الأكثر قربًا في إطار النظرية الكلاسيكية في القياس، وهو الدرجة الكلية، وبذلك تُصبح الفقرة عالية التمييز هي الفقرة التي تُسهم في بناء الدرجة

الكلية بصورة منظمة وغير عشوائية، وتُحافظ على نمط متسق عبر مدى الدرجات كله، لا في طرفيه فقط (Lord & Novick, 1968). وهذا المنطق يجعل الطريقة الارتباطية أقرب إلى فكرة الاتساق الداخلي؛ لأن الاتساق الداخلي يقوم أساسًا على تجانس العلاقات بين مكونات الاختبار (Nunnally & Bernstein, 1994).

ومن الناحية المفاهيمية، تفترض الطريقة الارتباطية أن الدرجة الكلية تمثل مؤشرًا معقولًا للبناء المقاس، وأن العلاقة بين الفقرة والدرجة الكلية تحمل دلالة على "انتماء" الفقرة للبناء، غير أن هذه الدلالة مشروطة بجودة الدرجة الكلية نفسها (من حيث الثبات وطول الاختبار) وبافتراضات البنية، مثل: أحادية البعد أو على الأقل غلبة بُعد رئيس، وبخصائص العينة (AERA et al., 2014؛ Lord & Novick, 1968). لذلك لا يُنظر إلى معاملات ارتباط الفقرة بالمجموع بوصفها أرقامًا نهائية، بل بوصفها مؤشرات تفسيرية تتطلب قراءة في ضوء شروط القياس.

2- أنواع معاملات الارتباط المستخدمة:

تتعدد معاملات الارتباط التي تُستعمل في تحليل الفقرات عند تقدير تمييز الفقرة بالطريقة الارتباطية، ويعود ذلك إلى طبيعة درجات الفقرة (غالبًا ثنائية في الفقرات الموضوعية) وطبيعة الدرجة الكلية (كمية)، ويمكن تنظيم أكثر المعاملات شيوعًا في ثلاثة أنماط رئيسة: الارتباط الثنائي النقطي، والارتباط الثنائي المتسلسل، وارتباط الفقرة بالمجموع المصحح.

ولنبدأ بالارتباط الثنائي النقطي فهو يُستخدم هذا المعامل عندما تكون درجة الفقرة ثنائية (مثل: صحيح/ خطأ أو واحد/ صفر) بينما تكون الدرجة الكلية كمية، ويُمثِّل في جوهره تطبيقًا لمعامل ارتباط بيرسون على متغير ثنائي وآخر كمي، بحيث يُعبِّر عن مدى اختلاف متوسط الدرجة الكلية بين من أجابوا الفقرة إجابة صحيحة ومن أخطؤوا فيها مع مراعاة التشتت العام للدرجات (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991). وتنبع أهميته من أنه يلتقط اتساق الفقرة عبر كامل العينة بدل الاقتصار على طرفين، ولذلك يُعد مؤشرًا شائعًا في تقدير تمييز الفقرة داخل الاختبارات التحصيلية والمقاييس النفسية ذات الفقرات الثنائية (القصابي، 2020). اما الارتباط الثنائي المتسلسل يُستخدم عندما يُفترض أن استجابة الفقرة الثنائية تعكس في الأصل متغيرًا كامنًا مستمرًا تم "تقطيعه" عند حد معين، أي أن الثنائية ليست طبيعة أصلية للسمة وإنما نتيجة قرار تصنيفي في القياس أو التصحيح، وبناءً على هذا الافتراض، يسعى هذا المعامل إلى تقدير الارتباط بين المتغير الكامن المستمر والدرجة الكلية، لا بين الثنائية الظاهرة والدرجة الكلية فحسب (Lord & Novick, 1968). غير أن قوة هذا المعامل تتوقف على مدى معقولية افتراض المتغير الكامن المستمر وعلى ملاءمة افتراضات التوزيع، وهي افتراضات قد لا تكون مضمونة في كل سياقات القياس التربوي، ما يستدعي استخدامه بحذر وقراءة نتائجه في ضوء طبيعة الفقرة والعينة (Crocker & Algina, 1986).

وأخيرًا ارتباط الفقرة بالمجموع المصحح تُعد هذه الصيغة استجابة مباشرة لمشكلة منهجية في الارتباط بين الفقرة والدرجة الكلية: إذا كانت درجة الفقرة جزءًا من الدرجة الكلية، فإن الارتباط قد يرتفع اصطناعيًا بسبب "تداخل" الفقرة في المحك الذي تُقارن به، وهو ما قد يقود إلى تقديرات مضللة خصوصًا عند قِصر الاختبار

(Henrysson, 1963). ولهذا يُحسب ارتباط الفقرة مع مجموع درجات الاختبار بعد حذف الفقرة نفسها، فيصبح المحك هو "بقية الاختبار" لا الاختبار كله، وقد بيّن Henrysson (1963) أن إدراج الفقرة داخل المجموع يميل إلى رفع الارتباط على نحو قد يكون خادعًا، واشتق صيغًا تصحيحية تُقلِّل هذا الأثر، مع ملاحظة أن بعض الصيغ تُظهر حساسية لطول الاختبار بينما تسعى أخرى إلى تقليل هذه الحساسية (Henrysson, 1963). ويُعد هذا التوجه متسقًا مع منطق الاتساق الداخلي: نحن نريد أن نعرف مدى اتساق الفقرة مع بقية الاختبار، لا مدى اتساقها مع نفسها ضمن المجموع.

3- خصائص الطريقة الارتباطية:

تتميز الطريقة الارتباطية بعدة مزايا تجعلها محورية في تحليل الفقرات ضمن النظرية الكلاسيكية في القياس، خاصة عندما يكون الهدف بناء اختبار متسق داخليًا وقابل للاستدلال، أبرز هذه المزايا أنها تستخدم معلومات العينة كلها، فهي لا تستبعد فئة الوسط، ولا تعتمد على قرار تقطيع يختزل التوزيع إلى طرفين، وهذا يجعلها – من حيث المبدأ – أقرب إلى تمثيل كيفية عمل الفقرة عبر مدى الدرجات الكلي، ويعزز قدرتها على التقاط أنماط الاتساق التي قد لا تظهر عند الاقتصار على طرفين (Crocker & Algina, 1986).

كما تتسق الطريقة الارتباطية مع هدف الاتساق الداخلي؛ لأن معاملات ارتباط الفقرة بالمجموع تعكس مباشرةً مقدار مساهمة الفقرة في البنية التراكمية للدرجة الكلية، فإذا كانت الفقرة تقيس البناء نفسه، فإنها تميل إلى أن ترتبط إيجابيًا بالدرجة الكلية، وإذا كانت تقيس شيئًا مختلفًا أو تتأثر بعوامل عرضية، فإن ارتباطها قد يضعف أو يصبح غير منتظم (Nunnally & Bernstein, 1994؛ AERA et al., 2014). ولهذا تُستخدم معاملات ارتباط الفقرة بالمجموع بوصفها مؤشرات أولية على جودة الفقرة وانسجامه مع البناء.

في المقابل، تحمل الطريقة الارتباطية حدودًا منهجية لا بد من إبرازها، **أول** هذه الحدود تأثرها بطول الاختبار، فالدرجة الكلية ليست معيارًا ثابت الخصائص؛ إذ إن طول الاختبار يؤثر في ثباتها وفي تباينها، ما ينعكس بدوره على معاملات الارتباط، وعندما يكون الاختبار قصيرًا، تصبح الدرجة الكلية أقل استقرارًا وأكثر عرضة لخطأ القياس العشوائي، فتضعف قدرة الارتباط على التقاط الاتساق الحقيقي بين الفقرة والبناء (Nunnally & Bernstein, 1994؛ Lord & Novick, 1968). ويزداد الأمر وضوحًا إذا حُسب الارتباط دون تصحيح، لأن إدراج الفقرة في المجموع يمكن أن يرفع الارتباط بصورة أكبر كلما قصر الاختبار، وهو ما أبرزه Henrysson (1963). و**ثاني** الحدود مشكلة إدراج الفقرة في المجموع، وقد سبقت الإشارة إليها، فالاعتماد على ارتباط غير مصحح قد يقود إلى قرارات حذف أو إبقاء على أساس تقدير متأثر ببنية الحساب، لا بجودة الفقرة فحسب (Henrysson, 1963). ولذلك تُعد الصيغة المصححة معيارًا منهجيًا أكثر اتساقًا مع الهدف التحليلي عندما تكون القرارات دقيقة أو عالية الأثر. و**ثالث** الحدود أن معاملات الارتباط – بحكم طبيعتها – تتأثر بخصائص العينة مثل: تقييد المدى وتجانس الدرجات، فإذا كانت العينة متقاربة جدًا في مستوياتها، تقل تباينات الدرجة الكلية، فيضعف الارتباط، حتى لو كانت الفقرة جيدة في مجتمع أكثر تباينًا، ما يحد من تعميم

الحكم على الفقرة دون مراعاة سياق العينة (Lord & Novick, 1968). وتظهر هنا قيمة التمييز بوصفه مؤشرًا يعتمد على "تباين حقيقي" في الأداء، لا على وجود بند جيد بمعزل عن خصائص العينة.

4- اعتبارات تفسيرية:

يُعد تفسير معاملات ارتباط الفقرة بالمجموع من أكثر المراحل حساسية؛ لأن الارتباط رقم واحد قد يخفي وراءه أسبابًا متعددة. فـــ الارتباط المنخفض قد يشير فعلًا إلى ضعف الفقرة في التمييز، مثل: أن تكون غامضة، أو أن مفتاح الإجابة غير مضبوط، أو أن الفقرة يعتمد على مهارة لغوية أو معرفية لا تتفق مع البناء المستهدف، لكنه قد يكون أيضًـــا نتيجة ظروف قياس لا تتصـــل بجودة الفقرة ذاتها، مثل: تجانس العينة أو قِصـــر الاختبار أو ملاءمة الصـــعوبة (Ebel & Frisbie, 1991؛ Lord & Novick, 1968). ومن ثم، فإن القراءة العلمية للارتباط المنخفض تستوجب طرح بدائل تفسيرية قبل اتخاذ حكم نهائي.

ومن الحالات التي قد يكون فيها الارتباط منخفضًـــا على نحو "مضـــلل" حالة تقييد المدى: إذا كان أغلب المفحوصـين في مسـتوى متقارب، فإن الارتباطات عمومًا تميل إلى الانخفاض لأن العلاقات الخطية تحتاج إلى تباين كافٍ في المتغيرين لتظهر. كذلك قد ينخفض الارتباط عندما تكون الفقرة شديدة السهولة أو شديدة الصعوبة؛ إذ تقل تباينات اســـتجاباتها، فتضـــعف قدرتها على إظهار علاقة منظمة مع الدرجة الكلية، حتى لو كان محتواها صـــحيحًا ومهمًا لأغراض محكية (Crocker & Algina, 1986). وفي المقابل، قد يبدو الارتباط مرتفعًا بصـــورة غير مســـتحقة عندما لا يُســـتخدم التصـــحيح؛ لأن إدراج الفقرة في المجموع يرفع العلاقة حســـابيًا (Henrysson, 1963).

ويتصل تفسير الارتباط أيضًا بمسألة أحادية البعد وتعدد الأبعاد. فإذا كان الاختبار يقيس بُعدًا واحدًا بصورة غالبية، يُتوقع أن تكون معاملات ارتباط الفقرة بالمجموع أكثر انتظامًا، وأن تعكس بالفعل انتماء الفقرات للبناء، أما إذا كان الاختبار متعدد الأبعاد، فقد ينخفض ارتباط فقرة جيدة لأنها تقيس بُعدًا فرعيًا لا يهيمن على الدرجة الكلية، أو لأنها تتســق مع بعد مختلف عن البعد الذي تســهم فيه بقية الفقرات، وفي هذه الحالة لا ينبغي تفســير الارتباط الضعيف بوصفه فشلًا مباشرًا للفقرة، بل بوصفه مؤشرًا لاحتمال عدم تجانس البناء أو الحاجة إلى تحليل بنيوي أعمق (AERA et al., 2014؛ Nunnally & Bernstein, 1994).

كما ينبغي الحذر من جعل الارتباط "حكمًا وحيدًا"، فالمعايير الحديثة تؤكد أن صـــدق الاســـتدلالات يعتمد على تجميع أدلة متعددة، وأن مؤشرات تحليل الفقرات ينبغي أن تُقرأ مع أدلة المحتوى، وأدلة البناء، وسياق استخدام الاختبار (AERA et al., 2014). وعليه، فإن الطريقة الارتباطية قوية في ربط الفقرة بالدرجة الكلية، لكنها لا تُغني عن تقييمات المحتوى ولا عن فهم طبيعة البناء المقاس.

5- متى تكون الطريقة الارتباطية أنسب نظريًا؟

تكون الطريقة الارتباطية أنسب نظريًا عندما يكون الهدف بناء اختبار تُفسَّر درجته الكلية بوصفها مؤشرًا على بناء واحد أو بنية متجانســة، وعندما يُراد تعظيم الاتســاق الداخلي ودعم قابلية الاســتدلال بالدرجة الكلية، ففي الاختبارات المعيارية أو الاختبارات التي تُســتخدم للمقارنة بين الأفراد، يصــبح الاتســاق الداخلي شــرطًا مهمًا،

وتصبح معاملات ارتباط الفقرة بالمجموع – خاصة المصححة – أدوات منسجمة مع هذا الهدف لأنها تُظهر مدى إسهام الفقرة في بنية الدرجة الكلية بصورة منظمة (Nunnally & Bernstein, 1994؛ Crocker & Algina, 1986).

وتزداد ملاءمة الطريقة الارتباطية عندما تتوفر عينة ذات حجم مناسب وتباين معقول؛ لأن ذلك يعزز استقرار معاملات الارتباط ويقلل من أثر التقلبات العشوائية، كما أنها تكون أكثر منطقية عندما يُراد الحكم على الفقرة عبر مدى الأداء كله، لا عبر طرفين فقط، وعندما يكون من المهم اكتشاف الفقرات التي تعمل جيدًا في الوسط مثلما تعمل في الأطراف (Ebel & Frisbie, 1991).

وأخيرًا، تبرز أفضلية الطريقة الارتباطية في السياقات التي تتطلب قراءة دقيقة لتأثير طول الاختبار وبنية حساب الدرجة الكلية، إذ يتيح استخدام ارتباط الفقرة بالمجموع المصحح تقليل الأثر المنهجي لإدراج الفقرة في المجموع، وهو ما يجعل الاستدلال بشأن جودة الفقرة أقرب إلى الهدف السيكومتري: تقدير مدى اتساق الفقرة مع بقية الاختبار لا مع نفسها ضمن الاختبار (Henrysson, 1963؛ Bazaldua et al., 2017). وبذلك يمكن النظر إلى الطريقة الارتباطية – ضمن شروطها – بوصفها إطارًا تقديريًا أكثر اتساقًا مع منطق بناء الاختبارات التي تسعى إلى تجانس الفقرات، وتعزيز دقة الدرجة الكلية.

المقارنة النظرية في ضوء حجم العينة وطول الاختبار

1- إطار المقارنة: ما الذي نقارنه تحديدًا؟

تستند المقارنة النظرية بين طريقة المجموعتين المتطرفتين والطريقة الارتباطية في تمييز الفقرة إلى أن كلتيهما تسعيان إلى تقدير "قدرة الفقرة على التفريق"، لكنهما تفعلان ذلك عبر منطقين مختلفين: منطق الفروق بين طرفين من توزيع الدرجات، ومنطق الاتساق المنتظم بين أداء الفقرة والدرجة الكلية، ولأن اختلاف المنطق يقود إلى اختلاف في خصائص التقدير، فإن إطار المقارنة في هذا الفصل يُبنى على أربعة محاور مترابطة: الاستقرار والدقة، التحيز المحتمل، الحساسية لتوزيع الدرجات، وقابلية التفسير واتخاذ القرار (Crocker & Algina, 1986؛ Lord & Novick, 1968).

أولًا: الاستقرار والدقة يشيران إلى مدى ثبات تقدير تمييز الفقرة لو أُعيد التطبيق على عينة مماثلة من المجتمع نفسه، وإلى مدى قرب التقدير من "التمييز الحقيقي" في المجتمع المفترض. داخل النظرية الكلاسيكية في القياس، يتأثر الاستقرار بخطأ القياس العشوائي، وبحجم العينة، وبقدرة الدرجة الكلية على تمثيل البناء المقاس (Nunnally & Bernstein, 1994). وبما أن طريقة المجموعتين تُسقط جزءًا كبيرًا من البيانات، بينما تعتمد الارتباطية على كامل العينة، فإن السؤال المركزي هنا: أيهما يوفّر تقديرًا أكثر استقرارًا تحت قيود حجم العينة وطول الاختبار؟

ثانيًا: التحيز المحتمل يعني انحراف التقدير بصورة منهجية في اتجاه معين بسبب طريقة الحساب أو خصائص العينة أو بنية الاختبار، فطريقة المجموعتين قد تُضخِّم الفروق أحيانًا بسبب اختيار الطرفين، أو تُقلِّلها بسبب سوء التصنيف عندما تكون الدرجة الكلية غير مستقرة، أما الطريقة الارتباطية فقد تتأثر بتداخل الفقرة داخل الدرجة

الكلية إذا لم يُستخدم الارتباط المصحح، فتظهر تقديرات أعلى مما تستحقه الفقرة (Henrysson, 1963). وعليه، فإن التحيز هنا لا يُفهم بوصفه خطأً عشوائيًا، بل بوصفه أثرًا منهجيًا يُمكن توقعه نظريًا.

ثالثًا: الحساسية لتوزيع الدرجات تعني مدى تأثر مؤشر التمييز بشكل توزيع الدرجات الكلية (الانحراف، التشتت، تقييد المدى)، وبخصائص صعوبة الاختبار، فطريقة المجموعتين تعتمد على "وجود طرفين" واضحين، بينما تعتمد الطريقة الارتباطية على تباين كافٍ في الدرجة الكلية وفي استجابات الفقرة كي تظهر العلاقة المنتظمة (Lord & Novick, 1968). لذلك تصبح خصائص التوزيع جزءًا من شروط صلاحية المقارنة.

رابعًا: قابلية التفسير واتخاذ القرار تتعلق بمدى وضوح معنى المؤشر لمتخذ القرار التربوي أو الباحث، ومدى ارتباطه بقرارات محددة مثل: حذف بند أو تعديله أو الإبقاء عليه، تمتلك طريقة المجموعتين ميزة تفسيرية مباشرة لأنها تُترجم إلى فرق واضح بين مجموعة عليا ودنيا، بينما تملك الطريقة الارتباطية ميزة تفسيرية مرتبطة ببناء الاختبار من حيث الاتساق الداخلي، خصوصًا عند استخدام الارتباط المصحح (Ebel & Frisbie, 1991؛ Crocker & Algina, 1986). وتنبع أهمية هذا المحور من أن المؤشر ليس هدفًا في ذاته، بل أداة ضمن قرار قياسي.

2- أثر حجم العينة على تقدير التمييز في كل طريقة:

يُعد حجم العينة أحد أكثر العوامل تأثيرًا في مؤشرات تحليل الفقرات داخل النظرية الكلاسيكية في القياس، لأن هذه المؤشرات تُقدَّر إحصائيًا وتخضع لتقلبات المعاينة، ويمكن فهم أثر حجم العينة هنا من خلال مبدأ عام: كلما صغرت العينة اتسعت تقلبات التقدير، وكلما كبرت تحسن الاستقرار، بشرط ثبات خصائص المجتمع وبنية الاختبار (Nunnally & Bernstein, 1994).

في طريقة المجموعتين المتطرفتين يظهر أثر حجم العينة بصورة مضاعفة لسببين. السبب الأول أن هذه الطريقة تُقلِّص العينة فعليًا لأنها لا تستخدم جميع المفحوصين، بل تعتمد على طرفين من التوزيع، وهذا يعني أن العينة "المستخدمة في الحساب" أصغر من العينة الأصلية، وأن نسب الاستجابة الصحيحة داخل كل مجموعة تصبح أكثر عرضة للتقلب عندما يكون حجم كل مجموعة محدودًا. والسبب الثاني أن اختيار الطرفين يعتمد على الدرجة الكلية، فإذا كانت العينة صغيرة تصبح عملية الفرز أكثر حساسية لتغيرات طفيفة في الدرجات، وتزداد احتمالات سوء التصنيف، خاصة عندما تكون درجات متعددة متقاربة أو عندما تكون الدرجة الكلية منخفضة الثبات (Kelley, 1939؛ Lord & Novick, 1968). وبذلك لا يتوقف أثر صِغر العينة على زيادة الخطأ العشوائي في نسب الإجابة فقط، بل يمتد إلى آلية تكوين المجموعتين ذاتها.

أما الطريقة الارتباطية فتستخدم كامل العينة، وهو ما يمنحها من حيث المبدأ ميزة استقرار أعلى عند تساوي حجم العينة الأصلي؛ إذ إن تقدير العلاقة يستند إلى جميع البيانات بدل الاقتصار على جزء منها، لكن هذه الميزة ليست مطلقة؛ لأن معاملات الارتباط نفسها تتطلب حجم عينة مناسبًا كي تستقر، وتتأثر بشدة عندما يكون

التباين محدودًا أو عندما تكون استجابات الفقرة شديدة التركز (مثل: بند شديد السهولة أو شديد الصعوبة) (Lord & Novick, 1968 ؛Crocker & Algina, 1986). إضافة إلى ذلك، فإن الارتباط غير المصحح قد يتضخم بصورة منهجية بسبب إدراج الفقرة في الدرجة الكلية، وهو تضخم قد يظهر بشكل أوضح في الاختبارات القصيرة، لكنه يظل قضية منهجية لا ترتبط بحجم العينة وحده (Henrysson, 1963).

ومن هنا تُطرح المناقشة الأساسية: لماذا قد تكون الطريقة الارتباطية أكثر استقرارًا نظريًا عند نفس حجم العينة؟

يمكن تبرير ذلك بأن استخدام جميع البيانات يخفف من تقلبات التقدير مقارنة باستخدام طرفين فقط، وأن الارتباط يُقدِّم تلخيصًا لعلاقة منتظمة عبر مدى الدرجات بدل الاعتماد على نقطتين تمثلان الطرفين، كما أن الارتباط المصحح يوجّه النظر إلى اتساق الفقرة مع بقية الاختبار، وهو ما ينسجم مع طبيعة التحليل البنائي للاختبار (Henrysson, 1963؛ Nunnally & Bernstein, 1994).

لكن متى لا تكون الارتباطية أكثر استقرارًا؟

تضعف أفضلية الارتباطية عندما يحدث تقييد شديد في المدى بسبب تجانس العينة؛ إذ إن تقارب درجات المفحوصين يقلل التباين اللازم لظهور علاقة واضحة بين الفقرة والمجموع، فينخفض الارتباط، حتى لو كانت الفقرة جيدة في مجتمع أكثر تباينًا (Lord & Novick, 1968). كما قد تتراجع الاستفادة من الارتباطية عندما تكون بنية الاختبار متعددة الأبعاد بصورة قوية؛ إذ تصبح الدرجة الكلية خليطًا من مصادر تباين متعددة، فينخفض ارتباط بعض الفقرات لأنها تقيس بُعدًا فرعيًا مشروعًا لكنه غير مسيطر على المجموع (AERA et al., 2014). وفي هذه الحالة قد تبدو طريقة المجموعتين في ظاهرها "أكثر حدة" في إظهار الفرق بين طرفين، لكنها تظل معرضة لعدم الاستقرار إذا كانت العينة صغيرة أو إذا كان الفرز مبنيًا على درجة كلية مضطربة.

3- أثر طول الاختبار على التمييز:

يؤثر طول الاختبار في تمييز الفقرة عبر قنوات مختلفة في كل طريقة، لأن طول الاختبار يغير طبيعة الدرجة الكلية بوصفها معيارًا: فهو يؤثر في ثباتها وتباينها ودقة استخدامها كمحك داخلي. وفي النظرية الكلاسيكية في القياس يُعد تحسن الثبات مع زيادة طول الاختبار فكرة مركزية في فهم جودة الدرجة الكلية (Nunnally & Bernstein, 1994).

في الطريقة الارتباطية يتجلى أثر طول الاختبار بصورة مباشرة لأن الدرجة الكلية تدخل في حساب العلاقة، عندما يطول الاختبار (مع افتراض جودة فقراته وتجانه النسبي)، تتحسن استقراريه الدرجة الكلية وتقل نسبة الخطأ العشوائي فيها، وبذلك يصبح ارتباط الفقرة بالمجموع أكثر قدرة على التقاط الاتساق الحقيقي بين الفقرة والبناء المقاس، أما عندما يكون الاختبار قصيرًا جدًا، فإن الدرجة الكلية تكون "مزعجة" بمعنى أن نسبة الضوضاء فيها أعلى، فتضعف العلاقة المنتظمة بين الفقرة والمجموع، حتى لو كانت الفقرة مناسبة؛ لأن المحك نفسه غير مستقر (Lord & Novick, 1968؛ Nunnally & Bernstein, 1994).

ويضاف إلى ذلك أثر منهجي محدد: إذا لم يُستخدم الارتباط المصحح، فإن إدراج الفقرة داخل المجموع يرفع الارتباط بصورة اصطناعية، ويكون هذا الارتفاع أشد وضوحًا عندما يكون الاختبار قصيرًا؛ لأن مساهمة الفقرة في المجموع تكون أكبر نسبيًا مقارنة باختبار طويل، لذلك يُعد تصحيح ارتباط الفقرة بالمجموع خطوة أساسية لضبط أثر الطول، ولجعل المؤشر أقرب إلى الدلالة المرادة: اتساق الفقرة مع بقية الاختبار (Henrysson, 1963؛ Crocker & Algina, 1986).

أما في طريقة المجموعتين المتطرفتين فإن أثر طول الاختبار يكون غالبًا غير مباشر. فطول الاختبار يؤثر في دقة تصنيف الأفراد إلى مجموعة عليا ودنيا. فالاختبار الأطول – إذا كانت فقراته متسقة – يوفر درجة كلية أكثر تمييزًا بين الأفراد، ما يقلل من سوء التصنيف عند تكوين الطرفين ويجعل الفرق بينهما أكثر تمثيلًا للفروق الحقيقية على البناء المقاس. في المقابل، الاختبار القصير جدًا يجعل الفرز إلى عليا ودنيا أكثر حساسية للخطأ العشوائي، فتدخل حالات من ذوي المستوى الحقيقي المتوسط في إحدى المجموعتين، أو حالات من ذوي المستوى الحقيقي المرتفع في المجموعة الدنيا والعكس، ما يضعف الفرق في نسب الإجابة الصحيحة ويخفض مؤشر التمييز (Lord & Novick, 1968).

لكن حتى مع اختبار طويل، تظل طريقة المجموعتين متأثرة بفقدان معلومات الوسط؛ فهي لا تستفيد من التحسن في تمثيل الدرجات إلا بقدر ما ينعكس على وضوح الطرفين، بينما الارتباطية تستثمر التحسن عبر جميع مفحوصي العينة، ولهذا يُتوقع نظريًا أن يرفع طول الاختبار جودة تقدير التمييز في الطريقتين، لكن آلية التحسن وأثره تختلفان (Crocker & Algina, 1986؛ Ebel & Frisbie, 1991).

4– تفاعل حجم العينة وطول الاختبار: سيناريوهات تفسيرية:

لأن حجم العينة وطول الاختبار يتفاعلان في تحديد استقرار مؤشرات الفقرات، يمكن تنظيم المقارنة في أربعة سيناريوهات نظرية دون الحاجة إلى بيانات رقمية:

السيناريو الأول: عينة صغيرة مع اختبار قصير

يمثل هذا السيناريو الحالة الأكثر إشكالًا لكلا الطريقتين، ففي طريقة المجموعتين يتضاعف عدم الاستقرار: عينة صغيرة أصلًا ثم تقليصها إلى طرفين صغيرين، مع فرز يعتمد على درجة كلية قصيرة منخفضة الاستقرار، وفي الطريقة الارتباطية تتأثر العلاقة بضعف الدرجة الكلية، وباحتمال تضخيم الارتباط غير المصحح، في هذه الحالة يُرجح نظريًا أن تكون الارتباطية المصححة أفضل نسبيًا إذا توفرت شروط تباين معقول، لأنها لا تُسقط وسط العينة، لكنها تظل محدودة بسبب قصر الاختبار وصغر العينة. أما طريقة المجموعتين فتحتاج حذرًا شديدًا لأن تقلب نسب الإجابة في الطرفين قد يقود إلى أحكام متسرعة (Henrysson, 1963؛ Nunnally & Bernstein, 1994).

السيناريو الثاني: عينة صغيرة مع اختبار طويل

هنا يتحسن الوضع جزئيًا: طول الاختبار يرفع استقرار الدرجة الكلية، فيتحسن فرز الأفراد في طريقة المجموعتين، ويتحسن محك الارتباط في الطريقة الارتباطية، لكن صِغر العينة يبقى عاملًا محدِّدًا، خاصة لطريقة المجموعتين التي تقلص العينة المستخدمة، لذلك يُرجح نظريًا أن تبقى الطريقة الارتباطية المصححة أكثر اتساقًا من حيث الاستقرار، بينما يمكن لطريقة المجموعتين أن تُستخدم كمؤشر داعم إذا كانت العينة، رغم صغرها، تسمح بتكوين طرفين غير هزيلين في الحجم (Crocker & Algina, 1986؛ Lord & Novick, 1968).

السيناريو الثالث: عينة كبيرة مع اختبار قصير

في هذا السيناريو يخف أثر صِغر الطول جزئيًا بسبب وفرة البيانات، فتتحسن استقراريه نسب الإجابة في المجموعتين، ويستقر تقدير الارتباط بسبب كبر العينة، إلا أن قصر الاختبار يبقى مؤثرًا في دقة الدرجة الكلية وفي مشكلة إدراج الفقرة في المجموع. وعليه، تكون الطريقة الارتباطية المصححة مناسبة لأنها تقلل التضخيم، وتستفيد من حجم العينة الكبير، بينما تكون طريقة المجموعتين مفيدة لسهولة تفسيرها مع بقاء قيد أن الفرز يستند إلى درجة كلية قصيرة قد تحمل ضوضاء أعلى مما ينبغي (Henrysson, 1963؛ Ebel & Frisbie, 1991).

السيناريو الرابع: عينة كبيرة مع اختبار طويل

يُعد هذا السيناريو الأكثر ملاءمة لتحليل الفقرات في إطار النظرية الكلاسيكية في القياس، فحجم العينة الكبير يدعم استقرار التقديرات، وطول الاختبار يدعم ثبات الدرجة الكلية، ما يجعل كلتا الطريقتين أكثر صلاحية من حيث الاستقرار، ومع ذلك، تظل الارتباطية – خصوصًا المصححة – أكثر اتساقًا مع هدف بناء اختبار متجانس لأنها تستخدم كل البيانات وتُعبِّر عن اتساق الفقرة مع بقية الاختبار. أما طريقة المجموعتين فتظل مفيدة كقراءة "تفريقيه" مساندة تُظهر الفرق بين الطرفين بصورة سهلة الفهم (Crocker & Algina, 1986؛ Nunnally & Bernstein, 1994).

5– خلاصة المقارنة: نقاط قوة وقصور واتساقها مع هدف القياس:

تدل المقارنة النظرية على أن كل طريقة تُجسِّد معنى معينًا للتمييز، وأن اختيار الطريقة ينبغي أن يُبنى على هدف القياس وشروط التطبيق، لا على تفضيل شكلي لمؤشر بعينه.

فإذا كان الهدف اتخاذ قرار سريع بحذف بند أو إبقائه في سياق تربوي محدود، وكانت العينة متوسطة أو كبيرة بما يكفي لتكوين مجموعتين واضحتين، فإن طريقة المجموعتين المتطرفتين قد تكون مفيدة لأنها مباشرة التفسير وتكشف الفقرات التي تفشل بوضوح في التفريق بين الطرفين، غير أن هذه الفائدة مشروطة بالحذر من تقلب التقدير في العينات الصغيرة، وبالحذر من سوء الفرز عند قصر الاختبار أو انخفاض ثبات الدرجة الكلية، وبوعي أن إهمال الوسط قد يحجب أداء الفقرة عبر مدى الدرجات (Kelley, 1939؛ Ebel & Frisbie, 1991).

أما إذا كان الهدف تحسين الاتساق الداخلي وبناء اختبار يُراد لدرجته الكلية أن تكون قابلة للاستدلال بصورة معيارية، فإن الطريقة الارتباطية تبدو أكثر اتساقًا نظريًا؛ لأنها تقيس مدى اندماج الفقرة في بنية الاختبار عبر العينة كلها، ولأن الارتباط المصحح يحد من التحيز الناتج عن إدراج الفقرة في المجموع، ويجعل تفسير التمييز أقرب إلى "اتساق الفقرة مع بقية الاختبار" (Henrysson, 1963؛ Nunnally & Bernstein, 1994)، ومع ذلك، لا ينبغي أن يُفهم الارتباط المنخفض دائمًا بوصفه ضعفًا جوهريًا في الفقرة، فقد يكون انعكاسًا لتجانس العينة أو لتعدد أبعاد البناء أو لعدم ملاءمة صعوبة الاختبار، وهو ما يستدعي قراءة تمييز الفقرة ضمن منظومة أدلة أوسع تشمل المحتوى والبناء وشروط التطبيق (AERA et al., 2014؛ Lord & Novick, 1968).

وبذلك تُختتم المقارنة بخلاصة معيارية: طريقة المجموعتين مناسبة كأداة تفسيرية سريعة عندما تكون شروطها متحققة ويُراد إبراز الفروق الطرفية، بينما الطريقة الارتباطية المصححة أنسب عندما يكون المقصود بناء اختبار متجانس، وتعزيز الدقة، والاستدلال بالدرجة الكلية، خاصة في سياقات القياس المعياري أو القرارات الأعلى أثرًا، وفي كل الأحوال، يبقى حجم العينة وطول الاختبار عاملين حاكمين في استقرار التقدير، ويجب أن يُفهم تمييز الفقرة بوصفه مؤشرًا مشروطًا بسياق القياس لا قيمة مجردة منفصلة عنه (Crocker & Algina, 1986؛ علّام، 2000)

خلاصات نظرية وتوصيات بحثية:

1- الاستنتاجات النظرية المتوقعة:

تُفضي المقارنة النظرية بين طريقة المجموعتين المتطرفتين والطريقة الارتباطية في تقدير تمييز الفقرة إلى مجموعة استنتاجات تُبرز اختلاف "معنى التمييز" الذي تجسده كل طريقة، وتوضح كيف يتغير استقرار التقدير بتغير حجم العينة وطول الاختبار.

أولًا: يتمثل الفرق الجوهري في منطق القياس:

1) تُجسِّد طريقة المجموعتين المتطرفتين التمييز بوصفه فارقًا في الأداء بين طرفين من توزيع الدرجات الكلية، أي أنها تركز على "حدة التفريق" بين مرتفعي ومنخفضي الأداء (كيلي، 1939؛ إيبل وفرسبي، 1991).
2) بينما تُجسِّد الطريقة الارتباطية التمييز بوصفه اتساقًا منتظمًا بين أداء المفحوص على الفقرة وأدائه على الاختبار ككل عبر كامل مدى الدرجات، وهو ما يجعلها أقرب إلى منطق الاتساق الداخلي والانسجام البنائي للاختبار (كروكر وألجينا، 1986؛ نونالي وبرنستين، 1994).

ثانيًا: يختلفان في مصدر المعلومات المستخدمة:

1) طريقة المجموعتين المتطرفتين تُهمل عمدًا استجابات وسط التوزيع، وتبني تقديرها على جزء من العينة، ما يجعلها أكثر عرضة لفقد معلومات قد تكون مهمة لفهم سلوك الفقرة عبر مستويات الأداء المتوسطة (لورد ونوفيك، 1968).

2) الطريقة الارتباطية تستخدم جميع البيانات، وتمنح تقديرًا يعكس النمط العام للعلاقة بين الفقرة والدرجة الكلية، وهو ما يدعم "شموليّة" القراءة، خصوصًا عندما تكون بنية الاختبار متجانسة (كروكر وألجينا، 1986).

ثالثًا: يتضح أن استقرار التقدير يتغير مع حجم العينة (ن) عبر آليات مختلفة في الطريقتين:

1) في طريقة المجموعتين المتطرفتين ينشأ عدم الاستقرار من تقلب نسب الإجابة في مجموعتين أصغر حجمًا من العينة الأصلية، فضلًا عن حساسية الفرز عندما تكون الدرجات الكلية متقاربة أو منخفضة الثبات، ومع صِغر (ن) يتضخم أثر العشوائية وسوء التصنيف، فتزداد قابلية مؤشر التمييز للتذبذب بين عينة وأخرى (بريتان، 1972؛ لورد ونوفيك، 1968).

2) في الطريقة الارتباطية يُتوقع – نظريًا – تحسن الاستقرار مع زيادة (ن) لأن تقدير الارتباط يستند إلى كامل العينة. لكن هذا التحسن مشروط بوجود تباين كافٍ في الدرجات الكلية وبألا تكون استجابات الفقرة شديدة التركز (سهولة أو صعوبة مفرطة)، وبسلامة افتراضات بناء الاختبار، كما أن إدراج الفقرة داخل الدرجة الكلية قد يرفع التقدير بصورة منهجية إذا لم يُستخدم الارتباط المصحح، وهو شكل من أشكال عدم الدقة لا يرتبط بالعشوائية بقدر ارتباطه ببنية الحساب (هنريسون، 1963).

رابعًا: يتغير استقرار التقدير مع طول الاختبار لأن طول الاختبار يعيد تشكيل خصائص الدرجة الكلية التي تعتمد عليها الطريقتان:

1) في الطريقة الارتباطية يؤثر طول الاختبار مباشرةً عبر ثبات الدرجة الكلية وتباينها، فكلما كان الاختبار أطول وأكثر تجانسًا ارتفعت جودة الدرجة الكلية كمحك داخلي، وزادت قدرة الارتباط على التقاط الاتساق الحقيقي بين الفقرة وبقية الاختبار، أما قِصر الاختبار فيرفع نسبة الضوضاء في الدرجة الكلية، ويضعف العلاقة المنتظمة، وقد يجعل الارتباط غير المصحح أكثر عرضة للتضخم بحكم أن مساهمة الفقرة في المجموع تكون أكبر نسبيًا (هنريسون، 1963؛ نونالي وبرنستين، 1994).

2) في طريقة المجموعتين المتطرفتين يكون أثر طول الاختبار غالبًا غير مباشر: فالاختبار الأطول – إذا كان متجانسًا – يحسن دقة تصنيف الأفراد إلى مجموعتين عليا ودنيا لأن الدرجة الكلية تصبح أقل تأثرًا بخطأ القياس العشوائي، فتتراجع احتمالات سوء التصنيف، أما الاختبار القصير جدًا فيضعف فرز الأفراد ويزيد من اختلاط المستويات الحقيقية بين المجموعتين، فيتقلص الفرق المتوقع في أداء الفقرة حتى لو كانت الفقرة جيدة (لورد ونوفيك، 1968).

خامسًا: يتبين أن الفروق بين الطريقتين ليست تفاضلًا مطلقًا، بل تفاضلًا مرتبطًا بهدف القياس وبشروط التطبيق، فالمؤشر الأكثر "معنى" هو الذي ينسجم مع الغرض من الاختبار وطبيعة العينة وبنية الاختبار وأدلة الصدق المتاحة، وتدعم معايير بناء الاختبارات فكرة أن مؤشرات تحليل الفقرات ينبغي أن تُقرأ ضمن حزمة أدلة، لا بوصفها أحكامًا منفردة، وأن القرارات المتعلقة بالفقرة يجب أن تستند إلى مزيج من دليل المحتوى، ودليل البناء، ودليل الإحصاءات المناسبة لسياق الاستخدام (الجمعية الأمريكية للبحث التربوي وآخرون، 2014).

2- توصيات:

أولًا: توصيات لبناة الاختبارات

1. عند استخدام طريقة المجموعتين المتطرفتين، يُوصى بالنظر إليها بوصفها مؤشرًا تفريقيًا سريعًا يُناسب السياقات التربوية التطبيقية التي تتطلب قرارًا مباشرًا، مع ضرورة مراعاة أن هذا المؤشر يتأثر بقوة بحجم العينة الفعّال بعد التقسيم وبجودة الدرجة الكلية المستخدمة في الفرز (كيلي، 1939؛ بريتان، 1972).
2. عند استخدام الطريقة الارتباطية، يُوصى بالاعتماد على ارتباط الفقرة بالمجموع المصحح لأن ذلك يقلل الأثر المنهجي لإدراج الفقرة داخل المجموع، ويجعل التفسير أقرب إلى "اتساق الفقرة مع بقية الاختبار" لا مع نفسها ضمن الدرجة الكلية (هنريسون، 1963).
3. يُوصى بعدم اتخاذ قرار حذف أو إبقاء بند اعتمادًا على مؤشر واحد فقط، بل ينبغي الجمع بين مؤشرات الصعوبة والتمييز، ومراجعة مفتاح الإجابة، وصياغة البدائل، والتحقق من تمثيل المحتوى. فالفقرة قد يكون ضعيف التمييز لأسباب تتعلق بملاءمة صعوبته للعينة أو بتجانس العينة، لا بضعف المحتوى ذاته (إيبل وفرسبي، 1991؛ علّام، 2000).
4. يُوصى بقراءة نتائج التمييز في ضوء خصائص العينة، وبخاصة مدى تباينها وتجانسها؛ لأن تقييد المدى قد يخفض معاملات التمييز الارتباطية، ويجعل تفسيرها خارج سياق العينة مضللًا (لورد ونوفيك، 1968).
5. في الاختبارات ذات القرارات الأعلى أثرًا، يُوصى بتدعيم تحليل الفقرات بإجراءات تُقيّم تجانس البناء، لأن ضعف التمييز قد يكون علامة على تعدد الأبعاد، أو على وجود بعد فرعي مشروع يحتاج إلى تنظيم بنائي (الجمعية الأمريكية للبحث التربوي وآخرون، 2014؛ نونالي وبرنستين، 1994).

ثانيًا: توصيات للباحثين

1. يُوصى بتنفيذ دراسات نظرية ومحاكاة منهجية تُظهر كيف تتغير خصائص مؤشرات التمييز تحت شروط مختلفة من حجم العينة وطول الاختبار وصعوبة الفقرات، بما يساعد على صياغة قواعد تفسير أكثر دقة من الاعتماد على أحكام عامة
2. يُوصى بإجراء مقارنات منهجية بين مؤشرات التمييز في النظرية الكلاسيكية في القياس ومؤشرات التمييز في نظرية الاستجابة للفقرة، مع التركيز على ما إذا كانت الاستنتاجات المتعلقة بالفقرة تتغير عبر الإطارين ولماذا، دون افتراض تطابق الدلالة بين المؤشرات المختلفة
3. يُوصى بفحص أثر تعدد الأبعاد وتقييد المدى على معاملات ارتباط الفقرة بالمجموع، وبناء نماذج تفسيرية توضح الحالات التي يكون فيها الارتباط المنخفض "دليلًا على مشكلة"، والحالات التي يكون فيها "أثرًا لعينة أو لبنية اختبار"

المراجع:

القصابي، خليفة بن أحمد بن حميد. (2020). تحليل الفقرات في بناء المقاييس النفسية: الصدق الظاهري، صدق الفقرات، الصدق العاملي. *المجلة الدولية للدراسات التربوية والنفسية*، 8(3)، 541–555. https://doi.org/10.31559/EPS2020.8.3.1

علّام، صلاح الدين محمود. (2000). *القياس والتقويم التربوي والنفسي*. دار الفكر العربي للطباعة والنشر. القاهرة، مصر.

عودة، أحمد سليمان. (1985). *القياس والتقويم في العملية التدريسية*. دار الأمل للطباعة والنشر والتوزيع. عمان، الأردن.

ملحم، سامي محمد. (2002). *القياس والتقويم في التربية وعلم النفس*. دار المسيرة للطباعة والنشر والتوزيع. عمان، الأردن.

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Brooks/Cole Publishing Company.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Prentice Hall.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Bazaldua, D. A. L., Lee, Y.-S., Keller, B., & Fellers, L. (2017). Estimation bias in item discrimination. Asia Pacific Education Review, 18(4), 585–598. https://doi.org/10.1007/s12564-017-9502-1

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 32(2), 289–303. https://doi.org/10.1177/001316447203200204

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Lawrence Erlbaum Associates.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice Hall.

Gulliksen, H. (1950). Theory of mental tests. Wiley

Haladyna, T. M. (2004). Developing and validating multiple-choice test items (3rd ed.). Lawrence Erlbaum Associates.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. Psychometrika, 28(2), 211–218. https://doi.org/10.1007/BF02289590

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1), 17–24. https://doi.org/10.1037/h0057123

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151–160. https://doi.org/10.1007/BF02288391

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Thorndike, R. M., & Thorndike-Christ, T. (2010). Measurement and evaluation in psychology and education (8th ed.). Pearson.